The Cybernetic Bayesian Brain

From Interoceptive Inference to Sensorimotor Contingencies

Anil K. Seth

Is there a single principle by which neural operations can account for perception, cognition, action, and even consciousness? A strong candidate is now taking shape in the form of "predictive processing". On this theory, brains engage in predictive inference on the causes of sensory inputs by continuous minimization of prediction errors or informational "free energy". Predictive processing can account, supposedly, not only for perception, but also for action and for the essential contribution of the body and environment in structuring sensorimotor interactions. In this paper I draw together some recent developments within predictive processing that involve predictive modelling of internal physiological states (*interoceptive inference*), and integration with "enactive" and "embodied" approaches to cognitive science (*predictive perception of sensorimotor contingencies*). The upshot is a development of predictive processing that originates, not in Helmholtzian perception-as-inference, but rather in 20th-century cybernetic principles that emphasized homeostasis and predictive control. This way of thinking leads to (i) a new view of emotion as active interoceptive inference; (ii) a common predictive framework linking experiences of body ownership, emotion, and exteroceptive perception; (iii) distinct interpretations of active inference as involving disruptive and disambiguatory—not just confirmatory—actions to test perceptual hypotheses; (iv) a neurocognitive operationalization of the "mastery of sensorimotor contingencies" (where sensorimotor contingencies reflect the rules governing sensory changes produced by various actions); and (v) an account of the sense of subjective reality of perceptual contents ("perceptual presence") in terms of the extent to which predictive models encode potential sensorimotor relations (this being "counterfactual richness"). This is rich and varied territory, and surveying its landmarks emphasizes the need for experimental tests of its key contributions.

Keywords
Active inference | Counterfactually-equipped predictive model | Evolutionary robotics | Free energy principle | Interoception | Perceptual presence | Predictive processing | Sensorimotor contingencies | Somatic marker hypothesis | Synaesthesia

Author

Anil K. Seth
a.k.seth@sussex.ac.uk
University of Sussex
Brighton, United Kingdom

Commentator

Wanja Wiese
wawiese@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

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

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

An increasingly popular theory in cognitive science claims that brains are essentially prediction machines (Hohwy 2013). The theory is variously known as the Bayesian brain (Knill & Pouget 2004; Pouget et al. 2013), predictive processing (Clark 2013; Clark this collection), and the predictive mind (Hohwy 2013; Hohwy this collection), among others; here we use the term PP (predictive processing). (See Table 1 for a glossary of technical terms.) At its most fundamental, PP says that perception is the result of the brain inferring the most likely causes of its sensory inputs by minimizing the difference between actual sensory signals and the signals expected on the basis of continuously updated predictive models. Arguably, PP provides the most complete framework to date for explaining perception, cognition, and action in terms of fundamental theoretical principles and neurocognitive architectures. In this paper I describe a version of PP that is distinguished by (i) an emphasis on predictive modelling of in-

Table 1: A glossary of technical terms.

Allostasis	The process of achieving homeostasis.
Active inference	Classically conceived as the minimization of prediction error by performing actions that confirm sensory predictions. However, as argued in this paper, it may also involve the performance of actions to disconfirm current predictions or to disambiguate among competing perceptual hypotheses.
Counterfactually-equipped predictive model	A predictive or generative model that encodes not only the likely causes of current sensory inputs but also (and explicitly) the likely causes of fictive sensory inputs conditioned on possible but unexecuted actions.
Counterfactual richness	A predictive model is counterfactually rich if it encodes a rich repertoire of potential sensorimotor relations, i.e., relations between potential actions and their expected sensory consequences.
Exteroception/exteroceptive	The classic senses conveying signals originating in the external environment.
Free energy	An information-theoretic quantity that bounds or limits the surprise associated with encountering an input, given a generative/predictive model mapping causes to sensory inputs. Under fairly general assumptions, free energy is the long-run sum of prediction error.
Free energy principle (FEP)	The FEP says that organisms obey a fundamental imperative towards the avoidance of (information-theoretically) surprising events, according to which they must minimize the long-run average surprise of sensory states, since surprising sensory states are (in the long run) likely to reflect conditions incompatible with continued existence.
Homeostasis	Any regulative processes that enable a system to keep certain variables within specific bounds.
Interoception/interoceptive	The sense of the internal physiological condition of the body.
Interoceptive inference	The predictive modelling of internal physiological states.
Interoceptive sensitivity	A characterological trait that reflects individual sensitivity to interoceptive signals, usually operationalized via heartbeat detection tasks.
Perceptual presence	The sense of the subjective reality of the contents of perception.
PPSMC	Predictive Perception of SensoriMotor Contingencies. A new theory that integrates predictive processing with sensorimotor theory. It says that mastery of a sensorimotor contingency is equivalent to the induction and deployment of a counterfactually-equipped predictive model linking potential actions to their expected sensory consequences.
Predictive processing (PP)/predictive coding	A scheme, dating back at least to Hermann von Helmholtz, which conceives of perception as probabilistic inference on the causes of sensory signals. Predictive coding is one specific implementation of predictive processing that rests on algorithms developed in the setting of data compression.
Sensorimotor contingency (SMC)	SMCs describe ways in which sensory signals change given actions in specific contexts; they are "rules" describing sensorimotor dependencies.
Sensorimotor theory	A cognitive theory which says that visual experiences arises from an implicit knowledge or mastery of SMCs. On this theory, perception is an activity.

ternal physiological states and (ii) engagement with alternative frameworks under the banner of "enactive" and "embodied" cognitive science (Varela et al. 1993).

I first identify an unusual starting point for PP, not in Helmholtzian perception-as-inference, but in the mid 20th-century cybernetic theories associated with W. Ross Ashby (1952, 1956; Conant & Ashby 1970). Linking these origins to their modern expression in Karl Friston's "free energy principle" (2010), perception emerges as a *consequence* of a more fundamental imperative towards homeostasis and control, and not as a process designed to furnish a detailed inner "world model" suitable for cognition and action planning. The ensuing view of PP, while still fluently accounting for (exteroceptive) perception, turns out to be more naturally applicable to the predictive perception of internal bodily states, instantiating a process of *interoceptive inference* (Seth 2013; Seth et al. 2011). This concept provides a natural way of thinking of the neural substrates of emotional and mood experiences, and also describes a common mechanism by which interoceptive and exteroceptive signals can be integrated to provide a unified experience of body ownership and conscious selfhood (Blanke & Metzinger 2009; Limanowski & Blankenburg 2013).

The focus on embodiment leads to distinct interpretations of *active inference*, which in general refers to the selective sampling of sensory signals so as to improve perceptual predictions. The simplest interpretation of active inference is the changing of sensory data (via selective sampling) to conform to current predictions (Friston et al. 2010). However, by analogy with hypothesis testing in science, active inference can also involve seeking evidence that goes *against* current predictions, or that *disambiguates* multiple competing hypotheses. A nice example of the latter comes from self-modelling in evolutionary robotics, where multiple competing self-models are used to specify actions that are most likely to provide disambiguatory sensory evidence (Bongard et al. 2006). I will spend more time on this example later. Crucially, these different senses of active inference rest on the capacity of predictive models to encode *counterfactual* relations linking potential (but not necessarily executed) actions to their expected sensory consequences (Friston et al. 2012; Seth 2014b). It also implies the involvement of model comparison and selection—not just the optimization of parameters assuming a single model. These points represent significant developments in the basic infrastructure of PP.

The notion of counterfactual predictions connects PP with what at first glance seems to be its natural opponent: "enactive" theories of perception and cognition that explicitly reject internal models or representations (Clark this collection; Hutto & Myin 2013; Thompson & Varela 2001). Central to the enactive approach are notions of "sensorimotor contingencies" and their "mastery" (O'Regan & Noë 2001), where a sensorimotor contingency refers to a rule governing how sensory signals change in response to action. On this approach, the perceptual experience of (for example) redness is given by an implicit knowledge (mastery) of the way red things behave given certain patterns of sensorimotor activity. This mastery of sensorimotor contingencies is also said to underpin *perceptual presence*: the sense of subjective reality of the contents of perception (Noë 2006). From the perspective of PP, mastery of a sensorimotor contingency corresponds to the learning of a counterfactually-equipped predictive model connecting potential actions to expected sensory consequences. The resulting theory of PPSMC (Predictive Perception of SensoriMotor Contingencies), Seth 2014b) provides a much needed reconciliation of enactive and predictive theories of perception and action. It also provides a solution to the challenge of perceptual presence within the setting of PP: perceptual presence obtains when the underlying predictive models are *counterfactually rich*, in the sense of encoding a rich repertoire of potential (but not necessarily executed) sensorimotor relations. This approach also helps explain instances where perceptual presence seems to be lacking, such as in synaesthesia.

This is both a conceptual and theoretical paper. Space limitations preclude any significant treatment of the relevant experimental lit-

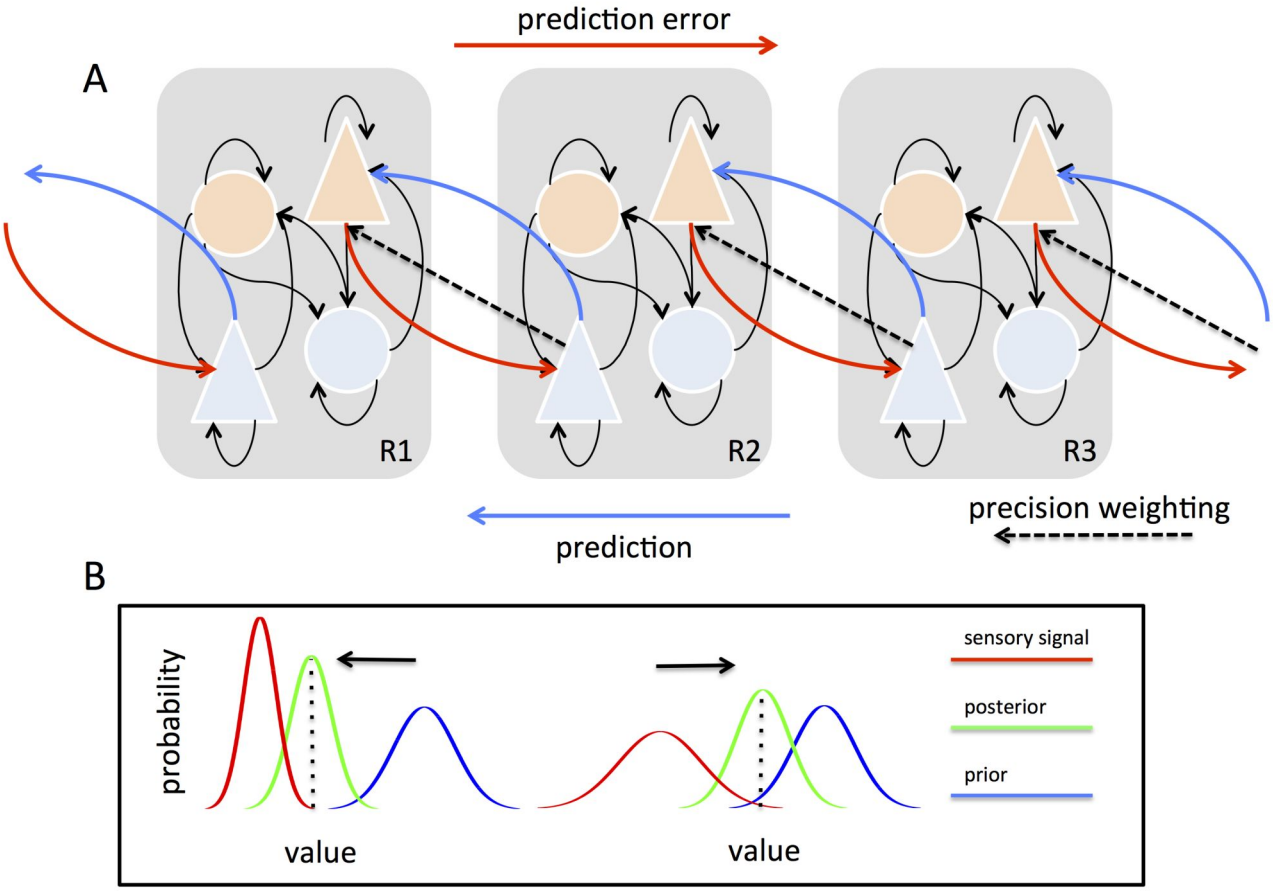

Figure 1: A. Schemas of hierarchical predictive coding across three cortical regions; the lowest on the left (R1) and the highest on the right (R3). Bottom-up projections (red) originate from "error units" (orange) in superficial cortical layers and terminate on "state units" (light blue) in the deep (infragranular) layers of their targets; while top-down projections (dark blue) convey predictions originating in deep layers and project to the superficial layers of their targets. Prediction errors are associated with precisions, which determine the relative influence of bottom-up and top-down signal flow via precision weighting (dashed lines). **B.** The influence of precisions on Bayesian inference and predictive coding. The curves show probability distributions over the value of a sensory signal (x-axis). On the left, high precision-weighting of sensory signals (red) enhances their influence on the posterior (green) and expectation (dotted line) as compared to the prior (blue). On the right, low sensory precision weighting has the opposite effect. Figure adapted from Seth (2013).

erature. However, even an exhaustive treatment would reveal that this literature so far provides only circumstantial support for the basics of PP, let alone for the extensions described here. Yet an advantage of PP theories is that they are grounded in concrete computational processes and neurocognitive architectures, giving us confidence that informative experimental tests can be devised. Implementing such an experimental agenda stands as a critical challenge for the future.

2 The predictive brain and its cybernetic origins

2.1 Predictive processing: The basics

PP starts with the assumption that in order to support adaptive responses, the brain must discover information about the external "hidden" causes of sensory signals. It lacks any direct access to these causes, and can only use information found in the flux of sensory signals them-

selves. According to PP, brains meet this challenge by attempting to predict sensory inputs on the basis of their own emerging models of the causes of these inputs, with prediction errors being used to update these models so as to minimize discrepancies. The idea is that a brain operating this way will come to encode (in the form of predictive or generative models) a rich body of information about the sources of signals by which it is regularly perturbed (Clark 2013).

Applied to cortical hierarchies, PP overturns classical notions of perception that describe a largely "bottom-up" process of evidence accumulation or feature detection. Instead, PP proposes that perceptual content is determined by top-down predictive signals emerging from multi-layered and hierarchically-organized generative models of the causes of sensory signals (Lee & Mumford 2003). These models are continually refined by mismatches (prediction errors) between predicted signals and actual signals across hierarchical levels, which iteratively update predictive models via approximations to Bayesian inference (see Figure 1). This means that the brain can induce accurate generative models of environmental hidden causes by operating only on signals to which it has direct access: *predictions* and *prediction errors*. It also means that even low-level perceptual content is determined via cascades of predictions flowing from very general abstract expectations, which constrain successively more fine-grained predictions.

Two further aspects of PP need to be emphasized from the outset. First, sensory prediction errors can be minimized either "passively", by changing predictive models to fit incoming data (perceptual inference), or "actively", by performing actions to confirm or test sensory predictions (active inference). In most cases these processes are assumed to unfold continuously and simultaneously, underlining a deep continuity between perception and action (Friston et al. 2010; Verschure et al. 2003). This process of active inference will play a key role in much of what follows. Second, predictions and prediction errors in a Bayesian framework have associated *precisions* (inverse variances, Figure 1). The precision of a prediction error is an indicator of its reliability, and hence can be used to determine its influence in updating top-down predictive models. Precisions, like mean values, are not given but must be inferred on the basis of top-down models and incoming data; so PP requires that agents have *expectations about precisions* that are themselves updated as new data arrive (and new precisions can be estimated). Precision expectations can therefore balance the influence of different prediction-error sources on the updating of predictive models. And if prediction errors have low (expected) precision, predictive models may overwhelm error signals (hallucination) or elicit actions that confirm sensory predictions (active inference).

A picture emerges in which cortical networks engage in recurrent interactions whereby bottom-up prediction errors are continuously reconciled with top-down predictions at multiple hierarchical levels—a process modulated at all times by precision weighting. The result is a brain that not only encodes information about the sources of signals that impinge upon its sensory surfaces, but that also encodes information about how its own actions interact with these sources in specifying sensory signals. *Perception* involves updating the parameters of the model to fit the data; *action* involves changing sensory data to fit (or test) the model; and *attention* corresponds to optimizing model updating by giving preference to sensory data that are expected to carry more information, which is called precision weighting (Hohwy 2013). This view of the brain is shamelessly model-based and representational (though with a finessed notion of representation), yet it also deeply embeds the close coupling of perception and action and, as we will see, the importance of the body in the mediation of this interaction.

2.2 Predictive processing and the free energy principle

PP can be considered a special case of the *free energy principle*, according to which perceptual inference and action emerge as a consequence of a more fundamental imperative towards the avoidance of "surprising" events (Friston 2005, 2009, 2010). On the free energy principle, or-

ganisms – by dint of their continued survival—must minimize the long-run average surprise of sensory states, since surprising sensory states are likely to reflect conditions incompatible with continued existence (think of a fish out of water). "Surprise" is not used here in the psychological sense, but in an information-theoretic sense—as the negative log probability of an event's occurrence (roughly, the unlikeliness of the occurrence of an event).

The connection with PP arises because agents cannot directly evaluate the (information-theoretic) surprise associated with an event, since this would require—impossibly—the agent to average over all possible occurrences of the event in all possible situations. Instead, the agent can only maintain a lower limit on surprise by minimizing the difference between actual sensory signals and those signals predicted according to a generative or predictive model. This difference is *free energy*, which, under fairly general assumptions, is the long-run sum of prediction error.

An attractive feature of the free energy principle is that it brings to the table a rich mathematical framework that shows how PP can work in practice. Formally, PP depends on established principles of Bayesian inference and model specification, whereby the most likely causes of observed data (*posterior*) are estimated based on optimally combining *prior expectations* of these causes with observed data, by using a (generative, predictive) model of the data that would be observed given a particular set of causes (*likelihood*). (See Figure 1 for an example of priors and posteriors.) In practice, because optimal Bayesian inference is usually intractable, a variety of approximate methods can be applied (Hinton & Dayan 1996; Neal & Hinton 1998). Friston's framework appeals to previously worked-out "variational" methods, which take advantage of certain approximations (e.g., Gaussianity, independence of temporal scales)—thus allowing a potentially neat mapping onto neurobiological quantities (Friston et al. 2006).[1]

1 Some challenging questions surface here as to whether prediction errors are used to update priors, which corresponds to standard Bayesian inference, or whether they are used to update the underlying generative/predictive model, which corresponds to learning.

The free energy principle also emphasizes *action* as a means of prediction error minimization, this being *active inference*. In general, active inference involves the selective sampling of sensory signals so as to minimize uncertainty in perceptual hypotheses (minimizing the entropy of the posterior). In one sense this means that actions are selected to provide evidence compatible with current perceptual predictions. This is the most standard interpretation of the concept, since it corresponds most directly to minimization of prediction error (Friston 2009). However, as we will see, actions can also be selected on the basis of an attempt to find evidence going against current hypotheses, and/or to efficiently disambiguate between competing hypotheses. These finessed senses of active inference represent developments of the free energy framework. Importantly, action itself can be thought of as being brought about by the minimization of *proprioceptive* prediction errors via the engagement of classical reflex arcs (Adams et al. 2013; Friston et al. 2010). This requires transiently low precision-weighting of these errors (or else predictions would simply be updated instead), which is compatible with evidence showing sensory attenuation during self-generated movements (Brown et al. 2013).

A more controversial aspect of the free energy principle is its claimed generality (Hohwy this collection). At least as described by Friston, it claims to account for adaptation at almost any granularity of time and space, from macroscopic trends in evolution, through development and maturation, to signalling in neuronal hierarchies (Friston 2010). However, in some of these interpretations reliance on predictive modelling is only implicit; for example the body of a fish can be considered to be an implicit model of the fluid dynamics and other affordances of its watery environment (see section 2.3). I am not concerned here with these broader interpretations, but will focus on those cases in which biological (neural) mechanisms plausibly implement explicit predictive inference via approximations to Bayesian computations—namely, the Bayesian brain (Knill & Pouget 2004; Pouget et al. 2013). Here, the free energy principle has potentially the greatest explanat-

ory power, especially given the convergence of empirical evidence (see Clark 2013 and Hohwy 2013 for reviews) and computational modelling showing how cortical microcircuits might implement approximate Bayesian inference (Bastos et al. 2012).

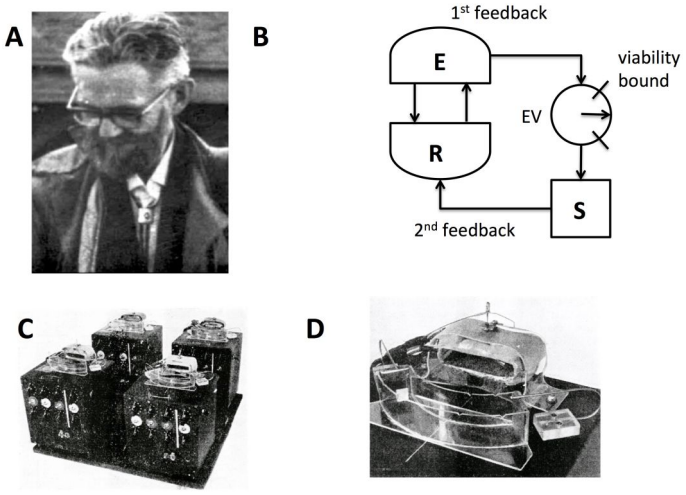

Figure 2: A. W. Ross Ashby, British psychiatrist and pioneer of cybernetics (1903–1972). **B.** A schematic of ultrastability, based on Ashby's notebooks. The system *R* homeostatically maintains its essential variables (EVs) within viability limits via first-order feedback with the environment *E*. When first-order feedback fails, so that EVs run out-of-bounds, second order "ultrastable" feedback is triggered so that *S* (an internal controller, potentially model-based) changes the parameters of R governing the first-order feedback. *S* continually changes *R* until homeostatic relations are regained, leaving the EVs again within bounds. **C.** Ashby's "homeostat", consisting of four interconnected ultrastable systems, forming a so-called "multistable" system. **D.** One ultrastable unit from the homeostat. Each unit had a trough of water with an electric field gradient and a metal needle. Instability was represented by the non-central needle positions, which on occurring would alter the resistances connecting the units via discharge through capacitors. For more details see Ashby (1952) and Pickering (2010).

