# Philosophy and Predictive Processing

# PP

Thomas Metzinger & Wanja Wiese (Eds.), ***Philosophy and Predictive Processing***

Thomas Metzinger & Wanja Wiese (Eds.)

*Philosophy and Predictive Processing*

Frankfurt am Main: MIND Group

*For Anja, Kerstin, and our Families*

**Imprint**



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

Production: Satzweiss.com Print Web Software GmbH, Saarbrücken

ISBN: 978-3-95857-138-9


**www.predictive-mind.net**

# Table of Contents

## Contributions

# Vanilla PP for Philosophers: A Primer on Predictive Processing

Wanja Wiese & Thomas Metzinger

**Keywords**

Active inference | Attention | Bayesian Inference | Environmental seclusion | Free energy principle | Hierarchical processing | Ideomotor principle | Perception | Perceptual inference | Precision | Prediction | Prediction error minimization | Predictive processing | Predictive control | Statistical estimation | Top-down processing

**Acknowledgments**

We are extremely grateful to Regina Fabry and Jakob Hohwy for providing very detailed and helpful feedback on a draft of this primer. Thanks also to Lucy Mayne and Robin Wilson for their valuable editorial help. Our student assistant Fabian Martin Römer deserves a special thanks for his professional and reliable help in correcting the formatting of all articles in this collection.

The goal of this short chapter, aimed at philosophers, is to provide an overview and brief explanation of some central concepts involved in predictive processing (PP). Even those who consider themselves experts on the topic may find it helpful to see how the central terms are used in this collection. To keep things simple, we will first informally define a set of features important to predictive processing, supplemented by some short explanations and an alphabetic glossary.

The features described here are not shared in all PP accounts. Some may not be necessary for an individual model; others may be contested. Indeed, not even all authors of *this collection* will accept all of them. To make this transparent, we have encouraged contributors to indicate briefly which of the features are *necessary* to support the arguments they provide, and which (if any) are *incompatible* with their account. For the sake of clarity, we provide the complete list here, very roughly ordered by how central we take them to be for "Vanilla PP" (i.e., a formulation of predictive processing that will probably be accepted by most researchers working on this topic). More detailed explanations will be given below. Note that these features do not specify individually necessary and jointly sufficient conditions for the application of the concept of "*predictive processing*". All we currently have is a semantic cluster, with perhaps some overlapping sets of jointly sufficient criteria. The framework is still developing, and it is difficult, maybe impossible, to provide theory-neutral explanations of all PP ideas without already introducing strong background assumptions. Nevertheless, at least features 1-7 can be regarded as necessary properties of what is called PP in this volume:

1. **Top-down Processing:** Computation in the brain crucially involves an interplay between top-down and bottom-up processing, and PP emphasizes the relative weighting of top-down and bottom-up signals in both perception and action.
2. **Statistical Estimation:** PP involves computing estimates of random variables. Estimates can be regarded as statistical hypotheses which can serve to explain sensory signals.
3. **Hierarchical Processing:** PP deploys hierarchically organized estimators (which track features at different spatial and temporal scales).
4. **Prediction:** PP exploits the fact that many of the relevant random variables in the hierarchy are predictive of each other.
5. **Prediction Error Minimization (PEM):** PP involves computing prediction errors; these prediction error terms have to be weighted by precision estimates, and a central goal of PP is to minimize precision-weighted prediction errors.
6. **Bayesian Inference:** PP accords with the norms of Bayesian inference: over the long term, prediction error minimization in the hierarchical model will approximate exact Bayesian inference.
7. **Predictive Control:** PP is action-oriented in the sense that the organism can act to change its sensory input to fit with its predictions and thereby minimize prediction error; among other benefits, this enables the organism to regulate its vital parameters (like levels of blood oxygenation, blood sugar, etc.).
8. **Environmental Seclusion:** The organism does not have direct access to the states of its environment and body (for a conceptual analysis of "direct perception", see Snowdon 1992), but infers them (by inferring the hidden causes of interoceptive and exteroceptive sensory signals). Although this is a basic feature of some philosophical accounts of PP (cf. Hohwy 2016; Hohwy 2017), it is controversial (cf. Anderson 2017; Clark 2017; Fabry 2017a; Fabry 2017b).
9. **The Ideomotor Principle:** There are "ideomotor" estimates; computing them underpins both perception and action, because they encode changes in the world which are registered by perception and can be brought about by action.
10. **Attention and Precision:** Attention can be described as the process of optimizing precision estimates.
11. **Hypothesis-Testing:** The computational processes underlying perception, cognition, and action can usefully be described as hypothesis-testing (or the process of accumulating evidence for the internal model). Conceptually, we can distinguish between passive and active hypothesis-testing (and one might try to match active hypothesis-testing with action, and passive hypothesis-testing with perception). It may however turn out that all hypothesis-testing in the brain (if it makes sense to say that) is active hypothesis-testing.
12. **The Free Energy Principle:** Fundamentally, PP is just a way of minimizing free energy, which on most PP accounts would amount to the long-term average of prediction error.

In the following, we do not assume any familiarity with PP or any mathematical background knowledge, and this introduction will, for the most part, be restricted to the conceptual basics of the PP framework. Having read this primer, one should be able to follow the discussion in the other papers of this collection. However, we would also strongly encourage readers to deepen their understanding of PP by reading (Clark 2016) and (Hohwy 2013), two excellent first philosophical monographs on this topic.

Wiese, W. & Metzinger T. (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 1. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573024

## 1 What Is Predictive Processing? Seven Core Features

Predictive processing (PP) is a framework involving a general computational principle which can be applied to describe perception, action, cognition, and their relationships in a single, conceptually unified manner. It is not directly a theory about the underlying neural processes (it is computational, not neurophysiological), but there are more or less specific proposals of how predictive processing can be implemented by the brain (see, e.g., Engel et al. 2001; Friston 2005; Wacongne et al. 2011; Bastos et al. 2012; Brodski et al. 2015). Moreover, it seems that at least some of the principles which can be applied to descriptions on subpersonal (e.g., computational or neurobiological) levels of analysis can also be applied to descriptions on the personal level (e.g., to agentive phenomena, the structure of reasoning, or phenomenological reports which describe the contents of consciousness). This is one reason why PP is philosophically interesting and relevant. If the theory is on the right track, then:

1. it may provide the means to build new conceptual bridges between theoretical and empirical work on cognition and consciousness,

2. it may reveal unexpected relationships between seemingly disparate phenomena, and

3. it may unify to some extent different theoretical approaches.

