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.

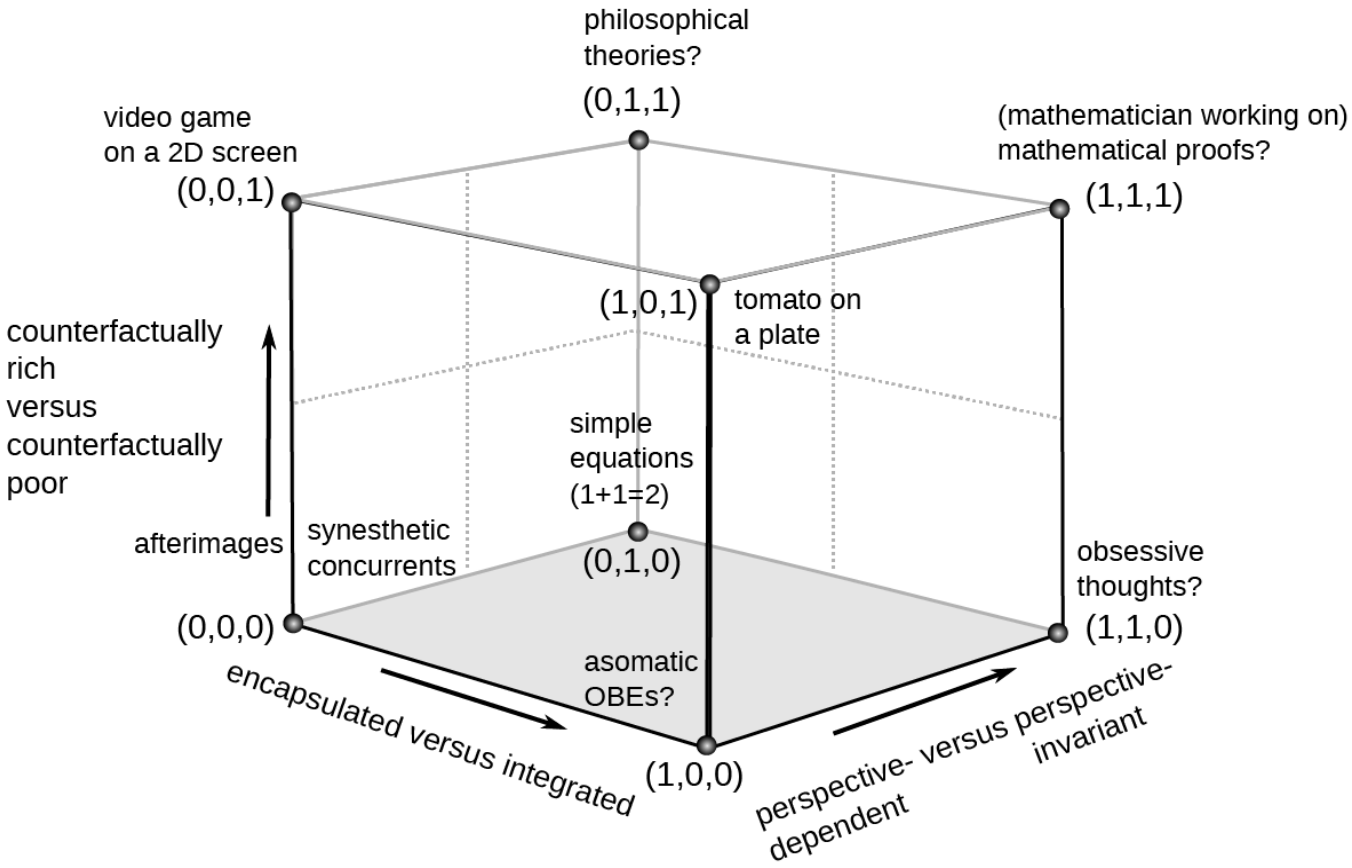

Figure 1: The figure illustrates how classes of experiences can be located in a cube, according to the extent to which they display counterfactual richness, perspective-independence, and causal integration (see main text for explanations). The cube (without the labels) is adapted from cube figures in 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.

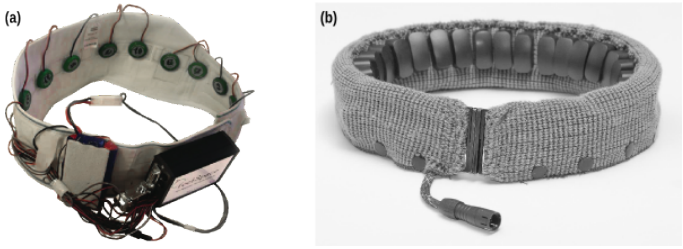

Figure 2: The figure shows two versions of the feelspace belt. (a) The original version used in Nagel et al. (2005). (b) The current version used in Karcher et al. (2012) and Kaspar et al. (2014). Images used with kind permission of Peter König.

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.