2.3 Predictive processing, free energy, and cybernetics

Typically, the origins of PP are traced to the work of the 19th Century physiologist Hermann von Helmholtz, who first formalized the idea of perception as inference. However, the Helmholtzian view is rather passive, inasmuch as there is little discussion of active inference or behaviour. The close coupling of perception and action emphasized in the free energy principle points instead to a deep connection between PP and mid-twentieth-century cybernetics. This is most obvious in the works of W. Ross Ashby (Ashby 1952; 1956; Conant & Ashby 1970) but is also evident more generally (Dupuy 2009; Pickering 2010). Importantly, cybernetics adopted as its central focus the *prediction and control of behaviour* in so-called teleological or purposeful machines.[2] More precisely, cybernetic theorists were (are) interested in systems that appear to have goals (i.e., teleological) and that participate in circular causal chains (i.e., involving feedback) coupling goal-directed sensation and action.

Two key insights from the first wave of cybernetics usefully anticipate the core developments of PP within cognitive science. These are both associated with Ashby, a key figure in the movement and often considered its leader, at least outside the USA (Figure 2).

The first insight consists in an emphasis on the homeostasis of internal *essential variables*, which, in physiological settings, correspond to quantities like blood pressure, heart rate, blood sugar levels, and the like. In Ashby's framework, when essential variables move beyond specific viability limits, adaptive processes are triggered that re-parameterize the system until it reaches a new equilibrium in which homeostasis is restored (Ashby 1952). Such systems are, in Ashby's terminology, *ultrastable*, since they embody (at least) two levels of feedback: a first-order feedback that homeostatically regulates essential variables (like a thermostat) and a second-order feedback that allostatically[3] re-organises a system's input–output relations when first-order feedback fails, until a new homeostatic regime is attained. In the most basic case, as implemented in Ashby's famous "homeostat" (Figure 2), this second-order feedback simply involves random changes to system

2 This underlines the close links between cybernetics and behaviourism. Perhaps this explains why cybernetics was so reluctant to bring phenomenology into its remit, an exclusion which, looking back, seems like a missed opportunity.

3 Allostasis: the process of achieving homeostasis.

parameters until a new stable regime is reached. The importance of this insight for PP is that it locates the function of biological and cognitive processes in generalizing homeostasis to ensure that internal essential variables remain within expected ranges.

Another way to summarize the fundamental cybernetic principle is to say that adaptive systems ensure their continued existence by successfully responding to environmental perturbations so as to maintain their internal organization. This leads to the second insight, evident in Ashby's *law of requisite variety*. This states that a successful control system must be capable of entering at least as many states as the system being controlled: "only variety can force down variety" (Ashby 1956). This induces a functional boundary between controller and environment and implies a minimum level of complexity for a successful controller, which is determined by the causal complexity of the environmental states that constitute potential perturbations to a system's essential variables. This view was refined some years later, in a 1970 paper written with Roger Conant entitled "Every good regulator of a system must be a model of that system" (Conant & Ashby 1970). This paper builds on the law of requisite variety by arguing (and attempting to formally show) that the nature of a controller capable of suppressing perturbations imposed by an external system (e.g., the world) must instantiate a model of that system. This provides a clear connection with the free energy principle, which proposes that adaptive systems minimize a limit on free energy (long-run average surprise) by inducing and refining a generative model of the causes of sensory signals. It also moves beyond Ashby's homeostat by implying that model-based controllers can engage in more successful multilevel feedback than is possible by random variation of higher-order parameters.

Putting these insights together provides a distinctive way of seeing the relevance of PP to cognition and biological adaptation. It can be summarized as follows. The purpose of cognition (including perception and action) is to maintain the homeostasis of essential variables and of internal organization (ultrastability). This implies the existence of a control mechanism with sufficient complexity to respond to (i.e., suppress) the variety of perturbations it encounters (law of requisite variety). Further, this structure must instantiate a model of the system to be controlled (good regulator theorem), where the system includes both the body and the environment (and their interactions). As Ashby himself tells us "[t]he whole function of the brain can be summed up in: error correction" (quoted in Clark 2013, p. 1). Put this way, perception emerges as a *consequence* of a more fundamental imperative towards organizational homeostasis, and not as a stage in some process of internal world-model construction. This view, while highlighting different origins, closely parallels the assumptions of the free energy principle in proposing a primary imperative towards the continued survival of the organism (Friston 2010).

It may be surprising to consider the legacy of cybernetics in this light. This is because many previous discussions of this legacy focus on examples which show that complex, apparently goal-directed behaviour can emerge from simple mechanisms interacting with structured bodies and environments (Beer 2003; Braitenberg 1984). On this more standard development, cybernetics challenges rather than asserts the need for internal models and representations: it is often taken to justify slogans of the sort "the world is its own best model" (Brooks 1991). In fact, cybernetics is agnostic with respect to the need for deployment of explicit internally-specified predictive models. If environmental circumstances are reasonably stable, and mappings between perturbations and (homeostatic) responses reasonably straightforward, then the good regulator theorem can be satisfied by controllers that only implicitly model their environments. This is the case, for instance, in the Watt governor: a device that is able exquisitely to control the output of (for instance) a steam engine, in virtue of its mechanism, and not through the deployment of explicit predictive models or representations (see Figure 3 and Van Gelder 1995; note that the governor can

be described as an implicit model since it has variables – e.g., eccentricity of the metal balls from the central column – which map onto environmental variables that affect the homeostatic target – engine output). However, where there exist many-to-many mappings between sensory states and their probable causes, as may be the case more often than not, it will pay to engage explicit inferential processes in order to extract the most probable causes of sensory states, insofar as these causes threaten the homeostasis of essential variables.

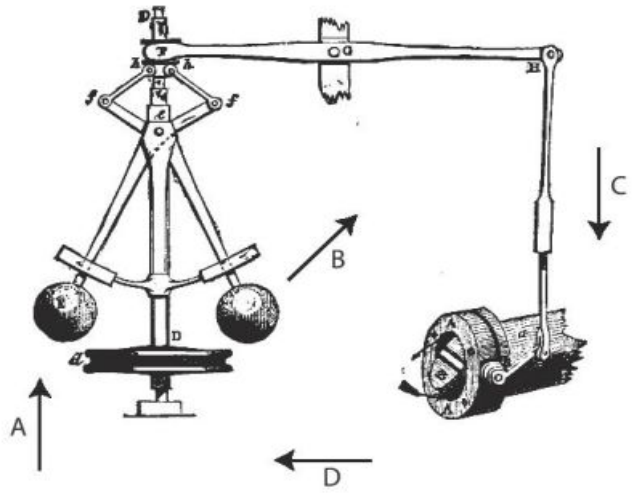

Figure 3: The Watt governor. This system, a central contributor to the industrial revolution, enabled precise control over the output of (for example) steam engines. As the speed of the engine increases, power is supplied to the governor (A) by a belt or chain, causing it to rotate more rapidly so that the metal balls have more kinetic energy. This causes the balls to rise (B), which closes the throttle valve (C), thereby reducing the steam flow, which in turn reduces engine speed (D). The opposite happens when the engine speed decreases, so that the governor maintains engine speed at a precise equilibrium.

In summary, rather than seeing PP as originating solely in the Helmholtzian notion of "perception as inference", it is fruitful to see it also as part of a process of model-based *predictive control* entailed by a fundamental imperative towards internal homeostasis. This shift in perspective reveals a distinctive agenda for PP in cognitive science, to which I shall now turn.

3 Interoceptive inference, emotion, and predictive selfhood

3.1 Interoceptive inference and emotion

Considering the cybernetic roots of PP, together with the free energy principle, leads to a potentially counterintuitive idea. This is that PP may apply more naturally to *interoception* (the sense of the internal physiological condition of the body) than to *exteroception* (the classic senses, which carry signals that originate in the external environment). This is because for an organism it is more important to avoid encountering unexpected interoceptive states than to avoid encountering unexpected exteroceptive states. A level of blood oxygenation or blood sugar that is unexpected is likely to be bad news for an organism, whereas unexpected exteroceptive sensations (like novel visual inputs) are less likely to be harmful and may in some cases be desirable, as organisms navigate a delicate balance between exploration and exploitation (Seth 2014a), testing current perceptual hypotheses through active inference (see section 5, below), all ultimately in the service of maintaining organismic homeostasis.

Perhaps because of its roots in Helmholtz, PP has largely been developed in the setting of visual neuroscience (Rao & Ballard 1999), with a related but somewhat independent line in motor control (Wolpert & Ghahramani 2000). Recently, an explicit application of PP to interoception has been developed (Seth 2013; Seth & Critchley 2013; Seth et al. 2011; see also Gu et al. 2013). On this theory of *interoceptive inference* (or equivalently *interoceptive predictive coding*), emotional states (i.e., subjective feeling states) arise from top-down predictive inference of the causes of interoceptive sensory signals (see Figure 4). In direct analogy to exteroceptive PP, emotional content is constitutively specified by the content of top-down interoceptive predictions *at a given time*, marking a distinction with the well-studied impact of expectations on *subsequent* emotional states (see e.g., Ploghaus et al. 1999; Ueda et al. 2003). Furthermore, interoceptive prediction errors can

be minimized by (i) updating predictive models (perception, corresponding to new emotional contents); (ii) changing interoceptive signals through engaging autonomic reflexes (autonomic control or active inference); or (iii) performing behaviour so as to alter external conditions that impact on internal homeostasis (allostasis; Gu & Fitzgerald 2014; Seth et al. 2011).

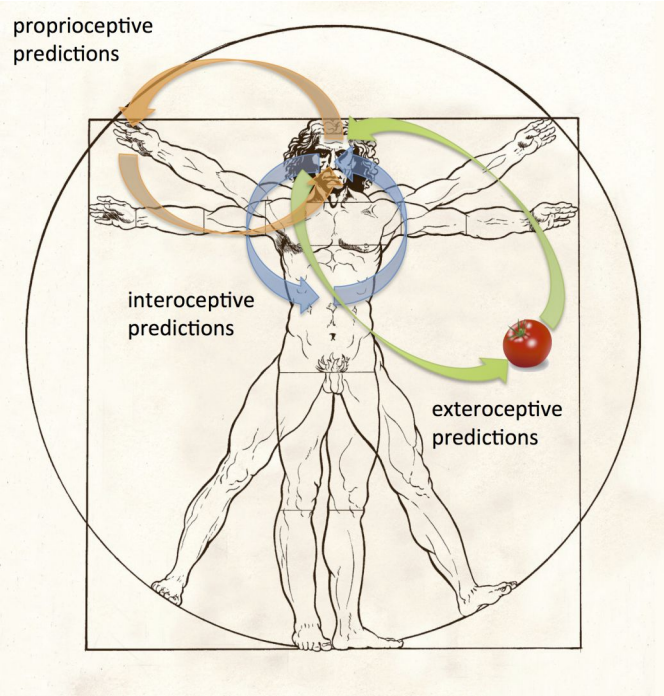

Figure 4: Inference and perception. Green arrows represent exteroceptive predictions and predictions errors underpinning perceptual content, such as the visual experience of a tomato. Orange arrows represent proprioceptive predictions (and prediction errors) underlying action and the experience of body ownership. Blue arrows represent interoceptive predictions (and prediction errors) underlying emotion, mood, and autonomic regulation. Hierarchically higher levels will deploy multimodal and even amodal predictive models spanning these domains, which are capable of generating multimodal predictions of afferent signals.

Consider an example in which blood sugar levels (an essential variable) fall towards or beyond viability thresholds, reaching unexpected and undesirable values (Gu & Fitzgerald 2014; Seth et al. 2011). Under interoceptive inference, the following responses ensue. First, interoceptive prediction error signals update top-down expectations, leading to subjective experiences of hunger or thirst (for sugary things). Because these feeling states are themselves surprising (and non-viable) in the long run, they signal prediction errors at hierarchically-higher levels, where predictive models integrate multimodal interoceptive and exteroceptive signals. These models instantiate predictions of temporal sequences of matched exteroceptive and interoceptive inputs, which flow down through the hierarchy. The resulting cascade of prediction errors can then be resolved either through autonomic control, in order to metabolize bodily fat stores (active inference), or through allostatic actions involving the external environment (i.e., finding and eating sugary things).

The sequencing and balance of these events is governed by relative precisions and their expectations. Initially, interoceptive prediction errors have high precision (weighting) given a higher-level expectation of stable homeostasis. Whether the resulting high-level prediction error engages autonomic control or allostatic behaviour (or both) depends on the precision weighting of the corresponding prediction errors. If food is readily available, consummatory actions lead to food intake (as described earlier, these actions are generated by the resolution of proprioceptive prediction errors). If not, autonomic reflexes initiate the metabolization of bodily fat stores, perhaps alongside appetitive behaviours that are predicted to lead to the availability of food, conditioned on performing these behaviours.[4]

3.2 Implications of interoceptive inference

Several interesting implications arise when considering emotion as resulting from interoceptive inference (Seth 2013). First, the theory generalizes previous "two factor" theories of emotion that see emotional content as resulting from an interaction between the perception of physiolo-

4 It is interesting to consider possible dysfunctions in this process. For example, if high-level predictions about the persistence of low blood sugar become abnormally strong (i.e., low blood sugar becomes chronically expected), allostatic food-seeking behaviours may not occur. This process, akin to the transition from hallucination to delusion in perceptual inference (Fletcher & Frith 2009), may help understand eating disorders in terms of dysfunctional signalling of satiety.

gical changes (James 1894) and "higher-level" cognitive appraisal of the context within which these changes occur (Schachter & Singer 1962). Instead of distinguishing "physiological" and "cognitive" levels of description, interoceptive inference sees emotional content as resulting from the multi-layered prediction of interoceptive input spanning many levels of abstraction. Thus, interoceptive inference integrates cognition and emotion within the powerful setting of PP.

The theory also connects with influential frameworks that link interoception with decision making, notably the "somatic marker hypothesis" proposed by Antonio Damasio (1994). According to the somatic marker hypothesis, intuitive decisions are shaped by interoceptive responses (somatic markers) to potential outcomes. This idea, when placed in the context of interoceptive inference, corresponds to the guidance of behavioural (allostatic) responses towards the resolution of interoceptive prediction error (Gu & Fitzgerald 2014; Seth 2014a). It follows that intuitive decisions should be affected by the degree to which an individual maintains accurate predictive models of his or her own interoceptive states; see Dunn et al. 2010, Sokol-Hessner et al. 2014 for evidence along these lines.

There are also important implications for disorders of emotion, selfhood, and decision-making. For example, anxiety may result from the chronic persistence of interoceptive prediction errors that resist top-down suppression (Paulus & Stein 2006). Dissociative disorders like alexithymia (the inability to describe one's own emotions), and depersonalization and derealisation (the loss of sense of reality of the self and world) may also result from dysfunctional interoceptive inference, perhaps manifest in abnormally low interoceptive precision expectations (Seth 2013; Seth et al. 2011). In terms of decision-making, it may be productive to think of addiction as resulting from dysfunctional active inference, whereby strong interoceptive priors are confirmed through action, overriding higher-order or hyper-priors relating to homeostasis and organismic integrity. It has even been suggested that autism spectrum disorders may originate in aberrant encoding of the salience or precision of interoceptive prediction errors (Quattrocki & Friston 2014). The reasoning here is that aberrant salience during development could disrupt the assimilation of interoceptive and exteroceptive cues within generative models of the "self", which would impair a child's ability to properly assign salience to socially relevant signals.

3.3 The predictive embodied self

The maintenance of physiological homeostasis solely through direct autonomic regulation is obviously limited: behavioural (allostatic) interactions with the world are necessary if the organism is to avoid surprising physiological states in the long run. The ability to deploy adaptive behavioural responses mandates the original Helmholtzian view of perception-as-inference, which has been the primary setting for the development of PP so far. A critical but arguably overlooked middle ground, which mediates between physiological state variables and the external environment, is the *body*. On one hand, the body is the material vehicle through which behaviour is expressed, permitting allostatic interactions to take place. On the other, the body is itself an essential part of the organismic system, the homeostatic integrity of which must be maintained. In addition, the experience of owning and identifying with a particular body is a key component of being a conscious self (Apps & Tsakiris 2014; Blanke & Metzinger 2009; Craig 2009; Limanowski & Blankenburg 2013; Seth 2013).

It is tempting to ask whether common predictive mechanisms could underlie not only classical exteroceptive perception (like vision) and interoception (see above), but also their integration in supporting conscious and unconscious representations of the body and self (Seth 2013). The significance of this question is underlined by realising that just as the brain has no direct access to causal structures in the external environment, it also lacks direct access to its own body. That is, given that the brain is in the business of inferring the causal sources of

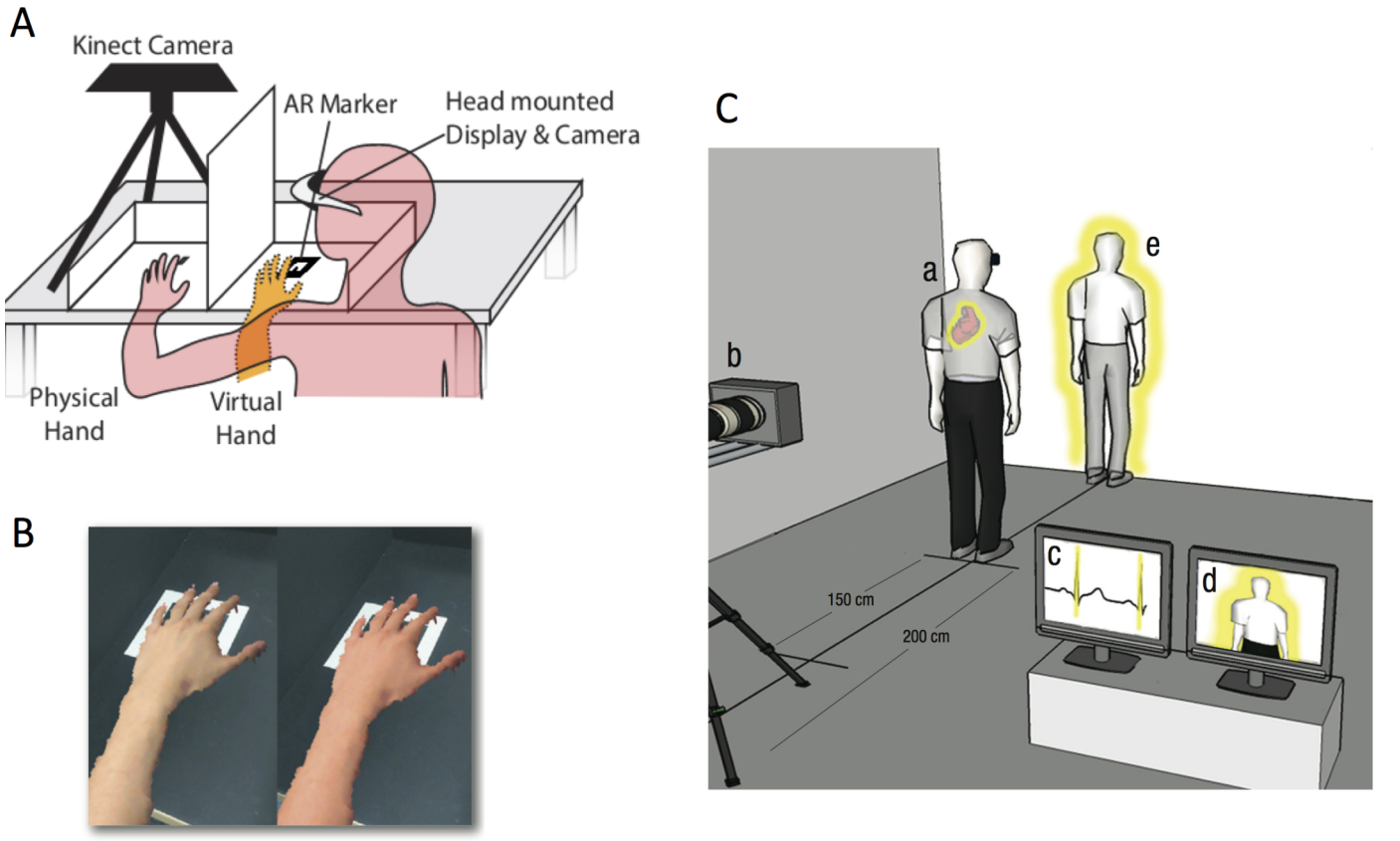

Figure 5: The interaction of interoceptive and exteroceptive signals in shaping the experience of body ownership. **A.** Set-up for applying cardio-visual feedback in the rubber hand illusion. A Microsoft Kinect obtains a real-time 3D model of a subject's left hand. This is re-projected into the subject's visual field using a head-mounted display and augmented reality (AR) software. **B.** The colour of the virtual hand is modulated by the subject's heart-beat. **C.** A similar set-up for the full-body illusion whereby a visual image of a subject's body is surrounded by a halo pulsing either in time or out of time with the heartbeat. Panels A and B are adapted from Suzuki et al. (2013); panel C is adapted from Aspell et al. (2013).

sensory signals, a key challenge emerges when distinguishing those signals that pertain to the body from those that originate from the external environment. A clue to how this challenge is met is that the physical body, unlike the external environment, constantly generates and receives internal input via its interoceptive and proprioceptive systems (Limanowski & Blankenburg 2013; Metzinger 2003). This suggests that the experienced body (and self) depends on the brain's best guess of the causes of those sensory signals most likely to be "me" (Apps & Tsakiris 2014), across interoceptive, proprioceptive, and exteroceptive domains (Figure 4).

There is now considerable evidence that the *experience of body ownership* is highly plastic and depends on the multisensory integration of body-related signals (Apps & Tsakiris 2014; Blanke & Metzinger 2009). One classic example is the *rubber hand illusion*, where the stroking of an artificial hand synchronously with a participant's real hand, while visual attention is focused on the artificial hand, leads to the experience that the artificial hand is somehow part of the body (Botvinick & Cohen 1998). According to current multisensory integration models, this change in the experience of body ownership is due to correlation between vision and touch overriding conflicting proprioceptive inputs (Makin et al. 2008). Through the lens of PP, this implies that prediction errors induced by multisensory conflicts will over time update self-related priors (Apps & Tsakiris 2014), with different signal sources (vision, touch, proprioception) each precision-weighted according to their expected reliability, and all in

the setting of strong prior expectations for correlated input.[5]

While the potential for exteroceptive multisensory integration to modulate the experience of body ownership has been extensively explored both for the ownership of body parts and for the experience of ownership of the body as a whole (for reviews, see Apps & Tsakiris 2014; Blanke & Metzinger 2009), only recently has attention been paid to interactions between interoceptive and exteroceptive signals. Initial evidence in this line of investigation was indirect, for example showing correlation between susceptibility to the rubber hand illusion and individual differences in the ability to perceive interoceptive signals ("interoceptive sensitivity", typically indexed by heartbeat detection tasks; Tsakiris et al. 2011). Other relevant studies have shown that body ownership illusions lead to temperature reductions in the corresponding body parts, perhaps reflecting altered active autonomic inference (Moseley et al. 2008; Salomon et al. 2013).