But what is PP in the first place? Here is a relatively early formulation of one of its key ideas:[1]

> Wenn die Anschauung sich nach der Beschaffenheit der Gegenstände richten müßte, so sehe ich nicht ein, wie man a priori von ihr etwas wissen könne; richtet sich aber der Gegenstand (als Objekt der Sinne) nach der Beschaffenheit unseres Anschauungsvermögens, so kann ich mir diese Möglichkeit ganz wohl vorstellen. (Kant 1998[1781/87], B XVII)[2]

One thing Kant emphasizes at this point in the *Critique of Pure Reason* is that our intuitions (*Anschauungen*), which constitute the sensory material on which acts of synthesis are performed, are not sense-data that are simply given (cf. Brook 2013, § 3.2). They are not just received, but are also partly shaped by the faculty of intuition (*Anschauungsvermögen*). In contemporary parlance, the idea can be expressed as follows:

> Classical theories of sensory processing view the brain as a passive, stimulus-driven device. By contrast, more recent approaches emphasize the constructive nature of perception, viewing it as an active and highly selective process. Indeed, there is ample evidence that the processing of stimuli is controlled by top-down influences that strongly shape the intrinsic dynamics of thalamocortical networks and constantly create predictions about forthcoming sensory events. (Engel et al. 2001, p. 704)

1 At this point, one might have expected a reference to Helmholtz' famous idea that perception is the result of unconscious inferences — we will refer to this passage below. Helmholtz' view on perception was heavily influenced by Kant (although Helmholtz seems to have emphasized the role of learning and experience more than Kant, see Lenoir 2006, pp. 201 & 203): "Dass die Art unserer Wahrnehmungen ebenso sehr durch die Natur unserer Sinne, wie durch die äusseren Dinge bedingt sei, wird durch die angeführten Thatsachen sehr augenscheinlich an das Licht gestellt, und ist für die Theorie unseres Erkenntnissvermögens von der höchsten Wichtigkeit. Gerade dasselbe, was in neuerer Zeit die Physiologie der Sinne auf dem Wege der Erfahrung nachgewiesen hat, suchte Kant schon früher für die Vorstellungen des menschlichen Geistes überhaupt zu thun, indem er den Antheil darlegte, welchen die besonderen eingeborenen Gesetze des Geistes, gleichsam die Organisation des Geistes, an unseren Vorstellungen haben." (Von Helmholtz 1855, p. 19). (Our translation: "These facts clearly show that the nature of our perceptions is as much constrained by the nature of our senses as by external objects. This is of utmost importance for a theory of our epistemic faculty. The physiology of the senses has recently demonstrated, by way of experience, exactly the same point that Kant earlier tried to show for the ideas of the human mind in general, by expounding the contribution made by the special innate laws of the mind — the organization of the mind, as it were — to our ideas.") An overview of PP's Kantian roots can be found in Swanson 2016.

2 "If intuition has to conform to the constitution of the objects, then I do not see how we can know anything of them a priori; but if the object (as an object of the senses) conforms to the constitution of our faculty of intuition, then I can very well represent this possibility to myself." (Kant 1998, B xvii)

This is what we here call the first feature of predictive processing: **Top-Down Processing**. As can be seen, the idea that perception is partly driven by top-down processes is not new (which is not to deny that dominant theories of perception have for a long time marginalized their role). The novel contribution of PP is that it puts an extreme emphasis on this idea, depicting the influence of top-down processing and prior knowledge as a *pervasive* feature of perception, which is not only present in cases in which the sensory input is noisy or ambiguous, but *all the time*.[3] According to PP, one's brain constantly forms statistical estimates, which function as representations[4] of what is currently out there in the world (feature #2, **Statistical Estimation**), and these estimates are hierarchically organized (tracking features at different spatial and temporal scales; feature #3, **Hierarchical Processing**).[5] The brain uses these representations to predict current (and future) sensory input and the source of it, which is possible because estimates at different levels of the hierarchy are *predictive* of each other (feature #4, **Prediction**). Mismatches between predictions and actual sensory input are not used passively to form percepts, but only to inform *updates* of representations which have already been created (thereby anticipating, to the extent possible, incoming sensory signals). The goal of these updates is to *minimize* the *prediction error* resulting from the prediction (feature #5, **Prediction Error Minimization (PEM)**), in such a way that updates conform to the norms of **Bayesian Inference** (feature #6; more on this below). The computational principle of PEM is a general principle to which all processing in the brain conforms (at all levels of the hierarchy posited by PP). From this, it is only a small step towards describing processing in the brain as a controlled online hallucination:[6]

> [A] fruitful way of looking at the human brain, therefore, is as a system which, even in ordinary waking states, constantly hallucinates at the world, as a system that constantly lets its internal autonomous simulational dynamics collide with the ongoing flow of sensory input, vigorously dreaming at the world and thereby generating the content of phenomenal experience. (Metzinger 2004[2003], p. 52)

Note that the contents of phenomenal experience are only part of what is, according to PP, generated through the hierarchically organized process of prediction error minimization (most contents will be unconscious). Summing up the first six core features described above, and adding the seventh feature, we can now give a concise definition of what is called predictive processing in this collection (we will enrich the definition with features 8-12 below):

3 Of course, it is an interesting question to what extent Kant himself saw active (top-down) influences on intuitions (*Anschauungen*) as a pervasive feature. At least some passages in the *Critique of Pure Reason* suggest that Kant laid more emphasis on the (top-down) influences exerted by our faculty of *cognizing* (the spontaneity of *concepts*):
"Unsere Erkenntnis entspringt aus zwei Grundquellen des Gemüts, deren die erste ist, die Vorstellungen zu empfangen (die Rezeptivität der Eindrücke), die zweite das Vermögen, durch diese Vorstellungen einen Gegenstand zu erkennen (Spontaneität der Begriffe); durch die erstere wird uns ein Gegenstand gegeben, durch die zweite wird dieser im Verhältnis auf jene Vorstellung (als bloße Bestimmung des Gemüts) gedacht." (Kant 1998[1781/87], B 74).
"Our cognition arises from two fundamental sources in the mind, the first of which is the reception of representations (the receptivity of impressions), the second the faculty for cognizing an object by means of these representations (spontaneity of concepts); through the former an object is given to us, through the latter it is thought in relation to that representation (as a mere determination of the mind)." (Kant 1998, B 74). However, a serious investigation of this question would have to focus on the influence of *unconscious* representations on the forming of intuitions (see Giordanetti et al. 2012).

4 The use of the word "representation" is not completely uncontroversial here. There is some debate about whether PP posits representations, and if so how best to describe them (see Gładziejewski 2016; Downey 2017; Dołega 2017). It is at least possible, however, to treat representationalist descriptions of the posits entailed by PP as a representational (or intentional) gloss (cf. Egan 2014; Anderson 2017). So, although we acknowledge that some would disagree, we believe it is useful to describe the estimates posited by PP as representations, at least for the purposes of this primer (even if some authors would argue that these posits are not representations in a strong sense).

5 This hierarchy of estimates entails a hierarchical generative model. A generative model is the joint distribution of a collection of random variables (see glossary). A hierarchical generative model corresponds to a hierarchy of random variables, where variables at non-adjacent levels are conditionally independent (this can, for instance, represent a hierarchy of causally related objects or events, see Drayson 2017). The hierarchy of estimates posited by PP tracks the values of a hierarchy of random variables. A heuristic illustration of a generative model can be found in the introduction to (Clark 2016). We are grateful to Chris Burr for reminding us to mention generative models.

6 Horn (Horn 1980, p. 373) ascribes the idea that "vision is a controlled hallucination" to Clowes (Clowes1971). The only published statement by Clowes which comes near this formulation seems however to be: "People see what they expect to see" (Clowes 1969, p. 379; cf. Sloman 1984). More recently, a similar idea has been put forward by Grush (Grush 2004, p. 395; he ascribes it to Ramesh Jain): "The role played by sensation is to constrain the configuration and evolution of this representation. In motto form, perception is a controlled hallucination process."

Predictive Processing (PP) is

- *hierarchical* predictive coding,
- involving *precison-mediated*
- prediction error minimization,[7]
- enabling predictive *control*.

Note that this definition already goes beyond what is often referred to as *predictive coding* (especially if predictive coding is just conceived as a computational strategy for data compression, cf. Shi and Sun 1999; Clark 2013a). Firstly, PP is hierarchical. Secondly, precision estimates can fulfil functional roles that go beyond just balancing prior assumptions and current sensory evidence in statistically optimal fashions (see Clark 2013b). Thirdly, PP is often described as action-oriented, in the sense that it enables **Predictive Control** (feature #7; cf. Seth 2015). This highlights the assumption, held by some, that action is in some sense more important than perception; although perception can be described as a process of gaining knowledge about the world, the main function of gaining this knowledge lies in enabling efficient, context-sensitive action, through which the organism successfully sustains its existence. This becomes evident when PP is considered in the wider context of Friston's free-energy principle (FEP).[8] Before elaborating on this, let us step back and take a look at the problem of perception, viewed from the perspective of predictive coding.

## 2 Predictive Processing and Predictive Coding

As for almost all features of PP (predictive processing), there are also prominent precursors of the PP view on perception. Consider the following statement by Helmholtz:

> Die psychischen Thätigkeiten, durch welche wir zu dem Urtheile kommen, dass ein bestimmtes Object von bestimmter Beschaffenheit an einem bestimmten Orte ausser uns vorhanden sei, sind im Allgemeinen nicht bewusste Thätigkeiten, sondern unbewusste. Sie sind in ihrem Resultate einem Schlusse gleich, insofern wir aus der beobachteten Wirkung auf unsere Sinne die Vorstellung von einer Ursache dieser Wirkung gewinnen, während wir in der That direct doch immer nur die Nervenerregungen, also die Wirkungen wahrnehmen können, niemals die äusseren Objecte. (Von Helmholtz 1867, p. 430)[9]

The problem of perception, as conceived here, has two aspects: (1) percepts are the result of an unconscious inferential process; (2) percepts present us with properties of external objects, although in fact we can only perceive the effects of external objects. A contemporary description of this idea can be found in Dennett's 2013 monograph, *Intuition Pumps and Other Tools for Thinking*. He characterizes the curious situation in which the brain finds itself, by likening it to the following fictional scenario:

> You are imprisoned in the control room of a giant robot. […] The robot inhabits a dangerous world, with many risks and opportunities. Its future lies in your hands, and so, of course, your own future as well depends on how successful you are in piloting your robot through the world. If it is destroyed, the electricity in this room will go out, there will be no more food in the fridge, and you will die. Good luck! (Dennett 2013, p. 102)

7 The first three parts of this definition correspond roughly with the definition offered by Clark (Clark 2013a, p. 202; Clark 2015, p. 5). In (Clark 2013a), Clark also introduces the notion of action-oriented PP (which incorporates the fourth aspect of the definition offered here). These four features are central too to Hohwy's exposition of prediction error minimization (see the first four chapters in Hohwy 2013).

8 More on this below. Note that it is possible to develop PP accounts without invoking FEP (so in a way PP is independent of FEP), but PP can be incorporated into FEP (see Friston and Kiebel 2009), so prediction error minimization can be construed as a way of minimizing free energy (which would then be a special case of FEP).

9 "The psychic activities that lead us to infer that there in front of us at a certain place there is a certain object of a certain character, are generally not conscious activities, but unconscious ones. In their result they are equivalent to a conclusion, to the extent that the observed action on our senses enables us to form an idea as to the possible cause of this action; although, as a matter of fact, it is invariably simply the nervous stimulations that are perceived directly, that is, the actions, but never the external objects themselves." (Von Helmholtz 1985[1925], p. 4).

The person inside the robot has only indirect access to the world, via the robot's sensors, and the effects of executed actions cannot be known but have to be inferred. This illustrates the feature we call **Environmental Seclusion** (feature #8). Environmental Seclusion is not a computational feature but an epistemological one, yet it appears in descriptions of the problems to which PP computations provide a solution.[10] To find out what the different signals received by the robot mean, the person inside has to form a hypothesis about their hidden causes. The problem of inferring the causes of sensory signals is an *inverse problem*, because it requires inverting the mapping from (external, hidden) causes to (sensory) effects. This is a difficult problem (to say the least), because the same effect could have multiple causes.[11] So even if the relationship between causes $C$ and effects $E$ could be described by a deterministic mapping, $f: C \rightarrow E$, the inverse mapping, $f^{-1}: E \rightarrow C$, would not usually exist. How does the brain solve this problem?

A first observation is that the cause of a sensory effect is underdetermined by the effect, so prior information has to be used to make a good guess about the hidden cause. Furthermore, if we know how the sensory apparatus is affected by external causes, it is easier to infer sensory effects, given information about external causes, than the other way around. So if we have some information about hidden causes, we can form a *prediction* of their sensory effects. This prediction can be compared to the actual sensory signal, and the extent to which the two differ from each other, i.e., the size of the *prediction error* gives us a hint as to the quality of our estimate of the hidden cause. We can update this estimate, compute a new prediction, again compare it with current sensory signals, and thereby (hopefully) minimize the prediction error. Ideally, it does not hurt if our first estimate of the hidden cause is really poor, as by constantly computing predictions and prediction errors, and by updating our estimate accordingly, we can become more and more confident that we have found a good representation of the hidden cause.

Let us illustrate this strategy with the following simple example. A teacher enters the classroom and finds a piece of paper on his desk, with the message "The teacher is an impostor. He doesn't even really exist." The message has been written with a fountain pen, in blue, which (as the teacher knows) excludes many of his students. To find the culprit, the teacher asks all students using fountain pens with blue ink to come to the front and, using their own pen, to write something on a piece of paper. As it turns out, this involves only three students, A, B, & C, and all use ink of different brands (which makes them distinguishable). The teacher can now form an educated guess about the hidden cause of the message ("The teacher is an impostor. He doesn't even really exist."): he assumes that student A is the culprit, and asks A to write down the same message. This can be seen as a prediction of the actual message, and by comparing them the teacher evaluates his guess about its hidden cause. If the ink looks the same there is no prediction error, and the estimate of the hidden cause does not have to be changed — the culprit has been found. If there is a difference, he can update his estimate, by assuming that, say, B has produced the message. By constantly forming predictions (messages written by the suspects) and comparing them with the actual sensory signal (the message on the desk), the teacher eventually minimizes the prediction error and finds the true culprit.