Emerging evidence now points more directly towards the predictive multisensory integration of interoceptive and exteroceptive signals in shaping the experience of body ownership. Two recent studies have taken advantage of so-called "cardio-visual synchrony" where virtual-reality representations of body parts (Suzuki et al. 2013) or the whole body (Aspell et al. 2013) are modulated by simultaneously recorded heartbeat signals, with the modulation either in-time or out-of-time with the actual heartbeat (Figure 5). These data suggest that statistical correlations between interoceptive (e.g., cardiac) and exteroceptive (e.g., visual) signals can lead to the updating of predictive models of self-related signals through (hierarchical) minimization of prediction error, just as happens for purely exteroceptive multisensory conflicts in the classic rubber hand illusion.

While these studies underline the plausibility of common predictive mechanisms underlying emotion, selfhood, and perception, many open questions nevertheless remain. A key challenge is to detail the underlying neural operations. Though a detailed analysis is beyond the scope of the present paper, it is worth noting that attention is increasingly focused on the insular cortex (especially its anterior parts) as a potential source of interoceptive predictions, and also as a comparator registering interoceptive prediction errors. The anterior insula has long been considered a major cortical locus for the integration of interoceptive and exteroceptive signals (Craig 2003; Singer et al. 2009); it is strongly implicated in interoceptive sensitivity (Critchley et al. 2004); it is sensitive to interoceptive prediction errors—at least in some contexts (Paulus & Stein 2006); and it has a high density of so-called "von Economo" neurons,[6] which have been frequently though circumstantially associated with consciousness and selfhood (Critchley & Seth 2012; Evrard et al. 2012).

3.4 Active inference, self-modeling, and evolutionary robotics

What role might *active* inference play in predictive self-modelling? Autonomic changes during illusions of body ownership (see above) are consistent with active inference; however they do not speak directly to its function. In the classic rubber hand illusion, hand or finger movements can be considered active inferential tests of self-related hypotheses. If these movements are not reflected in the "rubber hand", the illusion is destroyed—presumably because predicted visual signals are not confirmed (Apps & Tsakiris 2014). However, if hand movements are mapped to a virtual "rubber hand"—through clever use of virtual and augmented reality—the illusion is in fact strengthened, presumably because the multisensory correlation of peri-hand visual and proprioceptive signals constitutes a more stringent test of the perceptual hypothesis of ownership of the virtual hand (Suzuki et al. 2013). This introduces the idea that active inference is not simply about confirming sensory predictions but also involves seeking "disruptive" actions that are most informative with respect to testing current predictions,

5 Interestingly the expectation of perceptual correlations seems to be sufficient for inducing the rubber hand illusion (Ferri et al. 2013).

6 These are long-range projection neurons found selectively in hominid primates and certain other species.

and/or at disambiguating competing predictions (Gregory 1980). A nice example of how this happens in practice comes from *evolutionary robotics*[7]—which is obviously a very different literature, though one that inherits directly from the cybernetic tradition.

In a seminal 2006 study, Josh Bongard and colleagues described a four-legged "starfish" robot that engaged in a process much like active inference in order to model its own morphology so as to be able to control its movement and attain simple behavioural goals (Bongard et al. 2006). While there are important differences between evolutionary robotics and (active) Bayesian inference, there are also broad similarities; importantly, both can be cast in terms of model selection and optimization.

The basic cycle of events is shown in Figure 6. The robot itself is shown in the centre (A). The goal is to develop a controller capable of generating forward movement. The challenge is that the robot's morphology is unknown to the robot itself. The system starts with a range of (generic prior) potential self-models (B), here specified by various configurations of three-dimensional physics engines. The robot performs a series of initially random actions and evaluates its candidate self-models on their ability to predict the resulting proprioceptive afferent signals. Even though all initial models will be wrong, some may be better than others. The key step comes next. The robot evaluates new candidate actions *on the extent to which the current best self-models make different predictions as to their (proprioceptive) consequences*. These disambiguating actions are then performed, leading to a new ranking of self-models based on their success at proprioceptive prediction. This ranking, via the evolutionary robotics methods of mutation and replication, gives rise to a new population of candidate self-models. The upshot is that the system swiftly develops accurate self-models that can be used to generate controllers enabling movement (D). An interesting feature of this process is that it is highly resilient to unexpected perturbations. For instance, if a leg is removed then proprioceptive prediction errors will immediately ensue. As a result, the system will engage in another round of self-model evolution (including the co-specification of competing self-models and disambiguating actions) until a new, accurate, self-model is regained. This revised self-model can then be used to develop a new gait, allowing movement, even given the disrupted body (E, F).[8]

This study emphasizes that the operational criterion for a successful self-model is not so much its fidelity to the physical robot, but rather its ability to predict sensory inputs under a repertoire of actions. This underlines that predictive models are recruited for the control of behaviour (as cybernetics assumes) and not to furnish general-purpose representations of the world or the body.

The study also provides a concrete example of how actions can be performed, not to achieve some externally specified goal, but to permit inference about the system's own physical instantiation. Bayesian or not, this implies active inference. Indeed, perhaps its most important contribution is that it highlights how active inference can prescribe *disruptive* or *disambiguating* actions that generate sensory prediction errors under competing hypotheses, and not just actions that seek to confirm sensory predictions. This recalls models of attention based on maximisation of Bayesian surprise (Itti & Baldi 2009), and is equivalent to hypothesis testing in science, where the best experiments are those concocted on the basis of being most likely to falsify a given hypothesis (disruptive) or distinguish between competing hypotheses (disambiguating). It also implies that agents encode predictions about the likely sensory consequences of a range of potential actions, allowing the selection of those actions likely to be the most disruptive or disambiguating. This concept of a *counterfactually-equipped predictive model* bring us nicely to our next topic: so-called *enactive* cognitive science and its relation to PP.

7 Evolutionary robotics involves the use of population-based search procedures (genetic algorithms) to automatically specify control architectures (and/or morphologies) of mobile robots. For an excellent introduction see (Bongard 2013).

8 Videos showing the evolution of both gait and self-model are available from http://creativemachines.cornell.edu/emergent_self_models

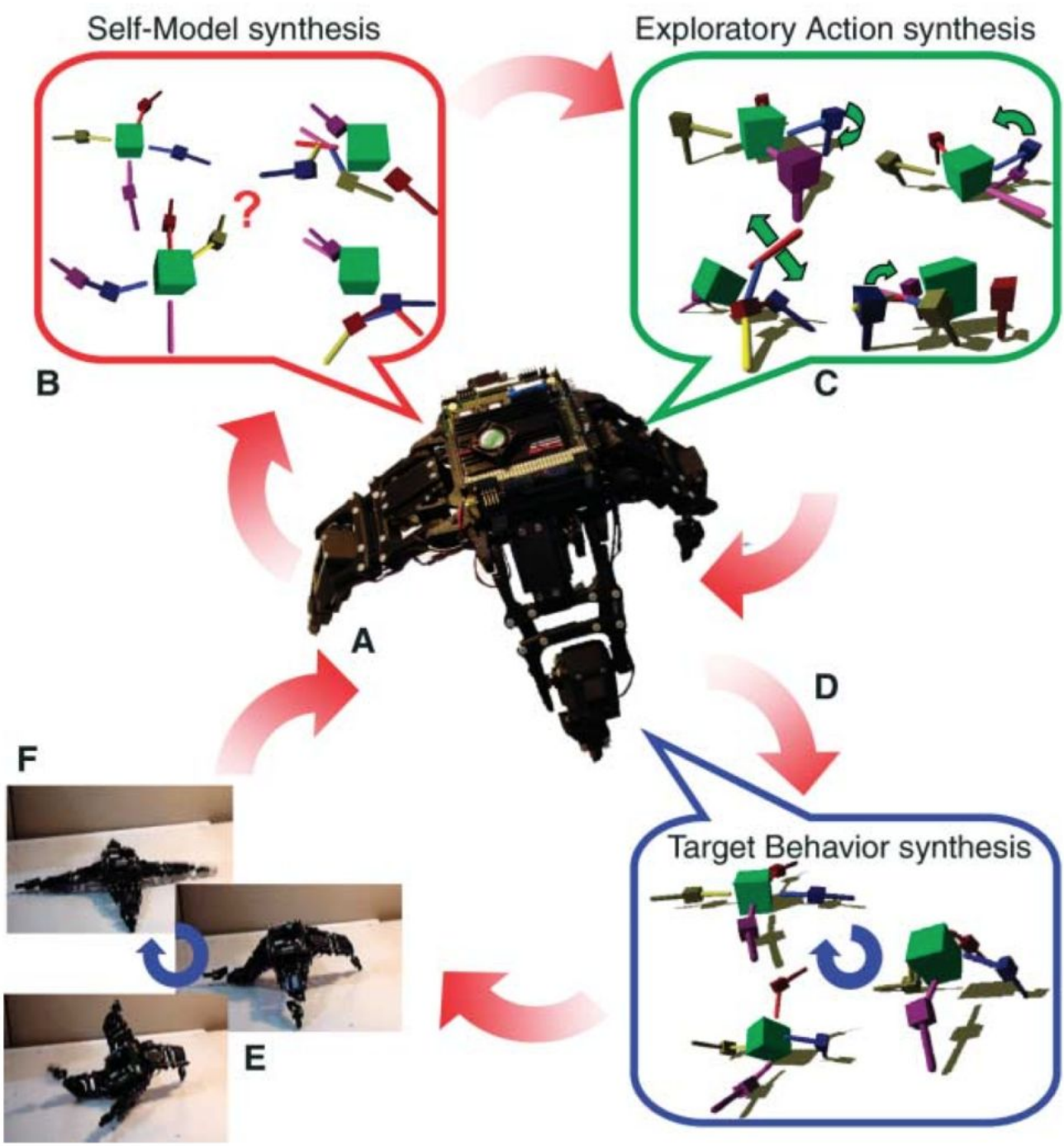

Figure 6: An evolutionary-robotics experiment demonstrating continuous self-modelling Bongard et al. (2006). See text for details. Reproduced with permission.

4 Predictive processing and enactive cognitive science

4.1 Enactive theories, weak and strong

The idea that the brain relies on internal representations or models of extra-cranial states of affairs has been treated with suspicion ever since the limitations of "good old fashioned artificial intelligence" became apparent (Brooks 1991). Many researchers of artificial intelligence have indeed returned to cybernetics as an alternative framework in which closely coupled feedback loops, leveraging invariants in brain-body-world interactions, obviate the need for detailed internal representations of external properties (Pfeifer & Scheier 1999). The evolutionary robotics methodology just described is

often coupled with simple dynamical neural networks in order to realize controllers that are tightly embodied and embedded in just this way (Beer 2003). Within cognitive science, such anti-representationalism is most vociferously defended by the movement variously known as "enactive" (Noë 2004), "embodied" (Gallese & Sinigaglia 2011), or "extended" (Clark & Chalmers 1998) cognitive science. Among these approaches, it is enactivism that is most explicitly anti-representationalist. While enactive theorists might agree that adaptive behaviour requires organisms and control structures that are systematically sensitive to statistical structures in their environment, most will deny that this sensitivity implies the existence and deployment of any "inner description" or model of these probabilistic patterns (Chemero 2009; Hutto & Myin 2013).

This tradition has weak and strong expressions. At the weak extreme is the truism that perception, cognition, and behaviour—and their underlying mechanisms—cannot be understood without a rich appreciation of the roles of the body, the environment, and the structured interactions that they support (Clark 1997; Varela et al. 1993). Weak enactivism is eminently compatible with PP, as seen especially with emerging versions of PP that stress embodiment through self-modelling and interoception, and which emphasize the importance of agent-environment coupling (embeddedness) through active inference. At the other extreme lie claims that explanations based on internal representations or models of any sort are fundamentally misguided, and that a new explicitly non-representational vocabulary is needed in order to make sense of the relations between brains, bodies, and the world (O'Regan et al. 2005). Strong enactivism is by definition incompatible with PP since it rejects the core concept of the internal model.

4.2 Sensorimotor contingency theory

A landmark in the strongly enactive approach is SMC (sensorimotor contingency) theory, which says that perception depends on the "practical mastery" of sensorimotor dependencies relevant to behaviour (O'Regan & Noë 2001). In brief, SMC theory claims that experience and perception are not things that are "generated" by the brain (or by anything else for that matter) but are, rather, "skills" consisting of fluid patterns of on-going interaction with the environment (O'Regan & Noë 2001). For instance, on SMC theory the conscious visual experience of redness is given by *the exercise of practical mastery of the laws governing how interactions with red things unfold* (these laws being the "SMC"s). The theory is not, however, limited to vision: the experiential quality of the softness of a sponge would be given by (practical mastery of) the laws governing its squishiness upon being pressed.

Two aspects of SMC theory deserve emphasis here. The first is that the concept of an SMC rightly underlines the close coupling of perception and action and the critical importance of ongoing agent-environment interaction in structuring perception, action, and behaviour. This is inherited from Gibsonian notions of perceptual affordance (Gibson 1979) and has certainly advanced our understanding of why different kinds of perceptual experience (vision, smell, touch, etc.) have different qualitative characters.

The second is that *mastery* of an SMC requires an essentially *counterfactual* knowledge of relations between particular actions and the resulting sensations. In vision, for instance, mastery entails an implicit knowledge of the ways in which moving our eyes and bodies would reveal additional sensory information about perceptual objects (O'Regan & Noë 2001). Here SMC theory has made an important contribution to our understanding of *perceptual presence*. Perceptual presence refers to the property whereby (in normal circumstances) perceptual contents appear as subjectively real, that is, as *existing*. For example, when viewing a tomato, we see it as real inasmuch as we seem to be perceptually aware of some of its parts (e.g., its back) that are not currently causally impacting our sensory surfaces. Looking at a picture of a tomato does not give rise to the same subjective impression of realness. But how can we be aware of parts of the tomato that, strictly speaking, we do not

see? SMC theory says the answer lies in our (implicit) mastery of SMCs, which relate potential actions to their likely sensory effects; and it is in this sense that we can be perceptually aware of parts of the tomato that we cannot actually see (Noë 2006).

SMC theory has often been set against naïve representationalist theories in cognitive science that propose such things as "pictures in the head" or that (like good-old-fashioned-AI) treat accurate representations of external properties as general-purpose goal states for cognition. This is all to the good. Yet by dispensing with implementation-level concepts such as predictive inference, it struggles with the important question of what exactly is going on in our heads during the exercise of mastery of a sensorimotor contingency. [9]

4.3 Predictive perception of sensorimotor contingencies

A powerful response is given by integrating SMC theory with PP, in the guise of PPSMC (Predictive Perception of SensoriMotor Contingencies; Seth 2014b). An extensive development of PPSMC is given elsewhere (see Seth 2014b plus commentaries and response). Here I summarize the main points. First, recall that under PP prediction errors can be minimized either by updating perceptual predictions or by performing actions, where actions are generated through the resolution of proprioceptive prediction errors. Also recall that PP is inherently hierarchical, so that at some hierarchical level predictive models will encode multimodal and even amodal expectations linking exteroceptive (sensory) and proprioceptive (motor) sensations. These models generate predictions about linked sequences of sensory and proprioceptive (and possibly interoceptive) inputs corresponding to specific actions, with predictions becoming increasingly modality-specific at lower hierarchical levels. These multi-level predictive models can therefore be understood as instantiating the implicit sub-personal knowledge of sensorimotor constructs underlying SMCs and their acquisition. Put simply, hierarchical active inference implies the existence of predictive models encoding information very much like that required by SMC theory.

The next step is to incorporate the notion of *mastery* of SMCs, which, as mentioned, implies an essentially counterfactual kind of implicit knowledge. The simple solution is to augment the predictive models that animate PP with counterfactual probability densities.[10] As introduced earlier (section 4.1), counterfactually-equipped predictive models encode not only the likely causes of current sensory input, but also the likely causes of fictive sensory inputs conditioned on possible but not executed actions. That is, they encode how sensory inputs (and their expected precisions) would change on the basis of a repertoire of possible actions (expressed as proprioceptive predictions), even if those actions are not performed. The counterfactual encoding of expected precision is important here, since it is on this basis that actions can be selected for their likelihood of minimizing the conditional uncertainty associated with a perceptual prediction. There is a mathematical basis for manipulating counterfactual beliefs of this kind, as shown in a recent model where counterfactual PP drives oculomotor control during visual search (Friston 2014; Friston et al. 2012).[11] Here the main point is that counterfactually-rich predictive models supply just what is needed by SMC theory: an answer to the question of what is going on inside our heads during the exercise of mastery of SMCs.

Counterfactual PP makes sense from several perspectives (Seth 2014b). As mentioned above, it provides a neurocognitive operationalisation of the notion of mastery of SMCs that is central to enactive cognitive science. In doing so it dissolves apparent tensions between enactive

9 At a recent symposium of the AISB society that focused on SMC theory, it was stated that "the main question is how to get the brain into view from an enactive/sensorimotor perspective. […] Addressing this question is urgently needed, for there seem to be no accepted alternatives to representational interpretations of the inner processes" (O'Regan & Dagenaar 2014).

10 See Beaton (2013) for a distinct approach to incorporating counterfactual ideas in SMC theory. Beaton's approach remains squarely within the strongly enactivist tradition.

11 There are also some challenges lying in wait here. For instance, it is not immediately clear how important assumptions like the Laplace approximation can generalize to the multimodal probability distributions entailed by counterfactual PP (Otworowska et al. 2014).

cognitive science and approaches grounded in the Bayesian brain, but only at the price of rejecting the strong enactivist's insistence that internal models or representations—of any sort—are unacceptable.[12] PPSMC also provides a solution to the challenge of accounting for perceptual presence within PP. The idea here is that perceptual presence corresponds to the *counterfactual richness* of predictive models. That is, perceptual contents enjoy presence to the extent that the corresponding predictive models encode a rich repertoire of counterfactual relations linking potential actions to their likely sensory consequences.[13] In other words, we experience normal perception as world-revealing precisely because the predictive models underlying perceptual content specify a rich repertoire of counterfactually explicit probability densities encoding the mastery of SMCs.

A good test of PPSMC is whether it can account for cases where normal perceptual presence is lacking. An important example is synaesthesia, where it is widely reported that synaesthetic "concurrents" (e.g., the inexistent colours sometimes perceived along with achromatic grapheme inducers) are not experienced as being part of the world (i.e., synaesthetes generally retain intact reality testing with respect to their concurrent experiences). PPSMC explains this by noticing that predictive models related to synaesthetic concurrents are counterfactually *poor*. The hidden (environmental) causes giving rise to concurrent-related sensory signals do not embed a rich and deep statistical structure for the brain to learn. In particular, there is very little sense in which synaesthetic concurrents depend on active sampling of their hidden causes. According to PPSMC, it is this comparative *counterfactual poverty* that explains why synaesthetic concurrents lack perceptual presence. SMC theory itself struggles to account for this phenomenon—not least because it struggles to account for synaesthesia in the first place (Gray 2003).

12 There is a more dramatic conflict with "radical" versions of enactivism, in which mental processes, and in some cases even their material substrates, are allowed to extend beyond the confines of the skull (Hutto & Myin 2013).

13 Presence may also depend on the hierarchical depth of predictive models inasmuch as this reflects object-related invariances in perception. For further discussion see commentaries and response to (Seth 2014b).

There are some challenges to thinking that perceptual presence uniquely depends on counterfactual richness. One might think that the more familiar one is with an object, the richer the repertoire of counterfactual relations that will be encoded. If so, the more familiar one is with an object, the more it should appear to be real. But *prima facie* it is not clear that familiarity and perceptual presence go hand-in-hand like this.[14] Also, some perceptual experiences (like the experience of a blue sky) can seem highly perceptually present despite engaging an apparently poor repertoire of counterfactual relations linking sensory signals to possible actions. An initial response is to consider that presence might depend not on counterfactual richness *per se*, but on a "normalized" richness based on higher-order expectations of counterfactual richness (which would be low for the blue sky, for instance). These considerations also point to potentially important distinctions between perceived *objecthood* and perceived *presence*, a proper treatment of which moves beyond the scope of the present paper.

5 Active inference

5.1 Counterfactual PP and active inference

Active inference has appeared repeatedly as an important concept throughout this paper. Yet it is more difficult to grasp than the basics of PP, which involve passive predictive inference. This is partly because several senses of active inference can be distinguished, which have not previously been fully elaborated.

In general, active inference can be harnessed to drive action, or to improve perceptual predictions. In the former case, actions emerge from the minimization of proprioceptive prediction errors through engaging classical reflex arcs (Friston et al. 2010). This implies the existence of generative models that predict time-varying flows of proprioceptive inputs (rather than just end-points), and also the transient reduction of expected precision of proprioceptive prediction

14 Thanks to my reviewers for raising this provocative point.

errors, corresponding to sensory attenuation (Brown et al. 2013).

In the latter case, actions are engaged in order to generate new sensory samples, with the aim of minimizing uncertainty in perceptual predictions. This can be achieved in several different ways, as is apparent by analogy with experimental design in scientific hypothesis testing. Actions can be selected that (i) are expected to *confirm* current perceptual hypotheses (Friston et al. 2012); (ii) are expected to *disconfirm* such hypotheses; or (iii) are expected to *disambiguate* between competing hypotheses (Bongard et al. 2006). A scientist may perform different experiments when attempting to find evidence against a current hypothesis than when trying to decide between different hypotheses. In just the same way, active inference may prescribe different sampling actions for these different objectives.

These distinctions underline that active inference *implies* counterfactual PP. In order for a brain to select those actions most likely to confirm, disconfirm, or decide between current predictive model(s), it is necessary to encode expected sensory inputs and precisions related to potential (but not executed) actions. This is evident in the example of oculomotor control described earlier (Friston et al. 2012). Here, saccades are guided on the basis of the expected precision of sensory prediction errors so as to minimize the uncertainty in current perceptual predictions. Note that this study retained the higher-order prior that only a single perceptual prediction exists at any one time, precluding active inference in its disambiguatory sense.