There are a lot of differences between this fictive scenario and the situation in which the brain finds itself. One is that the example involves personal-level agency (just like Dennett's giant robot thought

10 Here are some examples: "For example, during visual perception the brain has access to information, measured by the eyes, about the spatial distribution of the intensity and wavelength of the incident light. From this information the brain needs to infer the arrangement of objects (the causes) that gave rise to the perceived image (the outcome of the image formation process)." (Spratling 2016, p. 1 preprint).
"The first of these (the widespread, top-down use of probabilistic generative models for perception and action) constitutes a very substantial, but admittedly quite abstract, proposal: namely, that perception and […] action both depend upon a form of 'analysis by synthesis' in which observed sensory data is explained by finding the set of hidden causes that are the best candidates for having generated that sensory data in the first place." (Clark 2013a, p. 234; but see Clark in press, for a qualified view).
"Similarly, the starting point for the prediction error account of unity is one of indirectness: from inside the skull the brain has to infer the hidden causes of its sensory input" (Hohwy 2013, p. 220).

11 For this reason, the problem can also be described as an ill-posed problem (see Spratling 2016), but some authors would regard the problem of finding out how to solve this problem as ill-posed (see Anderson 2017).

experiment): the teacher tests the hypothesis that, say, student A is the culprit by asking A to write down a message. Furthermore, the number of possible hidden causes is finite, and the "prediction error" tells the teacher only that a particular student is not implicated. It does not contain any further information about the culprit; it only excludes one of the suspects. The brain cannot go through all possible hypotheses one by one, because there are (potentially) inifinite possible hidden causes in the world. Furthermore, the world is changing, so representations of hidden causes have to be dynamic: adapting to, and anticipating, all (relevant and predictable) changes in the environment. Finally, to be more realistic, the teacher example would have be extended such that the teacher forms predictions about *all* his sensory inputs *all* the time. Just as he could infer the causal interactions leading up to the note, he can infer all the causal goings-on around him all the time (including his own influence on the sensory stream).

## 3 Predictive Processing and Bayesian Inference

**Bayesian Inference** (feature #6) is a computational method to rationally[12] combine existing information, about which there is uncertainty, with new evidence. Here, uncertainty means that the information can be described in a probabilistic format, i.e., using a probability distribution. A very simple example would be a situation in which an agent is uncertain which of a finite number of hypotheses is true (as with the teacher above). Uncertainty would then be reflected by the fact that the agent assigns different probabilities to the hypotheses, without assigning a probability of 1 to any of them. But there can also be situations in which the agent's information is best modeled as being about an infinite number of possibilities ("hypotheses"), for instance, when the agent performs a noisy measurement of a quantity which could have any value in a continuous interval. In such a case, the information can be coded using a probability density function (i.e., a model), which assigns probabilities to regions (e.g., to sub-intervals). The question to which Bayesian inference provides a rational answer (using Bayes' rule) is the following: how should I update my model when I obtain new information? An example of new information would be information which an agent receives by performing a measurement (assuming the agent already has uncertain information about the quantity to be measured).

Formally, this update involves computing an *a posteriori distribution* (which is also just called the *posterior*). The posterior is obtained by combining an *a priori distribution* (also just called the *prior*) with a *likelihood*. The prior codes the information the agent already has; the likelihood codes how the domain about which the agent already has information is related to the domain of the new information obtained. A nice feature of Bayesian inference is that it can reduce uncertainty. Formally, this means that the posterior often has a lower variance — is more precise — than the prior.

Superficially speaking, there is no obvious connection between prediction error minimization (PEM) and Bayesian inference. In fact, it is not obvious why it would even be desirable to implement or approximate Bayesian inference using PEM. Nevertheless, there is one good reason. Recall that the inverse problem of perception is an ill-posed problem: sensory signals, considered as the effects of external events, cannot be mapped to the hidden states of the environment because for every sensory effect there are multiple possible external causes. In other words, there is uncertainty about the hidden causes. Given prior assumptions about these causes, and considering the sensory effects we measure as new evidence, Bayesian inference promises to give us a rational solution to the problem of how we should update our prior assumptions about hidden causes. In other words, what Bayesian inference can give us (at least in principle) is something like a "probabilistic inverse mapping". This function maps a measured sensory effect to the different (sets of) possible hidden causes, and indicates which possible causes are most likely the actual causes of sensory effects.

12 Here, "rationally" means in accordance with the axioms of probability, and with the definition of conditional probability; it can also be shown that Bayesian inference is optimal in an information-theoretic sense (see Zellner 1988).

But why do we need PEM if we have Bayesian inference? The answer is that Bayesian inference can be computationally complex, even intractable. In simple cases, it is possible to compute the posterior analytically; in other cases, it has to be approximated. In yet other cases, it may be possible to compute the posterior, but what one would really like to have is the *maximizer* of the posterior (for instance, a single hypothesis that is most likely, after having taken the new evidence into account). Finding the maximizer may again be computationally demanding and can require approximative methods. Some approximative methods involve prediction error minimization. While the motivation for Bayesian inference is independent of prediction error minimization, once Bayesian inference is regarded as a solution to the problem of perception, prediction error minimization can provide a solution to the problem of computing Bayesian updates.

Note that Bayesian inference also works for hierarchical models. Assuming that variables at non-adjacent levels in the hierarchy are conditionally independent, estimates can be updated in parallel at the different levels (cf. Friston 2003, p. 1342), which ideally yields a globally consistent set of estimates (in practice, things are complicated, as Lee and Mumford 2003, p. 1437, point out). Here, the idea is that most objects in the world do not directly influence each other causally, but they are still objects in the *same* world, which means that causal interactions between two arbitrary objects are usually *mediated* by other objects. Similarly, different features of a single object are not completely independent, because they are features of the *same* object, but this does not mean representations of these features must always be jointly processed. For instance, a blue disc can be represented by representing a certain color (blue) at a certain place, and a certain shape (a disc) at the same place. Information about the location of the color gives me information about the location of the shape. If I have a separate representation of the disc's location, however, I can treat the color and the shape as (conditionally) independent, i.e., given the disc's location, information about the color does not give me new information about the shape. Computationally, this allows for sparser representations, which may also be reflected by functional segregation in the brain (cf. Friston and Buzsáki 2016, who explore this with a focus on the temporal domain).