Several related ideas arise in connection with these new readings of active inference. Seeking disconfirmatory or disruptive evidence is closely related to maximizing Bayesian surprise (Itti & Baldi 2009). This also reminds us that the best statistical models are usually those that successfully account for the most variance with the fewest degrees of freedom (model parameters), not just those that result in low residual error *per se*. In addition, disambiguating competing hypotheses moves from Bayesian model selection and optimization to model comparison, where arbitration among competing models is mediated by trade-offs between accuracy and model complexity (Rosa et al. 2012).

The information-seeking (or "infotropic"[15]) role of active inference puts a different gloss on the free energy principle, which had been interpreted simply as minimization of prediction error. Rather, now the idea is that systems best ensure their long-run survival by inducing the *most predictive* model of the causes of sensory signals, and this requires disruptive and/or disambiguating active inference, in order to always put the current-best model to the test. This view helps dissolve worries about the so-called "dark room problem" (Friston et al. 2012), in which prediction error is minimized by predicting something simple (e.g., the absence of visual input) and then trivially confirming this prediction (e.g., by closing one's eyes).[16] Previous responses to this challenge have appealed to the idea of higher-order priors that are incompatible with trivial minimization of lower-level prediction errors: closing one's eyes (or staying put in a dark room) is not expected to lead to homeostatic integrity on average and over time (Friston et al. 2012; Hohwy 2013). It is perhaps more elegant to consider that disruptive and disambiguatory active inferences imply exploratory sampling actions, independent of any higher-order priors about the dynamics of sensory signals *per se*. Further work is needed to see how cost functions reflecting infotropic active inference can be explicitly incorporated into PP and the free energy principle.

5.2 Active interoceptive inference and counterfactual PP

What can be said about counterfactual PP and active inference when applied to *interoception*? Is there a sense in which predictive models underlying emotion and mood encode counterfactual associations linking fictive interoceptive signals (and their likely causes) to autonomic or allostatic controls? And if so, what phenomeno-

15 Chris Thornton came up with this term (personal communication).

16 The term "dark room problem" comes from the idea that a free-energy-minimizing (or surprise-avoiding) agent could minimize prediction error just by finding an environment that lacks sensory stimulation (a "dark room") and staying there.

logical dimensions of affective experience depend on these associations? While these remain open questions, we can at least sketch the territory.

We have seen that active inference in exteroception *implies* counterfactual processing, so that actions can be chosen according to their predicted effects in terms of (dis)confirming or disambiguating sensory predictions. The same argument applies to interoception. For active interoceptive inference to effectively disambiguate predictive models, or (dis)confirm interoceptive predictions, predictive models must be equipped with counterfactual associations relating to the likely effects of autonomic or (at higher hierarchical levels) allostatic controls. At least in this sense, interoceptive inference then also involves counterfactual expectations.

That said, there are likely to be substantial differences in how counterfactual active inference plays out in interoceptive settings. For instance, it may not be adaptive (in the long run) for organisms to continually attempt to disconfirm current interoceptive predictions, assuming these are compatible with homeostatic integrity. To put it colloquially, we do not want to drive our essential variables continually close to viability limits, just to check whether they are always capable of returning. This recalls our earlier point (section 4.1) that predictive control is more naturally applicable to interoception than exteroception, given the imperative of maintaining the homeostasis of essential variables. In addition, the causal structure of counterfactual associations encoded by interoceptive predictive models is undoubtedly very different than in cases like vision. These differences may speak to the substantial phenomenological differences in the kind of perceptual presence associated with these distinct conscious contents (Seth et al. 2011).

6 Conclusion

This paper has surveyed predictive processing (PP) from the unusual viewpoint of cybernetic origins in active homeostatic control (Ashby 1952; Conant & Ashby 1970). This shifts the perspective from perceptual inference as furnishing representations of the external world for the consumption of general-purpose cognitive mechanisms, towards model-based predictive control as a primary survival imperative from which perception, action, and cognition ensue. This view is aligned with the free energy principle (Friston 2010); however it attempts to account for specific cognitive and phenomenological properties, rather than for adaptive systems in general. Several implications follow from these considerations. Emotion becomes a process of active interoceptive inference (Seth 2013)—a process that also recruits autonomic regulation and influences intuitive decision-making through behavioural allostasis. A common predictive principle underlying interoception and exteroception also provides an integrative view of the neurocognitive mechanisms underlying embodied selfhood, in particular the experience of body ownership (Apps & Tsakiris 2014; Limanowski & Blankenburg 2013; Suzuki et al. 2013). In this view, the experience of embodied selfhood is specified by the brain's "best guess" of those signals most likely to be "me" across exteroceptive and interoceptive domains. From the perspective of cybernetics the embodied self is both that which needs to be homeostatically maintained and also the medium through which allostatic interactions are expressed.

A second influential line deriving from cybernetics sets PP within the broader context of model-based versus enactivist perspectives on cognitive science. On one hand, cybernetics has been cited in support of non-representational cognitive science in virtue of its showing how simple mechanisms can give rise to complex and apparently goal-directed behaviour by capitalizing on agent-environment interactions, mediated by the body (Pfeifer & Scheier 1999). On the other, the cybernetic legacy shows how PP can put mechanistic flesh on the philosophical bones of enactivism, but only by embracing a finessed form of representationalism (Seth 2014b). A key concept within enactive cognitive science is that of mastery of sensorimotor contingencies (SMCs). This concept is useful for understanding the qualitative character of distinct perceptual modalities, yet as expressed within enactivism it lacks a firm implementation basis. "Pre-

dictive Perception of SensoriMotor Contingencies" (PPSMC) addresses this challenge by proposing that SMCs are implemented by predictive models of sensorimotor relations, underpinned by the continuity between perception and action entailed by active inference. *Mastery* of sensorimotor contingencies depends on predictive models of counterfactual probability densities that specify the likely causes of sensory signals that *would* occur *were* specific actions taken. By relating PP to key concepts in enactivism, this theory is able to account for phenomenological features well treated by the latter, such as the experience of perceptual presence (and its absence in cases like synaesthesia).

Considering these issues leads to distinct readings of active inference, which at its most general implies the selective sampling of sensory signals to minimize uncertainty about perceptual predictions. At a finer grain, active inference can involve performing actions to confirm current predictions, to disconfirm current predictions, or to disambiguate competing predictions. These different senses rest on the concept of counterfactually-equipped predictive models; and they generalize the free energy principle to include Bayesian-model comparison as well as optimization and inference.

In summary, the ideas outlined in this paper provide a distinctive integration of predictive processing, cybernetics, and enactivism. This rich blend dissolves apparent tensions between internalist and enactivist (model-based and model-free) views on the neural mechanisms underlying perception, cognition, and action; it elaborates common predictive mechanisms underlying perception and control of self and world; it provides a new view of emotion as active interoceptive inference, and it shows how "counterfactual" predictive processing can account for the phenomenology of conscious presence and its absence in specific situations. It also finesses the concept of active inference to engage distinct forms of hypothesis testing that prescribe different sampling actions (one bonus is that the "dark room problem" is elegantly solved). At the same time, new and difficult challenges arise in validating these ideas experimentally and in distinguishing them from alternative explanations that do not rely on internally-realised inferential mechanisms.

Acknowledgements

I am grateful to the Dr. Mortimer and Theresa Sackler Foundation, which supports the work of the Sackler Centre for Consciousness Science. This work was also supported by ERC FP7 grant CEEDs (FP7-ICT-2009-5, 258749). Many thanks to Thomas Metzinger and Jennifer Windt for inviting me to make this contribution, and for the insightful and helpful reviewer comments they solicited. I'm also grateful to Kevin O'Regan and Jan Dagenaar for inviting me to speak at a symposium entitled "Consciousness without inner models?" (London, April 2014), which provided a feisty forum for debating some of the ideas presented here.

References

Adams, R. A., Shipp, S. & Friston, K. J. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, *218* (3), 611-643. 10.1007/s00429-012-0475-5

Apps, M. A. & Tsakiris, M. (2014). The free-energy self: A predictive coding account of self-recognition. *Neuroscience and Biobehavioral Reviews*, *41*, 85-97. 10.1016/j.neubiorev.2013.01.029

Ashby, W. R. (1952). *Design for a brain.* London, UK: Chapman and Hall.

——— (1956). *An introduction to cybernetics.* London, UK: Chapman and Hall.

Aspell, J. E., Heydrich, L., Marillier, G., Lavanchy, T., Herbelin, B. & Blanke, O. (2013). Turning the body and self inside out: Visualized heartbeats alter bodily self-consciousness and tactile perception. *Psychological Science*, *24* (12), 2445-2453. 10.1177/0956797613498395

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P. & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76* (4), 695-711. 10.1016/j.neuron.2012.10.038

Beaton, M. (2013). Phenomenology and embodied action. *Constructivist Foundations*, *8* (3), 298-313.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, *11* (4), 209-243. 10.1177/1059712303114001

Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, *13* (1), 7-13. 10.1016/j.tics.2008.10.003

Bongard, J. (2013). Evolutionary robotics. *Communications of the ACM*, *56* (8), 74-85. 10.1145/2493883

Bongard, J., Zykov, V. & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, *314* (5802), 1118-1121. 10.1126/science.1133687

Botvinick, M. & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, *391* (6669), 756-756. 10.1038/35784

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology.* Cambridge, MA: MIT Press.

Brooks, R. A. (1991). Intelligence without reason. In J. Mylopoulos & R. Reiter (Eds.) *Proceedings of the 12th international joint conference on artificial intelligence - volume 1* (pp. 569-595). San Francisco, CA: Morgan Kaufmann Publishers.

Brown, H., Adams, R. A., Parees, I., Edwards, M. & Friston, K. J. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, *14* (4), 411-427. 10.1007/s10339-013-0571-3

Chemero, A. (2009). *Radical embodied cognitive science.* Cambridge, MA: MIT Press.

Clark, A. (1997). *Being there. Putting brain, body, and world together again.* Cambridge, MA: MIT Press.

——— (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavior and Brain Sciences*, *36* (3), 181-204. 10.1017/S0140525X12000477

——— (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-21). Frankfurt a. M., GER: MIND Group.

Clark, A. & Chalmers, D. J. (1998). The extended mind. *Analysis*, *58* (1), 7-19. 10.1093/analys/58.1.7

Conant, R. & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1* (2), 89-97.

Craig, A. D. (2003). Interoception: The sense of the physiological condition of the body. *Current Opinion in Neurobiology*, *13* (4), 500-505. 10.1016/S0959

——— (2009). How do you feel now? The anterior insula and human awareness. *Nature Reviews Neuroscience*, *10* (1), 59-70. 10.1038/nrn2555

Critchley, H. D., Wiens, S., Rotshtein, P., Ohman, A. & Dolan, R. J. (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience*, *7* (2), 189-195. 10.1038/nn1176

Critchley, H. D. & Seth, A. K. (2012). Will studies of macaque insula reveal the neural mechanisms of self-awareness? *Neuron*, *74* (3), 423-426. 10.1016/j.neuron.2012.04.012

Damasio, A. (1994). *Descartes' error.* London, UK: Mac Millan.

Dunn, B. D., Galton, H. C., Morgan, R., Evans, D., Oliver, C., Meyer, M. & Dalgleish, T. (2010). Listening to your heart. How interoception shapes emotion experience and intuitive decision making. *Psychological Science*, *21* (12), 1835-1844. 10.1177/0956797610389191

Dupuy, J.-P. (2009). *On the origins of cognitive science: The mechanization of mind.* Cambridge, MA: MIT Press.

Evrard, H. C., Forro, T. & Logothetis, N. K. (2012). Von economo neurons in the anterior insula of the macaque monkey. *Neuron*, *74* (3), 482-489. 10.1016/j.neuron.2012.03.003

Ferri, F., Chiarelli, A. M., Merla, A., Gallese, V. & Costantini, M. (2013). The body beyond the body: Expectation of a sensory event is enough to induce ownership over a fake hand. *Proceedings of the Royal Society B: Biological Sciences*, *280* (1765), 20131140-20131140. 10.1098/rspb.2013.1140

Fletcher, P. C. & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10* (1), 48-58. 10.1038/nrn2536

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360* (1456), 815-836. 10.1098/rstb.2005.1622

——— (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, *13* (7), 293-301. 10.1016/j.tics.2009.04.005.

——— (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11* (2), 127-138. 10.1038/nrn2787

——— (2014). Active inference and agency. *Cognitive Neuroscience*, *5* (2), 119-121. 10.1080/17588928.2014.905517

Friston, K. J., Kilner, J. & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology - Paris*, *100* (1-3), 70-87. 10.1016/j.jphysparis.2006.10.001

Friston, K. J., Daunizeau, J., Kilner, J. & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biological Cybernetics*, *102* (3), 227-260. 10.1007/s00422-010-0364-z

Friston, K. J., Adams, R. A., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, *3* (151), 1-20. 10.3389/fpsyg.2012.00151

Friston, K. J., Thornton, C. & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, *3* (130), 1-7. 10.3389/fpsyg.2012.00130

Gallese, V. & Sinigaglia, C. (2011). What is so special about embodied simulation? *Trends in Cognitive Sciences*, *15* (11), 512-519. 10.1016/j.tics.2011.09.003

Gibson, J. J. (1979). *The ecological approach to visual perception.* Hillsdale, NJ: Lawrence Erlbaum.

Gray, J. A. (2003). How are qualia coupled to functions? *Trends in Cognitive Sciences*, *7* (5), 192-194. 10.1016/S1364-6613(03)00077-9

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *290* (1038), 181-197. 10.1098/rstb.1980.0090

Gu, X., Hof, P. R., Friston, K. J. & Fan, J. (2013). Anterior insular cortex and emotional awareness. *Journal of Comparative Neurology*, *521* (15), 3371-3388. 10.1002/cne.23368

Gu, X. & Fitzgerald, T. H. (2014). Interoceptive inference: Homeostasis and decision-making. *Trends in Cognitive Sciences*, *18* (6), 269-270. 10.1016/j.tics.2014.02.001

Hinton, G. E. & Dayan, P. (1996). Varieties of Helmholtz Machine. *Neural Networks*, *9* (8), 1385-1403. 10.1016/S0893

Hohwy, J. (2013). *The predictive mind.* Oxford, UK: Oxford University Press.

——— (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-22). Frankfurt a.M., GER: MIND Group.

Hutto, D. & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content.* Cambridge, MA: MIT Press.

Itti, L. & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49* (10), 1295-1306. 10.1016/j.visres.2008.09.007

James, W. (1894). The physical basis of emotion. *Psychological Review*, *1*, 516-529.

Knill, D. C. & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27* (12), 712-719. 10.1016/j.tins.2004.10.007

Lee, T. S. & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A,Optics, image science and vision*, *20* (7), 1434-1448. 10.1364/JOSAA.20.001434

Limanowski, J. & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neurosciences*, *7* (547), 1-20. 10.3389/fnhum.2013.00547

Makin, T. R., Holmes, N. P. & Ehrsson, H. H. (2008). On the other hand: Dummy hands and peripersonal space. *Behavioural Brain Research*, *191* (1), 1-10. 10.1016/j.bbr.2008.02.041

Metzinger, T. (2003). *Being no one.* Cambridge, MA: MIT Press.

Moseley, G. L., Olthof, N., Venema, A., Don, S., Wijers, M., Gallace, A. & Spence, C. (2008). Psychologically induced cooling of a specific body part caused by the illusory ownership of an artificial counterpart. *Proceedings of the National Academy of Sciences of the United States of America*, *105* (35), 13169-13173. 10.1073/pnas.0803768105

Neal, R. M. & Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan (Ed.) *Learning in Graphical Models* (pp. 355-368). Dordrecht, NL: Kluwer Academic Publishers.

Noë, A. (2004). *Action in perception.* Cambridge, MA: MIT Press.

——— (2006). *Experience without the head.* Clarendon, NY: Oxford University Press.

O'Regan, J. K. & Dagenaar, J. (2014). Consciousness without inner models: A sensorimotor account of what is going on in our heads. *Proceedings of the AISB.* http://doc.gold.ac.uk/aisb50/

O'Regan, J. K., Myin, E. & Noë, A. (2005). Skill, corporality and alerting capacity in an account of sensory consciousness. *Progress in Brain Research*, *150*, 55-68. 10.1016/S0079-6123(05)50005-0

O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, *24* (5), 939-1031.

Otworowska, M., Kwisthout, J. & van Rooj, I. (2014). Counterfactual mathematics of counterfactual predictive models. *Frontiers in psychology: Consciousness Research*, *5* (801), 1-2. 10.3389/fpsyg.2014.00801

Paulus, M. P. & Stein, M. B. (2006). An insular view of anxiety. *Biological psychiatry*, *60* (4), 383-387. 10.1016/j.biopsych.2006.03.042

Pfeifer, R. & Scheier, C. (1999). *Understanding intelligence.* Cambridge, MA: MIT Press.

Pickering, A. (2010). *The cybernetic brain: Sketches of another future.* Chicago, IL: University of Chicago Press.

Ploghaus, A., Tracey, I., Gati, J. S., Clare, S., Menon, R. S., Matthews, P. M. & Rawlins, J. N. (1999). Dissociating pain from its anticipation in the human brain. *Science*, *284* (5422), 1979-1981. 10.1126/science.284.5422.1979

Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. (2013). Probabilistic brains: Knowns and unknowns. *Nature Neuroscience*, *16* (9), 1170-1178. 10.1038/nn.3495

Quattrocki, E. & Friston, K. (2014). Autism, oxytocin and interoception. *Neuroscience and Biobehavioral Reviews*, *47C*, 410-430. 10.1016/j.neubiorev.2014.09.012

Rao, R. P. & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2* (1), 79-87. 10.1038/4580

Rosa, M. J., Friston, K. J. & Penny, W. (2012). Post-hoc selection of dynamic causal models. *Journal of Neuroscience Methods*, *208* (1), 66-78. 10.1016/j.jneumeth.2012.04.013

Salomon, R., Lim, M., Pfeiffer, C., Gassert, R. & Blanke, O. (2013). Full body illusion is associated with widespread skin temperature reduction. *Frontiers in Behavioral Neuroscience*, *7* (65), 1-11. 10.3389/fnbeh.2013.00065

Schachter, S. & Singer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review*, *69*, 379-399. 10.1037/h0046234

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17* (11), 565-573. 10.1016/j.tics.2013.09.007

——— (2014a). Interoceptive inference: From decision-making to organism integrity. *Trends in Cognitive Sciences*, *18* (6), 270-271. 10.1016/j.tics.2014.03.006

——— (2014b). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synaesthesia. *Cognitive Neuroscience*, *5* (2), 97-118. 10.1080/17588928.2013.877880.

Seth, A. K. & Critchley, H. D. (2013). Interoceptive predictive coding: A new view of emotion? *Behavioral and Brain Sciences*, *36* (3), 227-228.

Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, *2* (395), 1-16. 10.3389/fpsyg.2011.00395

Singer, T., Critchley, H. D. & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences*, *13* (8), 334-340. 10.1016/j.tics.2009.05.001

Sokol-Hessner, P., Hartley, C. A., Hamilton, J. R. & Phelps, E. A. (2014). Interoceptive ability predicts aversion to losses. *Cognition and Emotion*, 1-7. 10.1080/02699931.2014.925426

Suzuki, K., Garfinkel, S. N., Critchley, H. D. & Seth, A. K. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia*, *51* (13), 2909-2917. 10.1016/j.neuropsychologia.2013.08.014

Thompson, E. & Varela, F. J. (2001). Radical embodiment: Neural dynamics and consciousness. *Trends in Cognitive Sciences*, *5* (10), 418-425. 10.1016/S1364-6613(00)01750-2

Tsakiris, M., Tajadura-Jimenez, A. & Costantini, M. (2011). Just a heartbeat away from one's body: Interoceptive sensitivity predicts malleability of body-representations. *Proceedings. Biological sciences / The Royal Society*, *278* (1717), 2470-2476. 10.1098/rspb.2010.2547

Ueda, K., Okamoto, Y., Okada, G., Yamashita, H., Hori, T. & Yamawaki, S. (2003). Brain activity during expectancy of emotional stimuli: An fMRI study. *NeuroReport*, *14* (1), 51-55. 10.1097/01.wnr.0000050712.17082.1c

Van Gelder, T. (1995). What might cognition be if not computation? *Journal of Philosophy*, *92* (7), 345-381. 10.2307/2941061

Varela, F., Thompson, E. & Rosch, E. (1993). *The embodied mind: Cognitive science and human experience.* Cambridge, MA: MIT Press.

Verschure, P. F., Voegtlin, T. & Douglas, R. J. (2003). Environmentally mediated synergy between perception and behaviour in mobile robots. *Nature*, *425* (6958), 620-624. 10.1038/nature02024nature02024

Wolpert, D. M. & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3 Suppl*, 1212-1217. 10.1038/81497

Perceptual Presence in the Kuhnian-Popperian Bayesian Brain

A Commentary on Anil K. Seth

Wanja Wiese

Anil Seth's target paper connects the framework of PP (predictive processing) and the FEP (free-energy principle) to cybernetic principles. Exploiting an analogy to theory of science, Seth draws a distinction between three types of active inference. The first type involves confirmatory hypothesis-testing. The other types involve seeking disconfirming and disambiguating evidence, respectively. Furthermore, Seth applies PP to various fascinating phenomena, including perceptual presence. In this commentary, I explore how far we can take the analogy between explanation in perception and explanation in science.

In the first part, I draw a slightly broader analogy between PP and concepts in theory of science, by asking whether the Bayesian brain is Kuhnian or Popperian. While many aspects of PP are in line with Karl Popper's falsificationism, other aspects of PP conform to how Thomas Kuhn described scientific revolutions. Thus, there is both a sense in which the Bayesian brain is Kuhnian, and a sense in which it is Popperian. The upshot of these considerations is that falsification in PP can take many different forms. In particular, active inference can be used to falsify a model in more ways than identified by Seth.

In the second part of this commentary, I focus on Seth's PPSMCT (predictive processing account of sensorimotor contingency theory) and its application to perceptual presence, which assigns a crucial role to counterfactual richness. In my discussion, I question the significance of counterfactual richness for perceptual presence. First, I highlight an ambiguity inherent in Seth's descriptions of the target phenomenon (perceptual presence vs. objecthood). Then I suggest that counterfactual richness may not be the crucial underlying feature (of either perceptual presence or objecthood). Giving a series of examples, I argue that the degree of *represented causal integration* is an equally good candidate for accounting for perceptual presence (or objecthood), although more work needs to be done.