## 4 Predictive Processing and the Ideomotor Principle

So far, prediction error minimization has only been described as a way of generating percepts in accordance with sensory input. The primary role of prediction error minimization may not however be to infer hidden causes in the world, but to bring about changes in the world that help the agent stay alive (see section 7 below). Moreover, the primary target for such changes may not be the external but the internal environment of the agent, i.e., its body. In biological systems, organismic integrity is a top-level priority, because a stable organism (which can control its internal states) can survive in different environments, whereas an unstable organism may not survive even in friendly environments. This has been pointed out by Anil Seth:

> 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 […]. (Seth 2015, p. 9)

Clearly, the goal of interoceptive inference is not simply to infer the internal condition of the body, but to enable *predictive control* of vital parameters like blood oxygenation or blood sugar (feature #7). Seth

provides the following example. When the brain detects a decline in blood sugar through interoceptive inference, the resulting percept (a craving for sugary things) will lead to 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). (Seth 2015, p. 10)

Interoceptive prediction error minimization is therefore an illustrative example of how perception and action are coupled, according to PP. A goal of interoceptive PEM is to keep the organism's vital parameters (such as its blood sugar level etc.) within viable bounds, and this involves both accurately inferring the current state of these parameters and actively changing them (when necessary). Here is how Friston puts it (in terms of minimizing free energy, which under certain assumptions entails minimizing prediction error):

> Agents can suppress free energy by changing the two things it depends on: they can change sensory input by acting on the world or they can change their recognition density by changing their internal states. This distinction maps nicely onto action and perception […]. (Friston 2010, p. 129)

In short, the error between sensory signals and predictions of sensory signals (derived from internal estimates) can be minimized by changing internal estimates and by changing sensory signals (through action). What this suggests is that the same internal representations which become active in perception can also be deployed to enable action. This means that there is not only a common data-format, but also that at least some of the representations that underpin perception are numerically identical with representations that underpin action.

This assumption is already present in James' *ideomotor theory* (James 1890),[13] the core of which is summed up as follows by James: "[T]he idea of the movement M's sensory effects will have become an immediately antecedent condition to the production of the movement itself." (James 1890, p. 586; italics omitted). More recently, this has been picked up by *common coding* accounts of action representation (see Hommel et al. 2001; Hommel 2015; Prinz 1990).[14] The basic idea is always similar: The neural representations of hidden causes in the world overlap with the neural underpinnings of action preparation (which means parts of them are numerically identical). In other words, there are "ideomotor" representations, which can function both as percepts and as motor commands.[15]

Computationally, the **Ideomotor Principle** (feature #9) exploits a formal duality between action and perception. The duality is this: If I can perceptually access a state of affairs *p*, this means *p* has some perceivable consequences (or constituents) *c*; action is goal-oriented, so by performing an action I want to bring about some state of affairs *p*. This means action can also be described as a process in which the perceivable consequences *c* of *p* are brought about, and perception can be described as the process by which the causes of a hypothetical action (which brings about *p,* and thereby *c*) are inferred

13 Another precursor of the idea can be found in the works of Herbart (Herbart 1825, pp. 464 f.) and Lotze (Lotze 1852, pp. 313 f.).

14 A review of ideomotor approaches can be found in (Badets et al. 2014). A historical overview can be found in (Stock and Stock 2004).

15 Strictly speaking, ideomotor representations are sometimes just regarded as late (high-level) contributions to perception, and as the (early) precursors of action (in the following quotation, "TEC" denotes the theory of event coding (TEC)): "TEC does not consider the complex machinery of the 'early' sensory processes that lead to them. Conversely, as regards action, the focus is on 'early' cognitive antecedents of action that stand for, or represent, certain features of events that are to be generated in the environment (= actions). TEC does not consider the complex machinery of the 'late' motor processes that subserve their realization (i.e., the control and coordination of movements). Thus, TEC is meant to provide a framework for understanding linkages between (late) perception and (early) action, or action planning." (Hommel et al. 2001, p. 849)

(for a rigorous description of this idea, see Todorov 2009). The computational benefits of this dual perspective are reaped in the notion of *active inference* (developed by Friston and colleagues):

> In this picture of the brain, neurons represent both cause and consequence: They encode conditional expectations about hidden states in the world causing sensory data, while at the same time causing those states vicariously through action. [...] In short, active inference induces a circular causality that destroys conventional distinctions between sensory (consequence) and motor (cause) representations. This means that optimizing representations corresponds to perception or intention, i.e. forming percepts or intents. (Friston et al. 2011, p. 138) [16]

*Active* inference is often distinguished from *perceptual* inference. Since both are realized by minimizing prediction error, however, and since their implementations may not be neatly separable, *active inference* is also used as a more generic term, especially by Friston and colleagues. In the context of the free-energy principle (see below), it denotes the computational processes which minimize free energy and underpin both action and perception: "Active inference — the minimisation of free energy through changing internal states (perception) and sensory states by acting on the world (action)." (Friston et al. 2012a, p. 539).[17]

Common to both action and perception is (unconscious, approximatively Bayesian) inference. Since neural structures underpinning action and perception, respectively, are assumed to overlap, active and perceptual inference work in tandem.[18] This updated and enriched version of the **Ideomotor Principle** thereby provides a unifying perspective on action and perception, while its deeper implications and challenges are only beginning to be explored.[19]

## 5 Attention and Precision

One of the many fruitful ideas formulated in the PP framework is that the allocation of attention can be analyzed as the process of optimizing precision estimates (feature #10). This was first put forward by Karl Friston and Klaas Stephan (Friston and Stephan 2007) (two important papers extending this idea are Feldman and Friston 2010 and Hohwy 2012). Since precision estimates function as weightings of prediction error terms, the precision associated with a prediction error influences its impact on processing at other levels. This means that increasing estimated precision can enhance the depth of processing of a stimulus. Furthermore, precision estimates can be changed in a bottom-up and a top-down fashion: Bottom up, precision can be estimated as a function of obtained samples (e.g., as the inverse of the sample variance); top down, precision estimates can be modulated in contexts in which increases or decreases of precision are anticipated, or can function as goal-representations for mental action (Metzinger 2017). The difference between bottom-up and top-down changes in precision estimates can be linked to the difference between endogenous and exogenous attention (for details, see Feldman and Friston 2010 and Hohwy 2012).

Using this precision-optimization account of attention, it is possible to draw a connection between action and attention. Recall that according to the ideomotor principle, some neural structures are part of the neural underpinnings of both action and perception. Assume that neural structure *N* becomes active when I perceive a person scratching her chin and when I myself am about to scratch my chin.

16 This resonates with the "principle of reafference" (*Reafferenzprinzip*) of Holst and Mittelstaedt (Von Holst and Mittelstaedt 1950), which also stresses that the neural events accompanying perception can not only be regarded as effects of sensory signals but also as their causes, because they can influence sensory signals (through action).

17 See also (Clark 2016, p. 181) and (Burr 2017).

18 The same idea is exploited in recent work by Lake, Salakhutdinov, and Tenenbaum on concept learning: the system recognizes visual characters by inferring a "probabilistic program", which is a generative model that can be used to generate the visual input (cf. Lake et al. 2015, p. 1333).

19 For instance, (Wiese 2016) argues that, if the PP version of the ideomotor theory is on the right track, action is enabled by systematic misrepresentations; (Colombo 2017) argues that PP challenges the Humean theory of motivation, in that appeals to desire and value may not be necessary to account for social motivation, while minimizing social uncertainty may.

Following Friston et al. (Friston et al. 2011, p. 138), *N* could function both as a percept and as an intent, though it usually only functions as one of them. So, unless I suffer from echopraxia, perceiving a movement will not usually cause me to move in the same way (although there are situations in which persons do mimic each other to some extent, see Quadt 2017). This can be accounted for within the framework of PP by noting the following: the hypothesis that I am scratching my chin will yield proprioceptive and other sensory predictions (which describe, for example, the states of my muscles when my arm is moving). Unless I am in fact scratching my chin, these predictions will be in conflict with sensory signals, so there will be a large prediction error, which will lead to an update of the hypothesis that I am scratching my chin. In other words, the hypothesis cannot be sustained in the presence of such prediction errors. So to enable movement, precision estimates associated with sensory prediction errors must be cancelled out by top-down modulation. Combining this with the hypothesis that attention increases precision estimates, one could describe this as a process of *attending away* from somatosensory signals. Conversely, attending to sensory stimuli should impair normal movement (see Limanowski 2017).

This connection between action and attention is also exploited in accounts of self-tickling (see Van Doorn et al. 2014; Van Doorn et al. 2015). Deviances in precision estimates have been linked to attention and motor disorders, and raised in the context of autism and schizophrenia (see Gonzalez-Gadea et al. 2015; Palmer et al. 2015; Van de Cruys et al. 2014; Friston et al. 2014; Adams et al. 2016). This is only one example of how the PP approach may possess great heuristic fecundity and explanatory power for cognitive neuropsychiatry and related fields.

## 6 The Brain as a Hypothesis-Tester

We mentioned above that a person trapped in a giant robot (recall Dennett's thought experiment) has to form hypotheses about their environment. One reason why PP may seem appealing to some, if dubious to others, is that it applies personal-level descriptions of this kind to the computational level of description (cf. the paper "The hypothesis testing brain", Hohwy 2010, feature #11). However, such descriptions can at least be heuristically fruitful in trying to answer the question of why we perceive the world the way we do, in particular, the question of what the formal principles of perceptual organization are. Richard Gregory's classic paper "Perceptions as hypotheses" famously probes the idea that percepts *explain* sensory signals, and that they have *predictive power* (Gregory 1980, pp. 182, 186). Helmholtz already suggested an extended version of this idea, namely that movements could be regarded as experiments:

> [W]ir 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 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. (Von Helmholtz 1959[1879/1887], p. 39)[20]

A classical application in the present debate is saccadic eye movements, which are now conceptualized as an embodied form of hypothesis-testing (Friston et al. 2012b). Apart from these heuristic considerations, if PP is on the right track we can ask if the brain *literally* engages in inference. This question is answered in the affirmative by Alex Kiefer in his contribution to this collection (see Kiefer 2017). A

20 "We observe amid our own continuous 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 voluntary movements through which we change the appearance of things should be regarded as 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." (Own translation)

skeptical position is maintained by Jelle Bruineberg (see Bruineberg 2017 and Bruineberg et al. 2016). The more general issue of how folk psychology and PP are related, and to what extent the scientific usage of folk-psychological concepts may need to be revised, is discussed by Joe Dewhurst (see Dewhurst 2017).

## 7 Predictive Processing and Karl Friston's Free-Energy Principle

Consider the following tautology: Every organism which manages to stay alive for a certain time does not die during that time. Furthermore, staying alive entails running the risk of dying. This is not supposed to be a deep insight into the concept of life, but a seemingly superficial remark about living organisms. It nonetheless has some interesting implications. For every living organism, there are deadly situations which the organism must avoid to stay alive; and perhaps the more sophisticated an organism is, the more potentially deadly situations there are. Just think about the environments in which a bacterium can survive and compare them with those in which a human being can do so. If an organism has managed to stay alive for a certain time, this means it has (thus far) avoided any deadly situations.

If we make a list of possible situations in which an organism *could* find itself, and compare this list with the possible situations in which the organism is able to *survive*, we will find two things:

1. for most organisms (e.g., human beings), the second list will be drastically shorter than the first (because there are a lot of deadly situations); and

2. if we observe an organism which is capable of surviving for a decent amount of time, at a random point during its lifetime, it is very likely to be found in a situation from the second list (this echoes the tautology at the beginning of this section).

We can re-express these two observations in a slightly more technical way. Let us call the set of all possible states in which an organism could be its *state space*, where a state is defined by the current sensory signals received by the organism's sensory system. In principle, we can now define a probability distribution over this state space which assigns probabilities to the different regions in this space and describes how likely it is to find the organism in the respective regions during its lifetime. Certain regions will have a high probability (e.g., a fish is likely to be found in water); others will have a low probability (a fish is unlikely to be found outside of water). Furthermore, *most* regions of state space will have a low probability (because there are so many deadly situations). Formally, this means that the *entropy* of the probability distribution is low (it would be maximal if it assigned probabilities uniformly to the different regions of state space; see below for a simple formal example). With this probability distribution in hand, we can now make a bet on where in its state space the organism will be found, if observed at an arbitrary moment during its lifetime. Since the distribution assigns extremely low probabilities to most regions of state space, we can make a fairly precise guess (e.g., we can guess that a fish will be in water, that a freshwater fish will be in fresh water, and so on).

Now consider the following. Throughout the lifetime of the fish, we take repeated samples of its states and construct an *empirical distribution* using these samples. An empirical distribution assigns probabilities which reflect the *frequency* with which samples were (randomly) drawn from the different regions. As a simple example, think of a device which produces one of two numbers, 0 and 1, whenever a button is pushed, and the two numbers are produced with certain probabilities unknown to the agent. It could be that both numbers are produced with the same probability (0.5), or that one is produced much more frequently than the other (say, 0 is produced with probability 0.9, and 1 is produced with probability 0.1). Every time one presses the button, one notes which number has been produced (this is a single sample), and by counting how often each number is produced one can construct an empirical distribution using the relative frequencies. For instance, if 14 out of 100 samples are 0, and the other 86 samples are 1, the empirical distribution could assign the probability 0.14 to 0 and 0.86 to 1. The entropy of this distribution would be approximately 0.58.

But what is entropy in the first place? It is the average surprise of (in this case) the different outputs of the device. Here, surprise (also called "surprisal") is a technical notion for the negative logarithm of an event's probability. The average surprise (the entropy) is now computed as follows: $H = -[0.14 * \log(0.14) + 0.86 * \log(0.86)]$. If this quantity is low, it is because the surprise values of the individual outcomes are low (or at least most of them). So to have a low entropy, the surprise of states must be low at any time (or at least most of the time).