Keywords
Active inference | Binocular rivalry | Counterfactual richness | Cybernetics | Demarcation problem | Falsification | Free-energy principle | Naïve falsificationism | Objecthood | Paradigm change | Perceptual presence | Predictive processing | Rubber hand illusion | Scientific progress | Sensorimotor contingencies | Sophisticated falsificationism

Commentator

Wanja Wiese
wawiese@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Anil K. Seth
a.k.seth@sussex.ac.uk
University of Sussex
Brighton, United Kingdom

Editors

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

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

One of the relevant aspects of Seth's discussion is the way in which it highlights interesting links to theoretical precursors of PP. In doing so, he broadens the historical context in which the framework is usually situated. However, these considerations are not just relevant for the history of science, they also constitute a theoretical underpinning of several ways in which Seth has recently developed PP accounts of various phenomena. Due to limited space, I can only address some of these here. In particular, I will focus on his three interpretations of active

inference, and on his PP account of perceptual presence. In so doing, I will also try to take the analogy between explanation in perception and explanation in science a little further than it has previously been taken.

In section 2, I will briefly summarize Seth's view on the connection between cybernetics and the free-energy principle. One of the results of his considerations is that a distinction can be drawn between three types of active inference. The first type involves confirmatory hypothesis-testing. The other types involve seeking disconfirming and disambiguating evidence, respectively. Seth does not say much about what it takes to disconfirm or falsify a hypothesis or model. Furthermore, he seems to suggest that not all types of active inference he distinguishes are currently part of PP (at least in the version described by Karl Friston's FEP): "[t]hese points represent significant developments of the basic infrastructure of PP" (Seth 2014, p. 3).[1] In section 3, I will provide clarification of the notion of falsification by referring to the works of Karl Popper, Imre Lakatos, and Thomas Kuhn. I will also provide examples to show that different types of falsification are part and parcel of PP, not extensions of the basic infrastructure. In section 4, I point out an ambiguity in Seth's account of perceptual presence (perceptual presence vs. objecthood). After this, I suggest that counterfactual richness may not be the crucial underlying feature (of either perceptual presence or objecthood). Giving a series of examples, I argue that the degree of *represented causal integration* is an equally good candidate for accounting for perceptual presence (or objecthood), although more work needs to be done.

2 Cybernetics and the free-energy principle

In his very rich target paper, Anil Seth calls attention to one of the less well-considered precursors of PP: cybernetics. A central concept of cybernetics is the notion of homeostasis, which denotes an equilibrium of the system's parameters. This equilibrium is maintained by keeping the system's essential variables, like levels of blood oxygenation or blood sugar (cf. Seth this collection, p. 7), within a certain range (cf. ibid. pp. 7-8.). The process of achieving homeostasis is called allostasis (cf. ibid. p. 8). Cybernetic systems are teleological, i.e., goal-directed, because they are always trying to reach and preserve homeostasis. This suggests that control is more important than perception (cf. ibid. p. 9), and, as Seth emphasizes, it prioritizes interoceptive control over exteroceptive control: the main goal is to control the system's essential variables; interaction with the world is only necessary to the extent that it affects these variables (ibid. pp. 9-10.).

The principles of cybernetics fit astonishingly well to ideas motivating Karl Friston's FEP (which can, in some respects, be seen as a generalization of predictive processing).[2] The fundamental assumption behind this principle is that biological systems seek to "maintain their states and form in the face of a constantly changing environment" (Friston 2010, p. 127). This is obviously similar to the goal of achieving homeostasis.[3] Another focus of FEP is active inference, because action can reduce the surprisal of the agent's states (which is necessary to "resist a tendency to disorder", Friston 2009, p. 293); perceptual inference can only reduce the free-energy bound on surprise (Friston 2009, p. 294). This is in stark contrast with the Helmholtzian roots of PP, according to which action is primarily in the service of perception:

> [...] wir beobachten unter fortdauernder eigener Thätigkeit, und gelangen dadurch zur Kenntniss des Bestehens eines gesetzlichen Verhältnisses zwischen unseren Innervationen und dem Präsentwerden der verschiedenen Eindrücke aus dem Kreise

1 Unless stated otherwise, all page numbers refer to the target paper by Anil Seth.

2 It is more general, because predictive processing only plays a role in it if combined with the Laplace approximation (which entails, roughly, that probability distributions are approximated by Gaussian distributions). This approximation, however, also turns FEP into a more specific version, by assuming that the brain codes probability distribution as Gaussian distributions (which is not entailed by the general predictive processing framework discussed in Clark 2013, for instance).

3 In fact, the free-energy principle seems to be partly inspired by cybernetic ideas. Friston (2010, p. 127), for instance, cites Ashby (1947) when explaining the motivation for FEP.

> der zeitweiligen Präsentabilien. Jede unserer willkührlichen Bewegungen, durch die wir die Erscheinungsweise der Objecte abändern, ist als ein Experiment zu betrachten, durch welches wir prüfen, ob wir das gesetzliche Verhalten der vorliegenden Erscheinung, d.h. ihr vorausgesetztes Bestehen in bestimmter Raumordnung, richtig aufgefasst haben.[4] (Helmholtz 1959, p. 39)

According to this view, the main target of action is to find confirmatory evidence for internally-generated hypotheses. In short, the contrast between these two views can be described as "action as hypothesis-testing" versus "action as predictive control". Whereas the first seems to fit best to the Helmholtzian roots of PP (and puts action in the service of perception), the second seems to fit better to its cybernetic origins. Most notably, the free-energy principle combines both aspects, but assigns a pivotal role to action (perceptual inference only makes the free-energy bound on surprise tight, active inference leads to a further reduction of free energy, reducing surprise implicitly).

Seth compares model selection and optimization in evolutionary robotics to how these processes are implemented in active inference (pp. 14-15.). He cites the famous starfish robot developed by Josh Bongard, Victor Zykov, & Hod Lipson (2006) as an example. In a first phase, the robot generates multiple competing models of its own morphology and performs actions for which these models predict different sensory feedback. By comparing these predictions to the actual feedback, the starfish can thus exclude some of the possible models. When the robot has eliminated all but one model, a second phase starts and it uses this model to control its body and generate walking behavior (action as predictive control). Crucially, when the robot's morphology changes (when an experimenter removes one of its limbs), it can switch back to the first phase, re-creating competing models and using action to eliminate most of them (action as hypothesis-testing).

Seth points out that the second phase, in which the robot walks around, suggests that the main purpose of predictive models is to control behavior effectively, regardless of how accurately it represents the world or the body (p. 15). In the first phase, by contrast, exploratory actions are conducted in order to learn something about the body, not to reach a goal involving its environment (ibid.). As noted above, such instances of action conform more to Helmholtzian than to cybernetic roots (action as hypothesis-testing).

What this shows is that action can fulfill different purposes—not just theoretically, but also in real applications. The robot starfish uses action in at least two ways. Drawing on the often-noted analogy between PP and scientific practice (cf. Gregory 1980), Seth explores further purposes of action. This leads to a distinction between three types of active inference (pp. 18f.). The first involves active sampling to confirm predictions derived from currently active models; the second is employed to seek evidence that would disconfirm currently held hypotheses; the third involves sampling in order to disambiguate between alternative hypotheses (p. 19).

Crucially, Seth does not elaborate much on the notion of falsification or disconfirmation. He relates disconfirmation to Bayesian surprise (which formalizes the extent to which new evidence leads to a revision of prior representations, cf. Baldi & Itti 2010). Accordingly, he characterizes seeking falsifying evidence in terms of maximizing Bayesian surprise. However, the paper quoted in this context, Itti & Baldi (2009) only investigates the hypothesis that surprising information attracts attention, not that subjects act to maximize surprise. Friston et al. (2012, p. 6) clarify the relation between FEP and maximization of Bayesian surprise:

> The term Bayesian surprise can be a bit confusing because minimizing surprise per se (or maximizing model evidence) in-

4 "[...] we observe under constant own activity, and thereby achieve knowledge of the existence of a lawful relation between our innervations and the presence of different impressions of temporary presentations [Präsentabilien]. All of our willful movements through which we change the appearance of things should be considered an experiment, through which we test whether we have grasped correctly the lawful behavior of the appearance at hand, i.e. its supposed existence in determinate spatial structures." (My translation)

> volves keeping Bayesian surprise (complexity) as small as possible. This paradox can be resolved here by noting that agents expect Bayesian surprise to be maximized and then acting to minimize their surprise, given what they expect.

In the following section, I will clarify the notion of falsification, and discuss the ways in which it is used in PP. More specifically, I will illustrate various types of active inference by drawing a slightly broader analogy with theory of science. In particular, I will consider views put forward by Karl Popper and Thomas Kuhn, respectively. This will serve to help us get a handle on the general merits of confirmation and disconfirmation. Furthermore, both Popper's falsificationism and Kuhn's paradigm change can be related to aspects of predictive processing, which will hopefully lead to a better understanding of hypothesis-testing in PP. As a consequence, I invite Seth to provide a refined treatment of the relation between falsification and active inference.

3 Is the Bayesian brain Kuhnian or Popperian?[5]

The free-energy principle subsumes the Bayesian brain hypothesis[6] (cf. Friston 2009, p. 294). According to this view, processing in the brain can usefully be described as Bayesian inference. This means that the brain implements a probabilistic model that is updated in light of sensory signals using Bayes' theorem. More specifically, the brain combines prior knowledge about hidden causes in the world with a measurement of likelihood describing how probable the observed (sensory) evidence is, given various possible hidden causes. The result is a distribution (posterior) that describes how probable various possible causes are, given the obtained evidence. The process of determining the posterior is often called *model inversion*. In FEP, this type of inference is approximated using variational Bayes, which establishes the connection to predictive processing (cf. footnote 2 above). FEP can thus either be seen as a particular instance of the Bayesian brain hypothesis, or as a generalization.

As mentioned above, it is often pointed out that perceptions in PP are analogous to scientific hypotheses. The Bayesian brain is thus a hypothesis-testing brain (this analogy is also referred to in titles of papers by Jakob Hohwy, see Hohwy 2010, 2012). Thanks to active inference, the Bayesian brain performs an active kind of hypothesis testing. The three types of active inference distinguished by Seth assign a role to both confirmation and disconfirmation (falsification). This dual role of active inference is also emphasized by (Friston et al. 2012, p. 19):

> The resulting active or embodied inference means that not only can we regard perception as hypotheses, but we could regard action as performing experiments that confirm or disconfirm those hypotheses.

Further exploration of the analogy to theory of science reveals a puzzle: as we will see, doubts can be raised regarding the idea that a theory gains merit when it is confirmed (or even regarding the very notion of theory confirmation). Does this mean that the Bayesian brain generates hypotheses in an unscientific way?

3.1 The Popperian Bayesian brain

3.1.1 Conceptual clarification: From naïve to sophisticated falsificationism

According to Popper, science advances mainly by seeking falsifying evidence. In fact, falsifiability is Popper's proposed solution to the demarcation problem, i.e., the problem of specifying the difference between science and pseudo-science. Scientific theories posit universal propositions (scientific laws) that can never be proven in a strict sense, because only finite observations can be made. The next observation could, in principle, always disconfirm a universal em-

5 It should be noted that Popper rejected interpretations of confirmation (or corroboration) in terms of probabilities (cf. Popper 2005[1934], ch. X), as well as Bayesian interpretations of probability theory (cf. Popper 2005[1934], ch. *XVII). Here, I only suggest that a useful analogy between Popper's theory of science and the Bayesian brain can be drawn.

6 Seth identifies PP and the Bayesian brain (cf. p. 1). I follow suit in this commentary.

pirical hypothesis. Hence, being verifiable cannot be a criterion for being scientific, because theories cannot be empirically verified (cf. Popper 2005[1934], pp. 16-17.). Conversely, it is possible to *falsify* a universal statement using a single empirical proposition:

> Diese Überlegungen legen den Gedanken nahe, als Abgrenzungskriterium nicht die Verifizierbarkeit, sondern die *Falsifizierbarkeit* des Systems vorzuschlagen; [...] *Ein empirisch-wissenschaftliches System muß an der Erfahrung scheitern können.* (Popper 2005[1934], p. 17)[7]

Scientific theories thus cannot, according to Popper, be verified, but only falsified. However, when attempts to falsify a hypothesis have failed, we can say that the theory has been *corroborated*—which still means that the theory could be falsified in the future (cf. Popper 2005[1934], ch. X).

How can we apply these ideas to predictive processing? First, we have to find an analogon to scientific theories. I suggest that models can be treated analogously to theories, because in PP, predictions or hypotheses are derived from models and then compared to bottom-up signals. This also fits the way in which Seth describes the starfish example (namely in terms of model selection). What does it mean that a model is falsified in PP?

The question is not a trivial one, as there seems to be a crucial disanalogy between hypothesis-testing in Popper's sense and hypothesis-testing in the Bayesian brain. The reason why scientific theories are falsifiable is that they allow deriving hypotheses deductively. This means if a hypothesis is falsified, the theory is falsified as well. By contrast, hypotheses in the Bayesian brain are not deductively entailed by the models from which they are derived: the relation between model and hypothesis is *probabilistic* (the hypothesis is more or less probable, given the model). Hence, when a hypothesis or prediction elicits a large prediction error, this does not falsify the model; rather, it calls for an update to the effect that the model becomes less likely. Furthermore, according to Popper, it does not make sense to say that such hypotheses are corroborated to a greater or lesser extent. For being corroborated means that attempts at falsification have failed. But if it is in principle impossible to falsify a hypothesis, then saying that it has been corroborated becomes empty—worse, such hypotheses are not even scientific hypotheses (cf. Popper 2005[1934], pp. 248-249.). This, then, constitutes the puzzle mentioned above: if hypotheses in PP are not falsifiable, does this mean the Bayesian brain is unscientific?

This conclusion—that no useful analogy to Popper's theory of science can be drawn—rests on a naïve understanding of falsification (as emphasized by Imre Lakatos, cf. Lakatos 1970).[8] A closer look at the notion of falsification reveals that the analogy can be upheld. Furthermore, it helps us gain a better grasp of the notion of falsification in the context of PP.

First of all, we can note that in actual scientific practice, it is not the case that scientists attempt to falsify an isolated, single hypothesis —and then try to come up with a new theory when the hypothesis has been falsified. Rather, scientists often operate with different versions of a theory at the same time, or seek to find the best parameters for a model. The outcomes of an empirical study are then used to eliminate some of the different theories or parameter ranges. This has already been acknowledged by Popper (cf. 2005[1934], p. 63., fn. 10). As Thomas Nickles puts it:

> According to Popper, at any time there may be several competing theories being proposed and subsequently refuted by failed empirical tests—rather like balloons being launched and then shot down, one by one. (2014)

The result of this falsification procedure is that some of the competing theories are eliminated. This can already be seen as a slight departure

7 "These considerations suggest proposing not verifiability, but falsifiability as a demarcation criterion; [...] An empirical-scientific system must be able to break down in the light of empirical evidence." (My translation)

8 I am grateful to Thomas Metzinger for pointing me to Lakatos' work on falsificationism.

from what Imre Lakatos calls naïve falsificationism: for the elimination may be based on a comparison, not on an isolated falsification procedure. If some of the theories are in some sense better than the others (for instance, by making more empirical predictions, or by being less complex), then they can be preferred without having *independent* reasons to reject the eliminated theories. However, Popper's falsificationism is even more sophisticated.

Popper noted that there were no theory-neutral empirical propositions. Descriptions of empirical facts are not immediately given, they are based on observations and involve interpretations (cf. Popper 2005[1934], p. 84, fn. 32). This means it is always possible to add auxiliary hypotheses to a theory, and thereby make the theory compatible with seemingly falsifying evidence. As a consequence, when it comes to determining whether a theory is scientific or not, we cannot consider an isolated theory, but must assume a diachronic stance, in which we consider how a theory is modified in the light of new evidence. Such modifications (e.g., auxiliary hypotheses) increase the empirical content of the theory (cf. Lakatos 1970, p. 183). As Popper puts it:

> Bezüglich der Hilfshypothesen setzen wir fest, nur solche als befriedigend zuzulassen, durch deren Einführung der 'Falsifizierungsgrad' des Systems [...] nicht herabgesetzt, sondern gesteigert wird; in diesem Fall bedeutet die Einführung der Hypothese eine Verbesserung: Das System verbietet mehr als vorher.[9] (Popper 2005[1934], p. 58)

When confronted with evidence that contradicts predictions, we are thus never forced to reject the theory from which the prediction has been derived. We may always modify the theory. But this modification must not be *ad hoc*. Auxiliary hypotheses that only make the theory compatible with the evidence, without having any additional value (without allowing new predictions), are not scientific.

Lakatos (1970) emphasizes that this entails a refined notion of falsificationism. He calls this sophisticated falsificationism (or sophisticated *methodological* falsificationism). A theory can only be falsified in this "sophisticated" manner when it has been replaced by a theory that:

1. has more empirical content (makes new predictions), and
2. makes at least one prediction that is empirically corroborated (cf. Lakatos 1970, pp. 183-184.).

3.1.2 Sophisticated falsification in the Bayesian brain

Popper's sophisticated falsificationism[10] can more easily be applied to predictive processing, because it does not require that we reject a model whenever its predictions yield large prediction errors. Instead, the model can be updated to achieve a better fit with the data. Furthermore, we find a counterpart for the insight that there are no theory-neutral observations: bottom-up signals are never treated as raw data, but as being (more or less) noisy. Hence, prediction errors are weighted by expected precisions. When the expected precision is extremely low, prediction errors will be attenuated. A low expected precision can thus be seen as analogous to an auxiliary hypothesis that makes the model compatible with otherwise contradicting evidence. What is more, it is not an *ad hoc* move, because the precision estimate itself is also constantly being updated in light of the evidence. Similarly, when a model generates a significant amount of prediction error, but is strongly supported by a higher-level model with high prior probability, a relatively high amount of prediction error may not lead to a major revision of the model.

9 "Regarding such auxiliary hypotheses we stipulate that we allow only those hypotheses for which the 'degree of falsifiability' of the system is not decreased, but increased; in this case the introduction of auxiliary hypotheses means an improvement: The system prohibits more than before." (My translation)

10 Lakatos (1970) points out that Popper himself never made a sharp distinction between naïve and sophisticated falsificationism, but that he accepted the assumptions underlying sophisticated falsificationism, at least in parts of his work—whereas the person Karl Popper may have been more of a naïve than a sophisticated falsificationist.

Model competition in PP can also be seen as an instance of sophisticated falsificationism. Competition need not be resolved by eliminating those models that yield the largest prediction errors (as in the starfish robot). Instead, it may be that some models make more specific *counterfactual* predictions. Indeed, this seems to be the main rationale behind active inference in FEP.

According to the formalization provided in Friston et al. (2012, p. 4), active inference involves minimizing the entropy of a counterfactual density. This density links future internal states and hidden controls to hidden states, which cause sensory states; hidden controls are hidden states that can be changed by action (Friston et al. 2012, p. 3). A density has low entropy, roughly, if it assigns high values to a relatively small subset of states, and low values to most other sets of states. Predictions based on a probability density with very low entropy can thus be made with a high level of confidence, because most other possibilities are more or less ruled out (due to the low values assigned to them by the density). Formally, this is reflected in the proposition that the negative entropy of the counterfactual density is a monotonic function of the precision of counterfactual beliefs (Friston et al. 2012, p. 4).

The entropy of the counterfactual density is minimized with respect to hidden controls. In effect, this is a selection process, in which a model (here: a counterfactual density) is selected that has minimal entropy. The other models are eliminated, because they have higher entropies. We can say they are falsified in the sense of sophisticated falsificationism (but not in the sense of naïve falsificationism).

Another way in which model competition can be resolved without naïve falsification can be illustrated by the famous "wet lawn" example (cf. Pearl 1988). Suppose you enter your garden and find that the lawn is wet. There are at least two models that can explain this: either your sprinkler has been on during the night or it has rained. Let us assume that both models are initially equally likely (i.e., they have the same prior probability). When you now observe that your neighbor's garden is also wet, the rain model is corroborated, because it makes the strong prediction that the neighbor's lawn is wet (i.e., the conditional probability that the neighbor's lawn is wet, given that it has rained, is high). The other model is not incompatible with this evidence, but it is not supported by it as much (because the conditional probability that the neighbor's lawn is wet, given that your sprinkler has been on, is not as high). In other words, it has been explained away. As Jakob Hohwy puts it:

> The Rain model accounts for all the evidence leaving no evidence behind for the Sprinkler model to explain. Even though the Sprinkler model did increase its probability in the light of the first observation, it seems intuitive right to say that its probability is now returned to near its prior value. The model has been explained away. (2010, p. 137)

Explaining away is another example of sophisticated falsification. Even when two or more models are compatible with the evidence (and with each other), there can be reason to prefer one of them and reject the others.

The clarification in this section should have shown that there is more to falsification than just "disconfirming" a hypothesis, and that competition between models can be resolved in different ways, not only in the way exemplified by the starfish robot. Furthermore, different types of sophisticated falsificationism are part and parcel of predictive processing.

Does this mean that the Bayesian brain is Popperian? This conclusion would be premature. The above can at best show that there are many situations in which the Bayesian brain is a sophisticated falsificationist. But there may be situations in which not even sophisticated falsification is possible or necessary. In the following section, I will argue that predictive processing also has Kuhnian aspects.

3.2 The Kuhnian Bayesian brain

According to Kuhn, scientific research develops in different recurring phases. Most of the time,

scientists work within an established paradigm, in which implications of theories are explored and puzzles are solved (cf. Kuhn 1962, ch. IV). In this phase, falsification or confirmation do not play a role:

> Normal science does and must continually strive to bring theory and fact into closer agreement, and that activity can easily be seen as testing or as a search for confirmation or falsification. Instead, its object is to solve a puzzle for whose very existence the validity of the paradigm must be assumed. Failure to achieve a solution discredits only the scientist and not the theory. (cf. Kuhn 1962, p. 80)

At some stage, however, there will be anomalies, i.e., empirical observations that cannot be explained within the current paradigm. When these anomalies accumulate, scientists will try to explore new concepts and methods. If, using new concepts and methods, previously unexplainable anomalies can be accounted for, a scientific revolution can result, through which a new paradigm is established. Kuhn shares the sophisticated falsificationist's insight that theories are never rejected in isolation:

> [...] the act of judgment that leads scientists to reject a previously accepted theory is always based upon more than a comparison of that theory with the world. The decision to reject one paradigm is always simultaneously the decision to accept another, and the judgment leading to that decision involves the comparison of both paradigms with nature *and* with each other. (cf. Kuhn 1962, p. 77)

This shows that Kuhn's theory is in some respects in line with sophisticated falsificationism —but he goes beyond it, in that he doubts that a paradigm that has been adopted instead of another is always better or closer to the truth. The reason for this is that he claims competing paradigms to be incommensurable (cf. also Feyerabend 1962), which means that they typically use radically different concepts and methods (cf. Oberheim & Hoyningen-Huene 2013, §1). A new paradigm that becomes dominant is thus not marked by being closer to the truth, but mainly by constituting a departure from the old paradigm (cf. Kuhn 1962, pp. 170-171). This seems to entail that scientific progress need not be a process in which theories approximate the truth to an ever higher degree.

Can we find an analogon for such a transition from one paradigm to the other in predictive processing? Above, we saw that the sophisticated falsificationist assumes that scientific progress happens only when a theory makes new predictions, and thereby leads to the discovery of new states of affairs. This need not always be the case in the Bayesian brain. When a model is changed to minimize free-energy, this does not mean that the empirical content or predictive power has been increased. A particularly clear example of this can be found in perceptual phenomena like binocular rivalry.

In binocular rivalry (cf. Blake & Logothetis 2002), subjects are presented with two different images, one to the left eye, the other to the right eye, e.g., a face and a house. According to a predictive coding account put forward by Jakob Hohwy, Andreas Roepstorff & Karl Friston (2008), the brain generates two main competing models of what the stimuli depict, one corresponding to the face, the other corresponding to the house. However, only one of these models is consciously experienced at any given time (although there can be intermittent phases in which subjects report seeing a mixture of both stimuli, i.e., parts of the house and parts of the face at the same time, but usually non-overlapping). This means that the brain will tend to settle into one of two classes of states (one corresponding to perceiving the house, the other to perceiving the face). Since each of the models can only account for part of the visual input, both cause a significant amount of prediction error (cf. Hohwy et al. 2008, p. 691). Over time, the prior probability of the currently assumed model (house or face, respectively) will decrease, leading to a revision of the hypothesis, until the brain settles into a state corresponding to the other percept, at least temporarily (cf. Hohwy et al. 2008, pp.

692–694).[11] The crucial difference between this and cases like the wet lawn example or model selection in the starfish robot is that neither of the two competing models is in any sense better than the other (in terms of empirical content, simplicity, predictive power, etc.).

We can recast binocular rivalry in terms of Kuhnian paradigm changes. If we liken each of the two models (house/face) to a paradigm, we can say that perceiving a single object in binocular rivalry corresponds to the phase of normal science, in which many phenomena (inputs) can be explained. After some time, however, there are anomalies (increasing prediction error), which leads to a scientific crisis in which new directions are explored (intermittent phase in which no unified percept is generated), until a new form of scientific practice becomes dominant (scientific revolution), and a new phase of normal science (temporarily stable perception) is reached. The transition from one percept to the other does not go along with increased veridicality: neither of the two percepts is closer to the truth than the other.[12] This may also support the cybernetic idea that internal models are used in the pursuit of homeostasis, not to approximate the truth (as also noted by Seth this collection, p. 15).

There is another analogy between the Bayesian brain and Kuhn's theory of science. According to Kuhn, it is indeterminate whether an anomaly (an unexpected experimental result, for instance) is something that should be regarded as just another puzzle or as a reason to reject the whole paradigm:

> Excepting those that are exclusively instrumental, every problem that normal science sees as a puzzle can be seen, from another viewpoint, as a counterinstance and thus as a source of crisis. (Kuhn 1962, p. 79)

If it is treated as a puzzle, it yields questions like: how can we account for this phenomenon within our established framework? If it is treated as a counterinstance, a more fundamental solution is needed. This is analogous to the fact that whether two models in predictive processing are compatible or not depends on (hyper)priors (cf. FitzGerald et al. 2014, p. 2). When a hyper-prior has it that two models are incompatible, this can either lead to a competition, in which one of the models is eliminated, or it can lead to a revision of the hyper-prior. (Which of the two possibilities corresponds more to puzzle solving, and which to something more fundamental will depend on whether the lower-level models or the high-level prior initially have a higher probability.) This is illustrated by the RHI (rubber hand illusion).

In the RHI (Botvinick & Cohen 1998), the brain harbors two contradictory sensory models. According to the visual model, tactile stimulation occurs on the surface of the rubber hand. According to the proprioceptive model, the felt strokes occur at a different location (i.e., where the real hand is located). While there is, in and of itself, no contradiction between these models, it is likely that the brain has a prior that favors common-cause explanations of sensory signals. Relative to this prior, there is a tension between the models: they seem to indicate that the seen stroking and the felt touch occur at distinct locations, which is odd, because they occur synchronously (and the prior has it that synchronous effects have a common cause, which speaks against two distinct locations). As Jakob Hohwy puts it:

> [...] we have a strong expectation that there is a common cause when inputs co-occur in time. This makes the binding hypothesis of the rubber hand scenario a better explainer, and its higher likelihood

11 Two possible reasons why the probability of the currently assumed model decreases are offered by the authors: either there is a hyper-prior to the effect that the world changes (which is why a static hypothesis becomes less likely over time), or there are random effects that lead to multistability, such that neural dynamics switch from one basin of attraction to another (cf. Hohwy et al. 2008, p. 692).

12 In fact, it seems that the notion of incommensurability has been inspired by Gestalt switches (as in the perception of a Neckar cube), which are very similar to phenomena like binocular rivalry. However, Kuhn explicitly pointed out that there is a crucial difference between a Gestalt switch and a paradigm change: "[…] the scientist does not preserve the gestalt subject's freedom to switch back and forth between ways of seeing. Nevertheless, the switch of gestalt, particularly because it is today so familiar, is a useful elementary prototype for what occurs in full-scale paradigm shift" (1962, p. 85). I am grateful to Sascha Fink for drawing my attention to this statement.

> promotes it to determine perceptual inference and thereby resolve the ambiguity. (2013, p. 105)

Notice that the common-cause hypothesis (that the touch is felt where it is seen) only becomes the dominating hypothesis because the design of the study prevents subjects from confirming the distinct-causes hypothesis (e.g., by looking at their real hands). Because of the common-cause hypothesis, there is an ambiguity in the percepts. This ambiguity can be resolved in at least two ways: either by adjusting the lower-level (perceptual) models (to the effect that the felt touch occurs at the same location as the seen stroking); or by active inference (which in this case would lead to a rejection of the higher-level model corresponding to the common-cause hypothesis). The first way corresponds to puzzle solving, the second more closely to a paradigm change. Note that the analogy will be the stronger the more remote the hyper-prior is from the perceptual models.

I hope to have shown that the Bayesian brain has aspects that make it Popperian, as well as aspects that make it Kuhnian. At the very least, it should have become clear that falsification is a more complex concept than depicted in Seth's target paper (which seems to tend towards a more naïve form of falsificationism).

4 Perceptual presence

We have seen how fruitful analogies between PP and theory of science can be. As mentioned above, an early formulation of the analogy between perception and hypothesis-testing can be found in Richard Gregory's seminal paper "Perceptions as Hypotheses". There, we also find the suggestion that percepts *explain* sensory signals (cf. Gregory 1980, p. 13).[13]

How far can we take the analogy between explanation in perception and explanation in science? If we know what a good explanation is in science, does this give us a clue to the conditions under which percepts are experienced as real? Interestingly, there are accounts of scientific explanation that assign an essential role to counterfactual knowledge (cf. Waskan 2008). If someone purports to know why a certain event happened or why a phenomenon was observed, we expect her to also be able to tell us what *would* have happened if some of the initial conditions had been different. Similarly, when the Bayesian brain explains sensory signals by inferring their hidden causes, we would expect the brain's generative model to also have the resources to infer in what ways sensory signals would be different, had there been a change to their hidden causes.

This highlights the relevance of counterfactual models. Seth points out that counterfactuals play a crucial role in active inference. The consideration above may be another way to show the relevance of counterfactual models. Furthermore, it also highlights the usefulness of counterfactual *richness*. The richer a counterfactual model of hidden causes, the better the brain's explanation of sensory signals (all other things being equal). In general, we may also be inclined to say that the richer the counterfactual model, the higher the confidence that it helps track the *real* explanation of sensory signals. But does this mean it goes along with experienced *realness* (or *perceptual presence*)?

This is, basically, what Seth proposes in his PP account of perceptual presence (cf. Seth 2014). But what is perceptual presence in the first place? On the one hand, Seth characterizes the notion by contrasting examples. For instance, objects like a tomato possess perceptual presence, whereas afterimages do not. On the other hand, Seth provides the following characterization:

> In normal circumstances perceptual content is characterized by subjective veridicality; that is, the objects of perception are experienced as real, as belonging to the world. When we perceive the tomato we perceive it as an externally existing object with a back and sides, not simply as a specific view [...]. (2014, p. 98)

13 It should be noted that Gregory ascribes "far less explanatory power" (1980, p. 196) to perceptions than to scientific hypotheses.

The tomato is not perceived as a flat, red disc. Although you do not see the back and sides of the tomato in the same way that you see the front, there is still a sense in which both are *perceptually present* (cf. Noë 2006, p. 414). I shall now point to two ambiguities in Seth's description of the explanandum. This calls for a conceptual clarification, regarding which I shall make a tentative suggestion. After that, I shall argue that there may be possible counterexamples to Seth's hypothesis that perceptual presence correlates with the counterfactual richness of generative models.

4.1 Ambiguities in Seth's description of the explanandum

The tomato is not only experienced as perceptually present, it is also perceived as an *object* in the external world. In a commentary on Seth, Tom Froese (2014, p. 126) has therefore suggested that Seth conflates perceptual presence with experienced *objecthood*. This proposal has some plausibility, because the tomato is perceived as a real object, whereas afterimages are not experienced as objects (they are more like unstable colored shades). After all, even Seth admits, in his target paper, that it may be important to distinguish presence from objecthood (p. 18). This is one way in which Seth's definition of the explanatory target is ambiguous: is it about experienced *presence* or experienced *objecthood* (cf. also Seth 2014, pp. 105f.)? (This question becomes more pressing still when we consider the ethymology of "realness" or "reality": the Latin origin of the word is *res* (thing), which makes it a little confusing that Seth seems to identify perceptual presence with the sense of subjective reality, cf. Seth this collection, p. 2.)

Another ambiguity is related to the notion of a counterfactual model. In his target paper Seth defines a counterfactual model as a model encoding "how sensory inputs (and their expected precisions) would change on the basis of a repertoire of possible actions" (Seth this collection p. 17). On the one hand, one may ask if counterfactual models in the brain necessarily encode SMCs (sensorimotor contingencies). For the perception of a ripe tomato on a bush, it might be equally relevant to encode how sensory signals pertaining to the tomato would change if the wind were to blow the bush or if the tomato were to fall down. On the other hand, it is unclear how *explicit* a counterfactual representation has to be. Jakob Hohwy (2014) suggests that a rich causal structure could be modeled by extracting higher-order invariants (features that do not change if the tomato is dangling in the wind or has fallen down, for instance). Higher-order invariants are relatively perspective-independent.[14] The degree of perceptual presence would then correspond to the "depth of the inverted model"[15] (Hohwy 2014, p. 128). In his target paper, Seth notes that the depth of the model may indeed play a role (see footnote 13).

Two ambiguities are thus to be found in Seth's account. One concerns the characterization of the target phenomenon (experienced *realness* versus experienced *objecthood*). The other lies in the description of the represented causal structure: *counterfactual richness* versus *perspective-independence* of hidden causes. Counterfactual richness and causal "depth" are not completely independent. Below, I will give some examples that may be useful to explore the relationship between these two features. Furthermore, I will suggest that it could be helpful to consider another feature with respect to which the represented causal structure of objects may vary. This feature is the degree of

14 As I am using the term here, the depth of a model can be measured by its location in the predictive processing hierarchy (that is, whether it is high or low in the hierarchy). Estimates at higher levels track features that change more slowly (i.e., features that remain invariant when things change, for instance, when the subject changes her *perspective* on a perceptual object like a tomato by walking around the tomato or by turning it—hence the term "perspective-(in)dependence"). A model of a perceived object is deep when it represents features that change relatively slowly. Alternatively, one could stipulate that a model is deep when it represents features that change slowly *and* features that change more quickly. In fact, this may come closer to what Hohwy has in mind, but it blurs the distinction between perspective-dependence and causal integration. Hohwy writes: "[c]oncurrents are causes that do not interact on their own with other causes (presumably a fence won't occlude a concurrent)" (2014, p. 128). But encapsulated causes can be represented both at lower parts of the hierarchy (possible example: afterimages) and at higher parts of the hierarchy (possible example: certain conscious thoughts). This suggests that at least causal encapsulation can be dissociated from perspective-dependence and -independence.

15 The inverted model is the posterior distribution, the computation of which is based on the likelihood and the prior (see above).

causal encapsulation. For representations not only differ with respect to their counterfactual richness or their degree of perspective-dependence, but also with respect to the extent to which the represented causal structure is encapsulated or integrated. (In what follows, I will use the notion of a counterfactual model mainly in the sense in which Seth uses it: counterfactual models in this sense involve representations of possible bodily actions by the subject of experience.)

A phenomenal representation of a tomato on a plate is not only counterfactually rich and relatively perspective-dependent, the represented causal structure is also causally *integrated*.[16] It is, for instance, represented as being causally related to the plate, because it is experienced as lying *on* the plate (that is, it is not hovering above it). Furthermore, it is in possible causal contact with virtually all other objects in its vicinity (e.g., the subject's hands).

Contrast this with the experience of what is happening in a classical video game—say, a racing game. The player influences how the images on the two-dimensional screen change, because she has control over the vehicle. Hence, we can assume that representations of gaming sequences are (usually) counterfactually rich. At the same time, they are also perspective dependent (although they mainly depend on the *virtual* perspective from which objects are represented in the game). However, virtual objects in the game are experienced as causally encapsulated: although objects can interact with each other in the virtual world, they do not interact with most other parts of the player's environment. For instance, they will never break out of the screen and fly around in the room in which the player is sitting. Furthermore, they can only be influenced vicariously through a controller or keyboard. Thus there is not causal encapsulation in *every* respect (the virtual world is not experienced as completely disambiguated from the rest of the experienced world), but in *some* respects the encapsulation is rather strong (the virtual world is spatially bounded, e.g., with the screen as the limit). Note that many modern video games are less causally encapsulated, for instance when they are played on a touchscreen (or on devices with a three-dimensional screen, or in an immersive virtual reality).[17]

As mentioned above, causal integration and counterfactual richness are not completely independent. High counterfactual richness implies a certain degree of causal integration (at least in some respects), for it means that at least the subject can interact with the experienced object in some way—regardless of how separate the represented causal structure is from the rest of the subject's surroundings.

Similarly, highly perspective-invariant representations typically also involve the representation of an encapsulated causal structure. Abstract conscious thoughts, for instance, cannot be touched with the hand or other concrete objects. However, the implied encapsulation only holds in some respects. Sometimes thoughts can evoke strong emotions or a sequence of mental imagery. In certain obsessive-compulsive disorders, for instance, subjects will first have a thought ("My hands are dirty"), presumably followed by a feeling of disgust and the urge to wash the hands, which then leads to motor behavior (washing the hands); this, in turn, may be followed by the thought that the hands are still dirty. The content of the conscious thought is relatively perspective-invariant, and yet it involves, presumably, representations of causal structure that link it to concrete objects in the world.

As long as we interpret counterfactuals only as representations of sensorimotor contingencies, it may also seem that perspective-invariant[18] representations are counterfactually poor. However, if we include representations of possible *mental* actions and their effects, we can also conceive of counterfactually-rich perspective-invariant representations. A possible example is a philosophical argument or a theory, which someone can contemplate in their mind, being aware that there are several possible ways

16 Another possible term for this would be *causally open*, in the sense that it is represented as being in potential causal exchange with other objects in its surrounding. By integration, I thus do not mean integration *within* (or internal integration), but integration *with* other objects.

17 Thanks to Jennifer Windt for suggesting immersive video games as a further example.

18 Perspective-invariant representations are maximally perspective-independent.

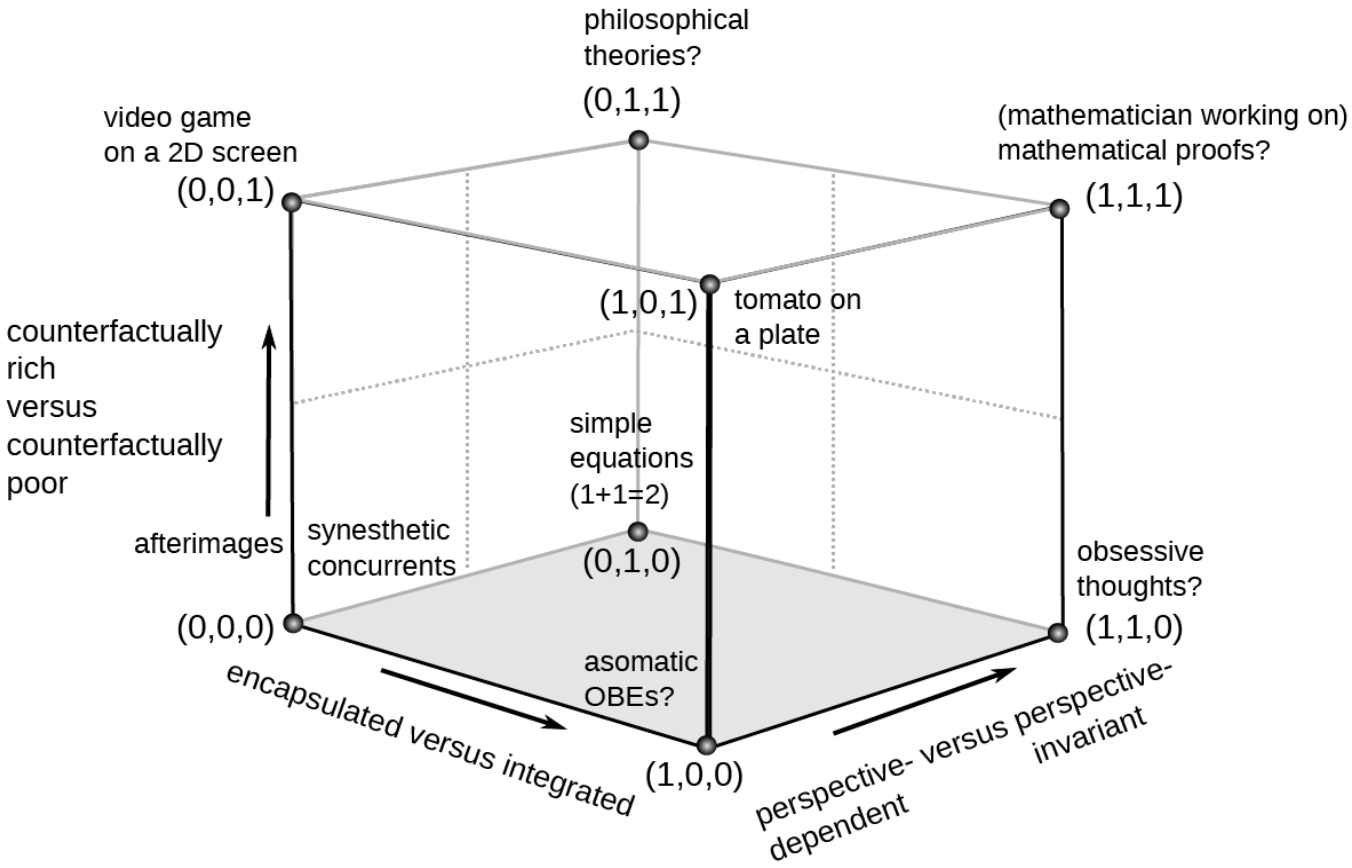

Figure 1: The figure illustrates how classes of experiences can be located in a cube, according to the extent to which they display counterfactual richness, perspective-independence, and causal integration (see main text for explanations). The cube (without the labels) is adapted from cube figures in Godfrey-Smith (2009); talks by Daniel Dennett brought this style of illustration to my attention.

in which the argument could be probed and attacked, or several important cases to which the theory could be applied.

Bearing in mind that the degree of causal encapsulation is not completely independent from the other two dimensions (counterfactual richness and perspective-invariance), we can depict different types of conscious experiences in a cube, where the three axes stand for the three dimensions described (see Figure 1). The most interesting locations in this cube are, of course, its eight corners, because they depict classes of experiences for which each of the three features is either completely absent or maximally pronounced. Finding examples of these "extremal experiences" is no easy task.[19] Even neural representations of synesthetic concurrents, Seth's prime example of counterfactually poor models, may, at first sight, seem to be located somewhere in the middle of the perspective-dependence axis.

Grapheme-color concurrents, for instance, are not simply triggered by graphic representations of glyphes, but by representations of abstract objects, i.e., graphemes, associated with certain glyphes (cf. Mroczko et al. 2009). Hence, it may seem that the hidden cause of the concurrent is not simply an object in the world, but also involves an abstract object, i.e., a grapheme, the representation of which is perspective-invariant. This would suggest that synesthetic concurrents cannot conclusively be placed in one of the cube's corners, because their represented hidden causes involve very high-level invariants.

On the other hand, one could object that the concurrent itself is represented in a rather perspective-dependent way. It may be part of a

19 In fact, it may be that the corners only constitute hypothetical endpoints. Thanks to Jennifer Windt for pointing this out.

causal network involving hidden causes that are represented in perspective-invariant ways, but the synesthetic percept itself is not a representation of an abstract hidden cause.[20] Hence, on second thought, it seems that concurrents, as in grapheme-color synesthesia, are in fact located close to the origin of our coordinate system: the representations involved are relatively perspective dependent, and they are counterfactually poor. At the same time, they are causally encapsulated, because they do not interact with physical objects (they cannot be touched, etc.).

4.2 Does counterfactual richness correlate with perceptual presence (or objecthood)?

What does this tell us about experienced "presence" or "objecthood"? Are all examples of counterfactually rich representations in the cube perceptually present, or are they associated with a high degree of objecthood? If so, this would support Seth's hypothesis that counterfactual richness correlates with perceptual presence (or objecthood). I believe that counterfactual richness can be dissociated both from perceptual presence and from objecthood. Olfactory experiences are, as argued by Michael Madary (2014), both counterfactually poor and perceptually present. This suggests that counterfactual richness does not correlate with perceptual presence. Similarly, experiences of classical video game sequences are counterfactually rich, but involve a low degree of perceptual presence; objects in the game are only experienced as virtual objects, not as real objects. Counterfactual richness and perceptual presence may therefore be doubly dissociable.