We can again apply this to the fish example. Most of the time, the fish will be in unsurprising states. Given exhaustive knowledge about the fish, we can in principle describe the regions of the state space in which the fish is likely to be found, and construct an "armchair" probability distribution that reflects this knowledge. Or we can observe the fish and note the relative frequencies with which it is found in different regions of its state space. In the long run, this empirical distribution should become more and more similar to the "armchair" distribution. (This is an informal description of the *ergodicity* assumption, which is a formally defined feature of certain random processes, see Friston 2009, p. 293.)

So far, we have observed the fish from the outside, from the observer's perspective. What happens if we change our point of view and observe the fish from "the animal's perspective" (as Eliasmith 2000, pp. 25 f., calls it)? The key difference is that we do not even know the fish's current state. An organism gains knowledge about its own current state by sensory measurements, but these measurements provide the organism with only partial and perhaps noisy information. What is more, the organism does not have access to the probability distribution relative to which the surprisal of its states can be computed. Here, the free-energy principle (FEP) provides a principled solution (feature #12).

The general strategy of FEP consists of two steps. The first is to try to match an internally coded probability distribution (a recognition distribution) to the true posterior distribution of the hidden states, given sensory signals. The second is to try to change sensory signals in such a way that the surprise of sensory and hidden states is low at any given time. This may seem to make matters even worse, because now there are two problems: How can the recognition distribution be matched to an unknown posterior, and how can the surprise of sensory signals be minimized, if the distribution relative to which the surprise is defined is unkown? The ingenuity of FEP consists in solving both problems by minimizing free energy. Here, free energy is an information-theoretic quantity, the minimization of which is possible from the animal's perspective (for details, see Friston 2008; Friston 2009; Friston 2010).

Explaining this requires a slightly more formal description (here, we will simplify matters; a much more detailed, but still accessible, explanation of the free-energy principle can be found in Bogacz 2015). Firstly, "matching" the recognition distribution to an unknown posterior is just an approximation to Bayesian inference in which the recognition distribution is assumed to have a certain form (e.g., Gaussian). This simplifies computations. Secondly, once an approximation to the true model is computed, free energy constitutes a tight bound on the surprisal of sensory signals. Hence, minimizing free energy by changing sensory signals will, implicitly, minimize surprisal.

Note that the connection between FEP and Bayesian inference is straightforward: minimizing free energy entails approximating the posterior distribution by the recognition distribution. If the recognition distribution is assumed to be Gaussian (with the famous bell-shaped probability density function), minimizing free energy entails minimizing precision-weighted prediction errors. So at least under this assumption (which is called the *Laplace assumption* ), there is also a connection between FEP and prediction error minimization. In fact, FEP can be regarded as the fundamental theory, which can combine the different features of predictive processing described above within a single, formally rigorous framework. However, it is debatable which of these features are actually entailed by FEP. As mentioned before, **Environmental Seclusion** is an example of a controversial feature (see Fabry 2017a; Clark 2017). Therefore, it could be helpful to look at specific aspects of this novel proposal not only from an empirical, but also from a conceptual and a metatheroetical perspective. This was one major motive behind our initiative, leading to the current collection of texts.