Trying to evaluate whether counterfactual richness correlates with phenomenal objecthood would presuppose that we know what phenomenal objecthood means. As I only have an intuitive grasp of what it means, I can only give a preliminary statement. To me, it seems that virtual objects in two-dimensional video games do not possess a high degree of phenomenal objecthood. But then again, even if a virtual tomato could be manipulated in various ways with a controller, the corresponding representation would probably not be as counterfactually rich as a representation corresponding to the experience of a real tomato. Hence, it is difficult to arrive at a definitive verdict.

A more promising path may involve the experience of objects in asomatic OBEs (out-of-body experiences) or asomatic dream experiences (Windt 2010; Metzinger 2013). Counterfactuals, as conceived of by Seth, always involve action on the part of a subject. Most, if not all, (non-mental) actions involve the body, so representing counterfactuals involves representing (parts of) the body. In asomatic OBEs and asomatic dream experiences, subjects do not identify with a body, but with an unextended point in space. I speculate that in such cases, representations of objects are less counterfactually rich.[21] This, however, does not necessarily mean that they are experienced as less present or as possessing less objecthood. There are still a lot of causal regularities involving external objects that may be tracked by models in the brain, even in the absence of an ordinary body representation. External objects can interact with each other, and counterfactual representations of possible causal processes may contribute to the experience of objecthood or perceptual presence. In particular, this is to be expected if none of the external objects are represented as causally encapsulated. If this bears out, it provides another reason to believe that counterfactual richness of generative models does not correlate with experienced objecthood. Let us now consider possible examples of other extremal experiences (in the corners of the cube) to investigate whether it is plausible to hypothesize that represented causal depth or causal encapsulation correlates with perceptual presence or objecthood.

The more perspective-invariant a representation, the more abstract it is. This also means that perspective-invariant representations typically involve an encapsulated causal structure. Thinking about a simple equation like

20 This may point to an aspect regarding which Hohwy's characterization of causal depth is ambiguous.

21 In fact, asomatic OBEs may be a better example than asomatic dream experiences, since such dreams typically lack concrete objects (cf. LaBerge & DeGracia 2000). I am grateful to Jennifer Windt for pointing this out.

"1+1=2" may be an example of this. There is no way in which the target of this representation can causally interact with the window behind my desk or the red bottle in front of the window. Furthermore, most (or all) bodily movements will not influence the way I experience the thought that one plus one equals two. Hence, it is arguably also a counterfactually poor representation.

When we move up, in the direction of counterfactually rich phenomenal representations, we arrive at representations that are counterfactually rich, perspective-invariant, and still causally encapsulated. Above, I mentioned conscious thoughts about philosophical arguments or theories as possible examples. Such thoughts may involve mental imagery and inner speech, and perhaps even complex phenomenal simulations involving counterfactual situations. It is not obvious whether it makes sense to say that such thoughts involve counterfactual representations linking possible mental actions to their effects. This is even harder without presupposing a developed theory of mental action (for recent proposals, cf. Proust 2013; Wu 2013).

Mental actions are goal-directed. Performing a mental action may therefore, at least in some cases, be followed by a representation of a situation in which the goal is realized (one possible example might be: remembering a name; represented situation: telling someone the name). In the case of a theory, a mental action could be considering whether a certain claim is true or not (or whether it is plausible). This may trigger thoughts like: "Assuming this is the case, what implications would this have? Are these implications plausible, or likely to be true? Are there possible counterexamples?" It might also involve trying to formulate something more clearly.

Furthermore, thinking about a theory or problem may involve conscious counterfactual thoughts of the form "If I gave up this assumption, there would not be a contradiction among the remaining hypotheses anymore", or "If the theory could account for this special case, it would be strengthened". One difference to conscious perception of concrete objects is, presumably, that such counterfactuals are *phenomenally* represented, whereas representations of SMCs are usually unconscious (and may impact on consciousness only indirectly).

Similar things apply to conscious thoughts about non-trivial mathematical expressions. For instance, if a mathematician sees the expression $(1 + x/n)^n$ she will probably think "If n tends to infinity, this expression will converge to e^x. Now, suppose the mathematician is investigating the asymptotic behavior of some complicated expression (e.g., she wants to find out what happens to a certain expression when n tends to infinity). While manipulating the terms on paper, she suddenly realizes that one factor contained in the expression is $(1 + x/n)^n$. As she is using pen and paper while thinking this, her brain will not only activate an abstract (but conscious) counterfactual thought, but probably also a representation of SMCs. These SMCs will involve taking the limit of the expression with which she started (i.e., $lim\ n \to \infty$), and this is now not only a mental action, but also a possible bodily action. She could write this down, and know that (if the limit exists) part of it would be e^x. Her mathematical investigation therefore involves:

- phenomenal representations regarding counterfactual mental actions;
- representations of SMCs (*embodied* versions of the above mentioned counterfactuals);
- a close coupling between writing, perceiving, and thinking.

The third point is especially important, because it suggests that for a mathematician working with pen and paper (or chalk and blackboard) the objects of her conscious thoughts are not causally encapsulated anymore. The causal structure represented while thinking about abstract concepts is intertwined with the causal structure represented while looking at written mathematical expressions. These causal relations are still relatively limited, but if the mathematician is completely absorbed in her work, the paper (or blackboard) may be all she is attending to in her environment at the moment, perhaps to the extent that she does not experi-

ence abstract relations represented by her notes as causally encapsulated anymore. It is conceivable that this aspect can be enhanced in virtual environments in which mathematical objects are not represented by writing on paper or blackboard, but by three-dimensional virtual objects that can be manipulated by touch or manual movements, for instance.[22] Contrary to what one might at first think, there may thus be cases in which high-degrees of perspective-invariance go along with both counterfactual richness and high degrees of causal integration.

Another class of abstract thoughts that may be experienced as causally integrated could be obsessive thoughts, like the thought that one's hands are contaminated with germs. Such thoughts may be triggered by specific events (like touching a door knob) and may go along with a fear of getting sick (because of the contamination). Subjects may also try to avoid touching objects that they fear might be contaminated. The reason for this is that the hidden cause represented by the obsessive thought, i.e., potential germ contamination, is not causally encapsulated. It is causally connected to concrete objects in the subjects' environment: things that are perceived as contaminated can cause a contamination of the hands; on the other hand, contaminated hands can infect other objects with germs. Furthermore, the inferred hidden cause (germ contamination) is relatively perspective-invariant. Subjects arguably do not imagine bacteria crawling on their hands, although the obsessive thought may go along with an altered perception of the hands. Finally, the model involved is probably counterfactually poor, as most actions do not change the alleged contamination (with the possible exception of washing the hands or touching allegedly contaminated objects; but here, the counterfactual effect is probably just an increase or decrease in the acuteness of the felt contamination). Therefore, I list obsessive thoughts as candidate examples of counterfactually poor, perspective-invariant representations the contents of which are represented as causally integrated.

22 This could be a case in which there is a particularly strong demand for the general ability of PP to combine "fast and frugal solutions" with "more structured, knowledge-intensive strategies" (Clark this collection).

4.3 Do perspective-invariance or represented causal integration correlate with perceptual presence (or objecthood)?

The examples given are certainly not uncontroversial and perhaps not all of them can be sustained in the light of further research. But hopefully the cube can still fulfill heuristic purposes, and can illustrate the need to clarify the relations between counterfactual richness, perspective-dependence, and causal integration. But assuming that the examples given are located in roughly the right places within the cube, what does this tell us about perceptual presence or experienced objecthood? Above, I dismissed Seth's hypothesis that counterfactual richness correlates with either presence or objecthood. Let us now briefly consider perspective-invariance and causal integration. If conscious thoughts involve causally-deep models (that represent perspective-invariant features), then it seems that the depth of the represented causal structure does not correlate with presence or objecthood. The thought that one plus one equals two does not possess a high degree of objecthood or perceptual presence. Hence, it seems that Hohwy's hypothesis that the depth of the generative model (the degree of perspective-independence) correlates with objecthood or presence should be dismissed as well. But the remaining candidate, causal integration, does not unequivocally correlate with either presence of objecthood (*if* the examples I gave make sense). The represented causal structure in obsessive thoughts need not be encapsulated, and still they are probably not accompanied by experienced objecthood or perceptual presence. Perhaps this shows that one ought first to clarify whether it even makes sense to talk about the phenomenology of objecthood or presence with respect to conscious thoughts.

4.4 How does perception change when new sensorimotor contingencies are learnt?

Another relevant question is whether increasing the degree of counterfactual richness, causal integ-

ration, or causal depth of a model just modifies (or enriches) the inferred hidden causes, or whether it leads to the perception of a new, possibly more abstract object. This relates to the question raised in the target paper, namely whether a person who is highly familiar with an object perceives it as more real (because she has mastery of more SMCs) than other persons (Seth this collection, p. 18). Interestingly, research on learning new SMCs tentatively suggests that it leads to the perception of new (more abstract) objects.

Under the lead of Peter König, cognitive scientists from Osnabrück have, in recent years, developed a compass belt that indicates to the person wearing it (while moving) changes in directions (cf. Kaspar et al. 2014). The aim of this project (called *feelspace*) is to study how perception in new sensory modalities can be enabled by sensory augmentation.[23] The belt (see Figure 2) contains several vibrators, which always signal the direction of magnetic north. Subjects who wear the belt for a couple of weeks learn new SMCs, e.g., related to how the vibrating signals change when they turn around. A straightforward application of Seth's PPSMCT suggests that the increased counterfactual richness simply goes along with an increased perceptual presence (for the belt, or the vibrations, or the hip / waist, etc). But the authors of the study cited report that perception changes in different ways:

> Initially the signal was predominantly perceived as tactile evolving to being perceived as location and direction information. Over time, the perception of tactile stimulation receded more and more into the background. Instead the subjects' reports focused more on changes in spatial perception. Furthermore, two months after the end of belt wearing the effects subjects reported – at least in the FRS questionnaire – diminished. (Kaspar et al. 2014, p. 59)

What changes is not just that SMCs for tactile stimulation on the skin where the belt is worn are learnt, but that these are connected to more abstract information (regarding location and direction). This also makes sense in comparison with other sensory modalities. Knowledge of auditory SMCs, for instance, does not increase the perception of the inner ear. When the brain learns the relevant SMCs, it thereby learns about the hidden causes of signals in the inner ear. In fact, this may be another reason to believe that counterfactual richness goes along with phenomenal objecthood.

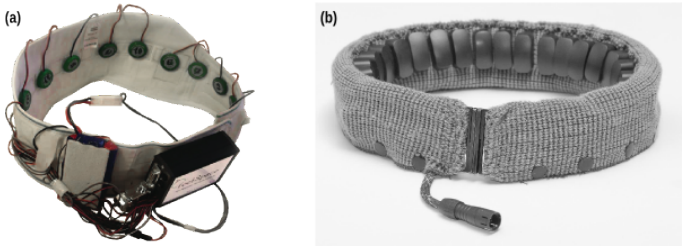

Figure 2: The figure shows two versions of the feelspace belt. (a) The original version used in Nagel et al. (2005). (b) The current version used in Karcher et al. (2012) and Kaspar et al. (2014). Images used with kind permission of Peter König.

This also suggests that when someone is more familiar with an object, the object itself need not become more real, but its connections to other objects might. The causal network in which it is embedded becomes more real. Perhaps the subject also experiences more abstract objects (corresponding to higher-level invariants).

All in all, I hope the examples given illustrate the need to provide a conceptually clearer account of counterfactual richness, causal depth, and causal integration. For at the moment it seems that they are too entangled to allow us to assess their potential relevance for experienced objecthood or presence in a rigorous way. Furthermore, it will be crucial to investigate how phenomenal properties are affected when there are *changes* in these three features (e.g., when counterfactual richness or causal integration is increased or decreased in a controlled way in a study).

5 Conclusion

I have tried to show that useful analogies between PP accounts and classical ideas in the-

23 For more information on the project, see: http://feelspace.cogsci.uni-osnabrueck.de/

ory of science run deeper than portrayed in Seth's target paper. Based on such analogies, I have argued that a proper treatment of active inference needs to be more sophisticated than Seth's threefold distinction. In particular, Seth blurs a whole range of ways in which models can be falsified.

Furthermore, I have suggested that Seth's predictive processing account of perceptual presence may profit from taking not just the counterfactual richness of generative models, but also their degree of perspective-dependence and their causal encapsulation into account (as mentioned above, this suggestion is inspired by Jakob Hohwy's work). I have proposed a way in which examples of possible combinations of these features can be explored, which may serve as a useful tool for future research.

Thomas Kuhn (1962, p. 88) writes that "normal science usually holds creative philosophy at arm's length, and probably for good reasons". I thus hope that research on predictive processing and consciousness has not yet reached the phase of normal science, so that this commentary can still make a humble contribution.

Acknowledgments

I am grateful to two anonymous reviewers, and to Jennifer Windt and Thomas Metzinger especially for providing a vast number of comments and remarks, which helped tremendously in revising the first draft of this paper. This comment was written with support by a scholarship from the Barbara Wengeler foundation.

References

Ashby, W. R. (1947). Principles of the self-organizing dynamic system. *The Journal of General Psychology*, *37* (2), 125-128. 10.1080/00221309.1947.9918144

Baldi, P., & Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks*, *23* (5), 649-666. 10.1016/j.neunet.2009.12.007

Blake, R. & Logothetis, N .K. (2002). Visual competition. *Nature Reviews Neuroscience*, *3* (1), 13-21.

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, *314* (5802), 1118-1121. 10.1126/science.1133687

Botvinick, M. & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, *391* (6669), 756-756. 10.1038/35784

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36* (3), 181-204. 10.1017/S0140525X12000477

——— (2015). Embodied Prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-21). Frankfurt a. M., GER: MIND Group.

Feyerabend, P. (1962). Explanation, reduction and empiricism. In H. Feigl & G. Maxwell (Eds.) *Scientific explanation, space, and time* (pp. 28-97). Minneapolis, MN: University of Minnesota Press.

FitzGerald, T. H., Dolan, R. J., & Friston, K. J. (2014). Model averaging, optimal inference and habit formation. *Frontiers in Human Neuroscience*, *8* (457), 1-11. 10.3389/fnhum.2014.00457

Friston, K. J. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13* (7), 293-301. 10.1016/j.tics.2009.04.005

——— (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11* (2), 127-138. 10.1038/nrn2787

Friston, K. J., Adams, R., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in Psychology*, *3* (151), 1-20. 10.3389/fpsyg.2012.00151

Froese, T. (2014). Steps toward an enactive account of synesthesia. *Cognitive Neuroscience*, *5* (2), 126-127. 10.1080/17588928.2014.905521

Godfrey-Smith, P. (2009). *Darwinian populations and natural selection.* Oxford, UK: Oxford University Press.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *290* (1038), 181-197. 10.1098/rstb.1980.0090

Hohwy, J. (2010). The hypothesis testing brain: some philosophical applications. In W. Christensen, E. Schier & J. Sutton (Eds.) *Proceedings of the 9th Conference of the Australasian Society for Cognitive Science* (pp. 135-144). Macquarie Centre for Cognitive Science. 10.5096/ASCS200922

——— (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3* (96), 1-14. 10.3389/fpsyg.2012.00096

——— (2013). *The predictive mind.* Oxford, UK: Oxford University Press.

——— (2014). Elusive phenomenology, counterfactual awareness, and presence without mastery. *Cognitive Neuroscience*, *5* (2), 127-128. 10.1080/17588928.2014.906399

Hohwy, J., Roepstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108* (3), 687-701. http://dx.doi.org/10.1016/j.cognition.2008.05.010

Itti, L. & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49* (10), 1295 – 1306. http://dx.doi.org/10.1016/j.visres.2008.09.007

Kärcher, S. M, Fenzlaff, S., Hartmann, D., Nagel, S. K., & König, P. (2012). Sensory augmentation for the blind. *Frontiers in Human Neuroscience*, *6* (37), 1-15. Frontiers Media SA. 10.3389/fnhum.2012.00037

Kaspar, K., König, S., Schwandt, J., & König, P. (2014). The experience of new sensorimotor contingencies by sensory augmentation. *Consciousness and Cognition*, *28*, 47-63. 10.1016/j.concog.2014.06.006

Kuhn, T. S. (1974). *The structure of scientific revolutions.* Chicago, IL: The University of Chicago Press.

LaBerge, S. & DeGracia, D. J. (2000). Varieties of lucid dreaming experience. In R. G. Kunzendorf & B. Wallace (Eds.) *Individual differences in conscious experience* (pp. 269-307). Amsterdam, NL: John Benjamins.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & Musgrave, A. (Eds.) *Criticism and the growth of knowledge* (pp. 91-196). Cambridge, UK: Cambridge University Press.

Madary, M. (2014). Perceptual presence without counterfactual richness. *Cognitive Neuroscience*, *5* (2), 131-132. 10.1080/17588928.2014.907257

Metzinger, T. K. (2013). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. *Frontiers in Psychology*, *4* (746). 10.3389/fpsyg.2013.00746

Mroczko, A., Metzinger, T., Singer, W., & Nikolić, D. (2009). Immediate transfer of synesthesia to a novel inducer. *Journal of Vision*, *9* (12), 1-8. 10.1167/9.12.25

Nagel, S. K., Carl, C.,Kringe, T., Märtin, R., & König, P. (2005). Beyond sensory substitution--learning the sixth sense. *Journal of Neural Engineering*, *2* (4), 13-26. 10.1088/1741-2560/2/4/R02

Nickles, T. (2014). Scientific revolutions. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/scientific-revolutions/

Noë, A. (2006). Experience without the head. In T. S. Gendler & J. Hawthorne (Eds.) *Perceptual experience* (pp. 411-434). Oxford, UK: Oxford University Press.

Oberheim, E. & Hoyningen-Huene, P. (2013). The incommensurability of scientific theories. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/incommensurability/

Pearl, J. (1988). Embracing causality in default reasoning. *Artificial Intelligence*, *35* (2), 259-271. 10.1016/0004-3702(88)90015-X

Popper, K. R. (2005[1934]). *Logik der Forschung.* Tübingen, GER: Mohr Siebeck.

Proust, J. (2013). Mental acts as natural kinds. In A. Clark, J. Kiverstein & T. Vierkant (Eds.) *Decomposing the will* (pp. 262-280). Oxford, UK: Oxford University Press.

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, *5* (2), 97-118. 10.1080/17588928.2013.877880

——— (2015). The Cybernetic Bayesian Brain. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1-25). Frankfurt a. M., GER: MIND Group.

Von Helmholtz, H. (1959). *Die Tatsachen in der Wahrnehmung. Zählen und Messen.* Darmstadt, GER: Wissenschaftliche Buchgesellschaft.

Waskan, J. (2008). Knowledge of counterfactual Interventions through cognitive models of mechanisms. *International Studies in Philosophy of Science*, *22* (3), 259-275. 10.1080/02698590802567308

Windt, J. M. (2010). The immersive spatiotemporal hallucination model of dreaming. *Phenomenology and the Cognitive Sciences*, *9* (2), 295-316. 10.1007/s11097-010-9163-1

Wu, W. (2013). Mental action and the threat of automaticity. In A. Clark, J. Kiverstein & T. Vierkant (Eds.) *Decomposing the will* (pp. 244-261). Oxford, UK: Oxford University Press.

Inference to the Best Prediction

A Reply to Wanja Wiese

Anil K. Seth

Responding to Wanja Wiese's incisive commentary, I first develop the analogy between predictive processing and scientific discovery. Active inference in the Bayesian brain turns out to be well characterized by abduction (inference to the best explanation), rather than by deduction or induction. Furthermore, the emphasis on control highlighted by cybernetics suggests that active inference can be a process of "inference to the best prediction", leading to a distinction between "epistemic" and "instrumental" active inference. Secondly, on the relationship between perceptual presence and objecthood, I recognize a distinction between the "world revealing" presence of phenomenological objecthood, and the experience of "absence of presence" or "phenomenal unreality". Here I propose that world-revealing presence (objecthood) depends on counterfactually rich predictive models that are necessarily hierarchically deep, whereas phenomenal unreality arises when active inference fails to unmix causes "in the world" from those that depend on the perceiver. Finally, I return to control-oriented active inference in the setting of interoception, where cybernetics and predictive processing are most closely connected.

Keywords
Abduction | Control-oriented active inference | Falsification | Objecthood | Presence

Author

Anil K. Seth
a.k.seth@sussex.ac.uk
University of Sussex
Brighton, United Kingdom

Commentator

Wanja Wiese
wawiese@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Editors

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

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

It is a pleasure to respond to Wanja Wiese's stimulating commentary (this collection), from which I learned a great deal. Much of what he says stands easily by itself, so here I select just a few key points which warrant further development in light of his analysis.

2 Active inference and hypothesis testing

A central claim in my target paper is that active inference, typically considered as the resolution of sensory prediction errors through action, should also (perhaps primarily) be considered as furnishing disruptive and/or disambiguatory evidence for perceptual hypotheses. This claim transparently calls on analogies with hypothesis testing in science (as well as on counterfactually-equipped generative models), and so invites comparisons with theoretical frameworks for scientific discovery, as Wiese nicely develops. In particular, Wiese notes that I do not "say much about what it takes to disconfirm or falsify a given hypothesis or model", inviting me to "provide a refined treatment of the relation between falsification and active inference" (this collection, p. 2). This is what I shall attempt in this first section.

2.1 The abductive brain

Wiese rightly says that a strict Popperian analogy for active inference is inappropriate since Popperian falsification relies on hypotheses that are derived deductively. Deductive inferences are *necessary inferences*, meaning that their falsification in turn falsifies the premises (theories) from which they derive. Active inference in the Bayesian brain is not deductive for two important reasons. First, as Wiese notes, Bayesian inference is inherently probabilistic so that competing hypotheses become more or less likely, rather than corroborated or falsified. Probabilistic weighting of hypotheses suggests a process of *induction* rather than deduction. Inductive inferences are *non-necessary* (i.e., they are not inevitable consequences of their premises) and are assessed by observation of outcome statistics, by analogy with classical statistical inference. Second, Bayesian reasoning pays attention not just to outcome frequencies but to properties of the explanation (hypothesis) itself, as captured by the slogan that (Bayesian) perception is the brain's "best guess" of the causes of its sensory inputs. This indicates that the Bayesian brain is neither deductive nor inductive but *abductive* (Hohwy 2014), where abduction is typically understood as "inference to the best explanation". In Bayesian inference, what makes a "best" explanation rests not only on outcome frequencies, but also on quantification of model complexity (models with fewer parameters are preferred), and by priors, likelihoods, as well as hyper-priors which may make some prior-likelihood combinations more preferable than others. Importantly, abductive (and inductive) processes are *ampliative*, meaning that they are capable of going beyond that which is logically entailed by their premises. This is important for the Bayesian brain, because the fecundity and complexity of the world (and body) requires a flexible and open-ended means of adaptive response.

So, the Bayesian brain is an abductive brain. But I would like to go further, recalling that active inference enables predictive *control* in addition to perception. This emphasis is particularly clear in the parallels with cybernetics and applications to interoception developed in the target article, where allostasic[1] control of 'essential variables' is paramount, and where predictive models are recruited towards this goal Conant & Ashby 1970; Seth 2013). In this light, active inference in the cybernetic Bayesian brain becomes a process of "inference to the best *prediction*", where the "best" predictions are those which enable control and homeostasis under a broad repertoire of perturbations.[2] It will be interesting to fully develop criteria for "best-making" in this control-oriented form of abductive inference.

2.2 Sophisticated falsificationism, active inference, and model disambiguation

Where does this leave us with respect to theories of scientific discovery? Strict Popperian falsification was already discounted as an analogy for active inference. At the other extreme, parallels with Kuhnian paradigm shifts also seem inappropriate since these are not based on inference whether deductive, inductive, or abductive. Also, such shifts are typically unidirectional: having dispensed with the Copernican worldview once, we are unlikely to return to it in the future. These two points challenge Wiese's analogy between paradigm shifts and perceptual transitions in bistable perception (see Wiese's footnote 12, this collection, p. 9). What best survives in this analogy is an appeal to hierarchical inference, where changes in "paradigm" correspond to alternations between hierarchically deep predictions, each of which recruit more fine-grained predictions which themselves each explain only part of the ongoing sensorimotor flux, under the hyper-prior that perceptual scenes must be self-consistent (Hohwy et al. 2008).

Wiese himself seems to favour Lakatos' interpretation of Popper, a "sophisticated falsificationism" where theories (perceptual hypotheses) can be modified rather than rejected outright, when predictions are not confirmed,

1 Allostasis: the process of achieving homeostasis.

2 There is an interesting analogy here to the overlooked "perceptual control theory" of William T. Powers, which says that living things control their perceived environment by means of their behavior, so that perceptual variables are the targets of control (1973).

and where hypotheses are not tested in isolation (more on this later). As Wiese shows, sophisticated falsification fits well with some aspects of Bayesian inference, like model updating. According to Lakatos, core theoretical commitments can be protected from immediate falsification by introducing "auxiliary hypotheses" which account for otherwise incompatible data (1970). The key criterion - in the philosophy of science sense - is that these auxiliary hypotheses are *progressive* in virtue of making additional testable predictions, as opposed to *degenerate*, which is when the core commitments become less testable.[3] This maps neatly to counterfactually-equipped active inference, where hierarchically deep predictive models spawn testable counterfactual sensorimotor predictions which are selected on the basis of precision expectations, and which lead to effective updating (rather than "falsification") of perceptual hypotheses. As Wiese notes, a good example of this is given by Friston and colleagues' model of saccadic eye movements (Friston et al. 2012). When it comes to model comparison, sophisticated falsification may even approximate some aspects of abductive inference: "Explaining away is another example of sophisticated falsification. Even when two or more models are compatible with the evidence … there can be reason to prefer one of them and reject the other" (Wiese this collection, p. 7). This strongly recalls Bayesian model comparison and "inference to the best explanation", if not its control-oriented "inference to the best prediction" form.

One important clarification is needed about Wiese's interpretation of model comparison, highlighting the critical roles of action and counterfactual processing. Wiese rightly emphasizes the important insight of Popper and Lakatos that hypotheses are never tested in isolation, mandating a process of comparison among competing models or hypotheses. However, he implies a sequential testing of each hypothesis: "balloons being launched and then shot done, one by one" (see Wiese this collection, p. 6). This is quite different from the interpretation of model comparison pursued in my target article, where multiple models are considered in parallel, and where counterfactual predictions are leveraged to select the action (or experiment) most likely to *disambiguate* competing models. In Bayesian terms this is reflected in a shift towards model comparison and averaging (FitzGerald et al. 2014; Rosa et al. 2012), as compared to inference and learning on a single model. Bongard and colleagues' evolutionary robotics example was selected precisely because it illustrates this point so well (Bongard et al. 2006). Here, repeated cycles of model selection and refinement lead to the prescription of novel actions that best disambiguate the current best models (note the plural). Indeed, it is the repeated refinement of disambiguatory actions that gives Bongard's starfish robot its compelling "motor babbling" appearance. To reiterate: different actions may be specified when the objective is to disambiguate multiple models in parallel, as compared to testing models one-at-a-time. In the setting of the cybernetic Bayesian brain this example is important for two reasons: it underlines the importance of counterfactual processing (to drive the selection of disambiguatory actions) and it emphasizes that predictive modelling can be seen as a means of control in addition to discovery, explanation, or representation. In this sense it doesn't matter how accurate the starfish self model is – what matters is whether it works.

2.3 Science as control or science as discovery?

The distinction between explanation and control returns us to the philosophy of science. Put simply, the views of Popper, Lakatos, and (less so) Kuhn, are concerned with how science reveals truths about the world, and how falsification of testable predictions participates in this process. Picking up the threads of abduction,

3 An important application of this idea is to the Bayesian brain itself as a scientific hypothesis. A concern about the Bayesian brain hypothesis is that it can be insulated from falsification by postulating convenient (typically unobservable) priors, much like adaptationist explanations in evolutionary biology can be critiqued as "just so" stories. The key question, not answered here, is whether neural mechanisms implement (approximations to) Bayesian inference, or whether Bayesian concepts merely provide a useful interpretative framework. In the former case one would require the Bayesian brain hypothesis to be progressive not degenerate.

control-oriented active inference, and "inference to the best prediction", we encounter the possibility that theories of scientific discovery might themselves appear differently when considered from the perspective of control. Historically, it is easy to see the narrative of science as a struggle to gain increasing control over the environment (and over people), rather than a process guided by the lights of increasing knowledge and understanding.[4] A proper exploration of this territory moves well beyond the present scope (see e.g., Glazebrook 2013). In any case, whether or not this perspective helps elucidate scientific practice, it certainly suggests important limits in how far analogies can be taken between philosophies of scientific discovery and the cybernetic Bayesian brain.

3 Perceptual presence and counterfactual richness

The second part of Wiese's commentary picks up on the issue of *perceptual presence*, which in my target article was associated with the "richness" of counterfactual sensorimotor predictions (see also Seth 2014, 2015b). Wiese makes a number of connected points. First, he rightly notes an ambiguity between objecthood and presence in perceptual phenomenology, as presented in my target article (Seth this collection) and in Seth (2014). Second, he introduces the notion of *causal encapsulation* as a third phenomenological dimension, complementing counterfactual richness and perspective dependence. He spends some time developing examples based on cognitive phenomenology and mental action to illustrate how these dimensions might relate. Here, I will focus on the relationship between presence and objecthood from the perspective of counterfactual predictive processing – or more specifically the theory of " Predictive Processing of SensoriMotor Contingencies" (PPSMC; Seth 2014, 2015b).[5]

4 The continually increasing pressure to justify research in terms of "impact" – especially when seeking funding – highlights one way in which an emphasis on control (rather than discovery) is realized in scientific practice.

5 See also my response (Seth 2015b) to commentaries on (Seth 2014), which focuses on this issue.

3.1 Presence and objecthood together

As Wiese notes, when visually perceiving a real tomato (figure 1A) there is both a sense of *presence* (the subjective sense of reality of the tomato) and of *objecthood* (the perception that a (real) object is the cause of sensations). Importantly, while distinct, these properties are not independent. There is a "world-revealing" dimension to perceptual presence which is closely aligned with the experience of an externally-existing object: "How can it be true … that we are perceptually aware, when we look at a tomato, of the parts of the tomato which, strictly speaking, we do not perceive. This is the puzzle of perceptual presence" (Noë 2006, p. 414).

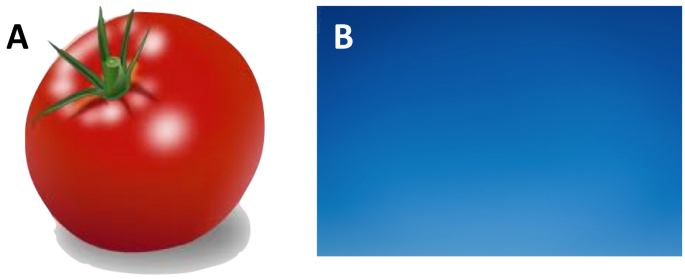

Figure 1: A. An image of a tomato. **B**. An image of a clear blue sky.

How does this object-related world-revealing presence come about? In predictive processing (and by extension PPSMC), objecthood depends on predictive models encoding hierarchically deep invariances that accommodate complex nonlinear mappings from (object-related, world-revealing) hidden causes to sensory signals (Clark 2013; Hohwy 2013). There is a reciprocal dependency here between hierarchical depth and counterfactual richness, because (i) hierarchically deep invariances in generative models enable precise predictions about rich repertoires of counterfactual sensorimotor mappings, and (ii) counterfactual richness can scaffold the acquisition of hierarchically deep invariant predictions. One might even say that hierarchically deep invariances are partly *constituted* by (possibly latent) predictions of counterfactually rich sensorimotor mappings (Seth 2015b). These dependencies indicate that ob-

jecthood and world-revealing presence depend on *expectations about counterfactual richness*, rather than counterfactual richness *per se*. Altogether, counterfactually-informed active inference enables the extraction and encoding of hierarchically deep hidden causes of sensory signals. In virtue of hierarchical depth, these inferred causes will also be *perspective invariant*, in the sense that they will have been separated from those causes that depend on on actions (or other properties) of the perceiver (see Wiese this collection, p. 11). In short, to the extent that objecthood and perceptual presence go together, so do hierarchical depth (encoding world-revealing invariances) and (expected) counterfactual richness.

3.2 Presence and objecthood apart

So far so good, but it is evident that presence and objecthood do not *always* go together (Di Paolo 2014; Froese 2014; Madary 2014), a phenomenological fact which requires further analysis (Seth 2015b). Presence without objecthood is exemplified in vision by the experience of a uniform deep blue sky (Figure 1B), and is also characteristic of non-visual modalities like olfaction (Madary 2014). The visual impression of a blue sky, or the tang of briny sea air, both seem perceptually present but without eliciting any specific phenomenology of objecthood. At the same time, the corresponding predictive models are likely to be hierarchically shallow and counterfactually poor: there is not much I can do (besides closing my eyes or looking away) to alter the sensory input evoking a blue-sky experience, and the inferred hidden causes are unlikely to lie behind multiple inferential layers. Hierarchical shallowness may explain the lack of phenomenal objecthood, but why isn't there also a lack of perceptual presence?

Blue-sky-experiences (and olfactory scenes) actually *do* lack the world-revealing presence associated with objecthood. But they do not appear *phenomenally unreal* in the sense that perceptual afterimages and synaesthetic concurrents are experienced as unreal. In PPSMC, phenomenal unreality can arise from an inferential failure to separate hidden causes in the world, from those that depend on actions (or other properties) of the perceiver (Seth 2015b). This in turn emerges from violations of counterfactual predictions. For example, consider how saccadic eye movements engage counterfactual predictions. Perceptual afterimages track eye movements, violating counterfactual predictions associated with world-revealing hidden causes that rest on active inference. In contrast, counterfactual predictions associated with blue skies are less amenable to disconfirmation by eye movements, so (non-object-related) perceptual presence remains.[6]

Summarizing, perceptual presence, as an explanatory target, can be refined into (i) a *world-revealing presence* associated with objecthood and hierarchical depth, and (ii) a *phenomenal unreality* arising from a failure to inferentially separate hidden causes in the world from those associated with the perceiver. Both rely on counterfactual processing, and so both call on active inference. Perspective invariance is also implicated in objecthood (through hierarchical depth) and phenomenal unreality (through isolating worldy causes), suggesting that this dimension may not be as separable from counterfactual richness as proposed by Wiese (this collection, p. 13). But is that all there is to presence?

3.3 Causal encapsulation and embodiment

Wiese distinguishes three dimensions to perceptual presence: counterfactual richness (vs. poverty), perspective invariance (vs. dependence), and causal encapsulation (vs. integration). The third of these, causal encapsulation, is perhaps the hardest to pin down. The idea as I understand it, is that a representation (predictive model) is causally encapsulated if it is inferentially isolated from other hidden causes; by contrast it is causally *open* or *integrated* if it expresses a rich set of relations to other inferred

6 Phenomenal unreality on this story corresponds to a loss of "transparency" as described by (Metzinger 2003). For Metzinger, transparency is lost – and phenomenal unrealness results – when the "construction process" underlying perception becomes available for attentional processing. This maps neatly on a failure to inferentially unmix world-related from perceiver-related hidden causes – see Seth (2015b) for more on this.

causes. So, a predictive model underlying the experience of a tomato may be causally integrated with that underlying the experience of the table on which it lies, and the hand (maybe my hand), which is poised to reach out and pick it up. Here, there may be a relation between causal encapsulation/integration and the inferential unmixing of perceiver-related and world-related hidden causes: a failure to separate these causes would presumably prevent rich causal integration with other hidden causes in the world.

The concept of causal encapsulation highlights another interesting aspect of Wiese's commentary: the idea that counterfactual predictions may not always encode sensorimotor contingencies: "it might be equally relevant to encode how sensory signals pertaining to the tomato would change if the wind were to blow … or if the tomato were to fall down" (Wiese this collection, p. 11). While such extra-personal causal contingencies may be salient in many cases, I see them as secondary to sensorimotor body-related counterfactual predictions. By definition they do not involve active inference: I have to wait for the wind to change direction (though perhaps I might move to get a better view). This means that many central features of active inference discussed here – its relation to predictive control, homeostasis, and counterfactually-informed model disambiguation – do not apply.

The body re-emerges here as central, this time as a ground for the generation of counterfactual predictions. Specifically, bodily constraints shape counterfactual predictions since they place limits on how actions can be deployed in intervening upon the (inferred) causes of sensory input. This suggests that changing action repertoires would alter experiences of presence. Wiese raises out-of-body-experiences and dream experiences as a relevant context (this collection, p. 15), where subjects sometimes identify their first-person-perspective, not with a body, but with an unextended point in space. I agree with him that examining world-revealing presence in these situations would be fascinating, if extremely difficult in practice.

The body is of course not only a source of counterfactual predictions, but also the target of counterfactually-informed active inference, both for representation (exemplified by the rubber-hand-illusion, as mentioned by Wiese) and for control.[7] As emphasized in the target article, control-oriented active inference is particularly significant for *interoception*, where predictive modelling is geared towards allostasis and homeostasis rather than accurate representation (see also Seth 2013). Returning the focus to interoceptive inference raises a host of intriguing questions, which can only gestured at here. One may straightaway wonder how counterfactual aspects of interoceptive inference shape the "presence" of emotional and body-related experiences. Is it possible to have an emotional experience lacking in "affective presence" – and what is the phenomenological correlate of "objecthood" for interoceptive experience? Other interesting questions are how precision weighting sets the balance between representation versus control in active interoceptive inference, and what it means to isolate "wordly" causes when both the means and the targets of active inference are realized in the body. These are not just theoretical questions: advances in virtual reality (Suzuki et al. 2013) and in methods for measuring interoceptive signals (Hallin & Wu 1998) promise real empirical progress on these issues.

4 Conclusions

This response has been shaped by Wiese's perspicuous focus on the philosophy of science and on the phenomenology of perceptual presence. My response to the first topic was to frame the Bayesian brain in terms of *control-oriented ab-*

7 Wiese, when discussing König's FeelSpace project (Kaspar 2014), interprets PPSMC as saying that increased practice with the FeelSpace compass belt – and hence increased counterfactual richness– would lead to "increased perceptual presence (for the belt, or the vibrations, or the hip/waist, etc.)" (Wiese this collection, p. 17). I see things differently. The counterfactual predictions, while mediated by the belt, relate to hidden causes in the world (e.g., magnetic north). In fact, PPSMC says that FeelSpace practice would lead to hierarchically deep and counterfactually rich models of how "magnetic north" impacts on belt vibrations and the like, leading to increased world-revealing presence for these worldly causes but diminished perceptual presence of the tactile stimulation itself. Still, the FeelSpace project certainly provides a fertile empirical testbed for the ideas raised here.

duction, where falsification is replaced by "inference to the best prediction" as a criterion for progress. I also reinforced the dependency between active inference and counterfactual processing, which underpins the important case of disambiguatory active inference in Bayesian model comparison. With respect to perceptual presence I proposed a distinction between world-revealing presence and phenomenal unreality (Seth 2015b). World-revealing presence corresponds to objecthood and is associated with hierarchical depth, expected counterfactual richness, and perspective invariance of perceptual hypotheses. Phenomenal unreality transpires when perceptual inference fails to unmix world-related from perceiver-related causes; this corresponds to a loss of "phenomenal transparency" (Metzinger 2003) and depends on violation of counterfactual sensorimotor predictions. Space constraints prevented me considering Wiese's discussion of the "presence" of cognitive phenomenology, like abstract mathematical and philosophical thinking, in these terms. There is of course a rich literature in linking such phenomena to the body (Lakoff & Nunez 2001), and hence perhaps to active inference where the concept of a "mental action" becomes critical (O'Brien & Soteriou 2009). Space constraints also prevented Wiese from elaborating on interoception, which I consider the most interesting setting for control-oriented active inference, in virtue of the cybernetics-inspired emphasis on homeostasis and allostasis. Interesting questions emerge here about how counterfactual processing plays into the phenomenology of interoceptive experience.

Cognitive scientists have long argued for a continuity between perception and action (Dewey 1896). To close, I suggest thinking instead of a continuum between *epistemic* and *instrumental* active inference. This is simply the idea that active inference – a continuous process involving both perception and action – can be deployed with an emphasis on predictive control (instrumental), or on revealing the causes of sensory signals (epistemic). This process intertwines interoception, proprioception, and exteroception, and autonomic and motoric action, with the balance always delicately orchestrated by precision optimisation and counterfactual processing. Putting things this way provides a new way to link "life" and "mind" (Godfrey-Smith 1996) and may help reveal the biological imperatives underlying perception, emotion, and selfhood.

Acknowledgements

I am grateful to the Dr. Mortimer and Theresa Sackler Foundation, which support the work of the Sackler Centre for Consciousness Science. Many thanks to Thomas Metzinger, Jennifer Windt and the MIND group for inviting me to participate in this project, to Jakob Hohwy and Karl Friston for correspondence about abductive inference, and to Wanja Wiese for his excellent commentary.

References

Bongard, J., Zykov, V. & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science*, *314* (5802), 1118-1121. 10.1126/science.1133687

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioural and Brain Sciences*, *36* (3), 181-204. 10.1017/S0140525X12000477

Conant, R. & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1* (2), 89-97.

Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review*, *3*, 357-370.

Di Paolo, E. A. (2014). The worldly constituents of perceptual presence. *Frontiers in Psychology*, *5*. 10.3389/fpsyg.2014.00450

FitzGerald, T. H., Dolan, R. J. & Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in Human Neuroscience*, *8*. 10.3389/fnhum.2014.00457

Friston, K. J., Adams, R. A., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology*, *3*. 10.3389/fpsyg.2012.00151

Froese, T. (2014). Steps toward an enactive account of synesthesia. *Cognitive Neuroscience*, *5* (2), 126-127. 10.1080/17588928.2014.905521

Glazebrook, T. (Ed.) (2013). *Heidegger on science.* New York, NY: State University of New York Press.

Godfrey-Smith, P. G. (1996). Spencer and Dewey on life and mind. In M. Boden (Ed.) *The philosophy of artificial life* (pp. 314-331). Oxford, UK: Oxford University Press.

Hallin, R. G. & Wu, G. (1998). Protocol for microneurography with concentric needle electrodes. *Brain Research Protocols*, *2* (2), 120-132.

Hohwy, J. (2013). *The predictive mind.* Oxford, UK: Oxford University Press.

——— (2014). The self-evidencing brain. *Noûs*. 10.1111/nous.12062

Hohwy, J., Roepstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108* (3), 687-701. 10.1016/j.cognition.2008.05.010

Kaspar, K., König, S., Schwandt, J. & König, P. (2014). The experience of new sensorimotor contingencies by sensory augmentation. *Consciousness and Cognition*, *28*. 10.1016/j.concog.2014.06.006

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.) *Criticism and the growth of knowledge* (pp. 91-196). Cambridge, UK: Cambridge University Press.

Lakoff, G. & Nunez, R. (2001). *Where mathematics comes from: How the embodied mind brings mathematics into being.* New York, NY: Basic Books.

Madary, M. (2014). Perceptual presence without counterfactual richness. *Cognitive Neuroscience*, *5* (2), 131-133. 10.1080/17588928.2014.907257

Metzinger, T. (2003). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences*, *2*, 353-393.

Noë, A. (2006). Experience without the head. In T. Gendler & A. Hawthorne (Eds.) *Perceptual experience* (pp. 411-434). New York, NY: Clarendon / Oxford University Press.

O'Brien, L. & Soteriou, M. (Eds.) (2009). *Mental actions.* Oxford, UK: Oxford University Press.

Powers, W. T. (1973). *Behavior: The control of perception.* Hawthorne, NY: Aldine de Gruyter.

Rosa, M. J., Friston, K. J. & Penny, W. (2012). Post-hoc selection of dynamic causal models. *Journal of Neuroscience Methods*, *208* (1), 66-78. 10.1016/j.jneumeth.2012.04.013

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17* (11), 565-573. 10.1016/j.tics.2013.09.007

——— (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, *5* (2), 97-118. 10.1080/17588928.2013.877880

——— (2015). The cybernetic bayesian brain. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

——— (2015b). Presence, objecthood, and the phenomenology of predictive perception. *Cognitive Neuroscience*

Suzuki, K., Garfinkel, S. N., Critchley, H. D. & Seth, A. K. (2013). Multisensory integration across exteroceptive and interoceptive domains modulates self-experience in the rubber-hand illusion. *Neuropsychologia*, *51* (13), 2909-2917. 10.1016/j.neuropsychologia.2013.08.014

Wiese, W. (2015). Perceptual presence in the Kuhnian-Popperian Bayesian brain. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.