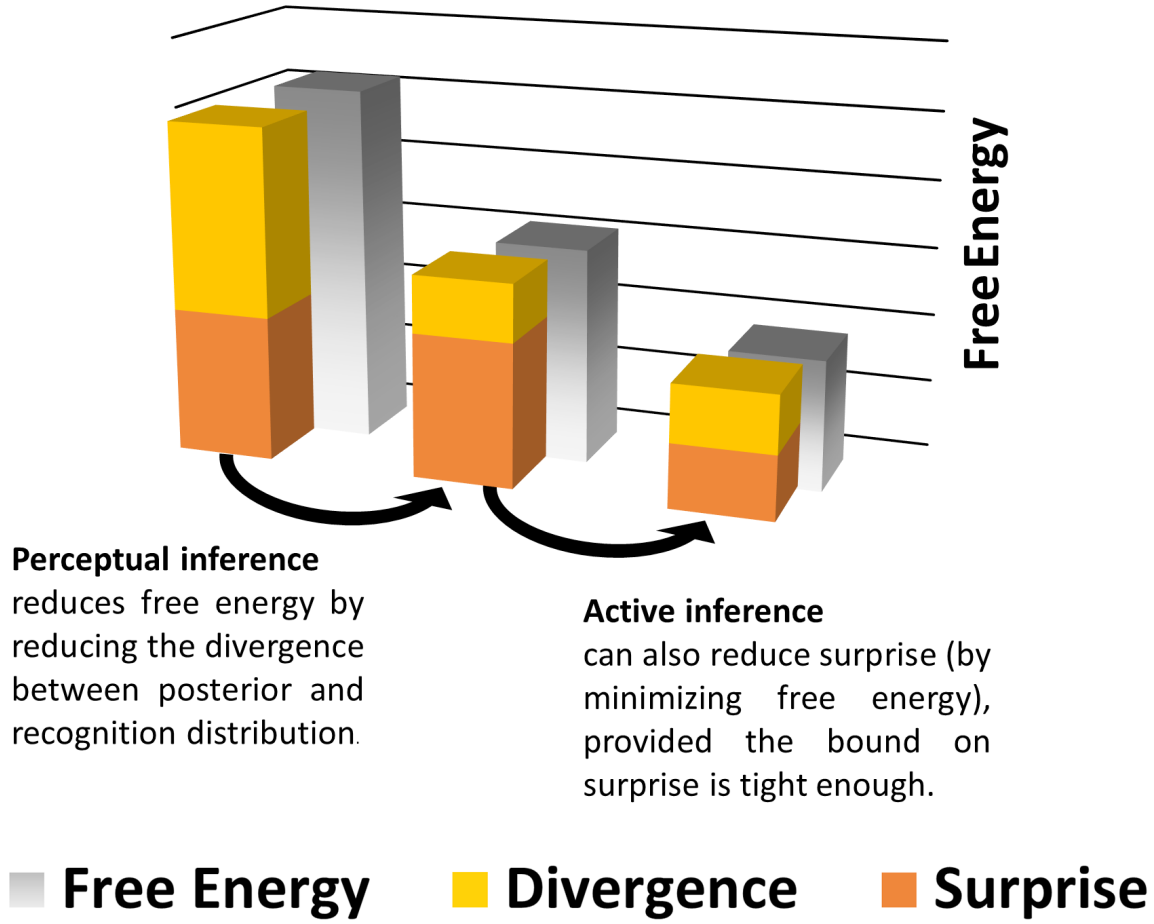


**Figure 1:** A schematic illustration of how minimizing free energy can, implicitly, minimize surprise. Initially, the recognition distribution will not match the true posterior distribution (of hidden causes, given sensory signals) very well. In order to improve the recognition distribution, it can be changed in such a way that the measured sensory signals become more likely, given this model (this means the model evidence is increased). One way to implement this is by minimizing prediction error. So the assumption is that sensory signals are unsurprising, and this should be reflected by the recognition distribution (i.e., the recognition distribution is altered in such a way that, relative to this distribution, sensory signals are unsurprising). Of course, it could be that the sensory signals are, relative to the true posterior, surprising. For this reason, the recognition distribution has to be tested. This is done, implicitly, by bringing about changes in the world that will, if the recognition distribution is adequate, lead to unsurprising sensory signals. This is active sampling. To some extent, sensory signals will always be surprising, so an adjustment of the recognition distribution will always be required, followed by active sampling, and a further adjustment of the recognition distribution, etc. So this bootstrapping process works through a continuous trial-and-error procedure, and depends on an intimate causal connection between the agent and its environment. Although the black arrows are meant to indicate a temporal sequence, there does not have to be a neat separation between perceptual inference and active inference, and the bootstrapping process could also start with bodily movements.

## Glossary

**Active inference**: 1. Computational process in which prediction error is minimized by acting on the world ("making the world more similar to the model"), as opposed to minimizing prediction error by changing the internal model, i.e. perceptual inference ("making the model more similar to the world"). 2. Also used as a generic term for the computational processes which underpin both action and perception, and, in the context of FEP, for all computational processes that minimize free energy.

**Bayesian inference**: Updating a model in accordance with Bayes' rule, i.e. computing the posterior distribution: $p(c|s) = p(s|c)p(c)/p(s)$. For an example, see (Harkness and Keshava 2017).

**Counterfactual model**: A counterfactual model is a conditional probability distribution that relates possible actions to possible future states (at least following Friston et al. 2012b).

**Estimator**: A statistical estimator is a function of random variables that are conceived as samples; so an estimator specifies how to compute an estimate from observed data. An estimate is a particular value of an estimator (which is computed when particular samples, i.e., realizations of random variables, have been obtained).

**"Explaining Away"**: The notion of "explaining away" is ambiguous. 1. Some authors write that sensory signals are explained away by top-down predictions (cf. Clark 2013a, p. 187). 2. Another sense in which the term is used is that competing hypotheses or models are explained away (cf. Hohwy 2010, p. 137). 3. A third sense is as in *explaining prediction error away* (cf. Clark 2013a, p. 187).

**Free energy**: In the context of Friston's FEP, free energy is not a thermodynamic quantity, but an information-theoretic quantity that constitutes an upper bound on surprisal. If this bound is tight, the surprisal of sensory signals can therefore be reduced if free energy is minimized by bringing about changes in the world.

**Gaussian distribution**: The famous bell-shaped probability distribution (also called the normal distribution). Its prominence is grounded in the central limit theorem, which basically states that many distributions can be approximated by Gaussian distributions.

**Generative model**: The joint probability distribution of two or more random variables, often given in terms of a prior and a likelihood: $p(s,c) = p(s|c)p(c)$. (Sometimes, only the likelihood $p(s|c)$ is called a "generative model".) The model is generative in the sense that it models how sensory signals $s$ are *generated* by hidden causes $c$. Furthermore, it can be used to *generate* mock sensory signals, given an estimate of hidden causes.

**Hierarchy**: PP posits a hierarchy of estimators, which operate at different spatio-temporal timescales (so they track features at different scales). The hierarchy does not necessarily have a top level (but it might have a center — think of the levels as rings on a disc or a sphere).

**Inverse problem**: From the point of view of predictive coding, the problem of perception requires inverting the mapping from hidden causes to sensory signals. This problem is difficult, to say the least, because there is not usually a unique solution, and sensory signals are typically noisy (which means that the mapping from hidden causes to sensory signals is not deterministic).

**Prediction**: A prediction is a deterministic function of an estimate, which can be compared to another estimate (the predicted estimate). Predictions are not necessarily about the future (note that a variable can be predictive of another variable if the first carries information about the second, i.e., if there is a correlation, cf. Anderson and Chemero 2013, p. 204). Still, many estimates in PP are also predictive in the temporal sense (cf. Butz 2017; Clark 2013c, p. 236).

**Precision**: The precision of a random variable is the inverse of its variance. In other words, the greater the average divergence from its mean, the lower the precision of a random variable (and vice versa).

**Random variable**: A random variable is a measurable function between a probability space and a measurable space. For instance, a six-sided die can be modeled as a random variable, which maps each of six equally likely events to one of the numbers in the set {1,2,3,4,5,6}.

**Surprisal**: An information-theoretic notion which specifies how unlikely an event is, given a model. More specifically, it refers to the negative logarithm of an event's probability (also just called "surprise"). It is important not to confuse this subpersonal, information-theoretic concept with the personal-level, phenomenological notion of "surprise".

## References


Adams, R. A., Huys, Q. J. & Roiser, J. P. (2016). Computational psychiatry: Towards a mathematically informed understanding of mental illness. *J Neurol Neurosurg Psychiatry, 87* (1), 53-63. https://dx.doi.org/10.1136/jnnp-2015-310737.

Anderson, M. L. (2017). Of Bayes and bullets: An embodied, situated, targeting-based account of predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Anderson, M. L. & Chemero, T. (2013). The problem with brain GUTs: Conflation of different senses of ''prediction" threatens metaphysical disaster. *Behavioral and Brain Sciences, 36* (3), 204–205.

Badets, A., Koch, I. & Philipp, A. M. (2014). A review of ideomotor approaches to perception, cognition, action, and language: Advancing a cultural recycling hypothesis. *Psychological Research, 80* (1), 1–15. https://dx.doi.org/10.1007/s00426-014-0643-8.

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. https://dx.doi.org/10.1016/j.neuron.2012.10.038.

Bogacz, R. (2015). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*. https://dx.doi.org/10.1016/j.jmp.2015.11.003.

Brodski, A., Paasch, G.-F., Helbling, S. & Wibral, M. (2015). The faces of predictive coding. *The Journal of Neuroscience, 35* (24), 8997-9006. https://dx.doi.org/10.1523/jneurosci.1529-14.2015.

Brook, A. (2013). Kant's view of the mind and consciousness of self. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*.

Bruineberg, J. (2017). Active inference and the primacy of the 'I can'. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Bruineberg, J., Kiverstein, J. & Rietveld, E. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*, 1–28. https://dx.doi.org/10.1007/s11229-016-1239-1.

Burr, C. (2017). Embodied decisions and the predictive brain. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Butz, M. V. (2017). Which structures are out there? Learning predictive compositional concepts based on social sensorimotor explorations. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Clark, A. (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36* (3), 181–204. https://dx.doi.org/10.1017/S0140525X12000477.

———(2013b). The many faces of precision (Replies to commentaries on ''Whatever next? Neural prediction, situated agents, and the future of cognitive science"). *Frontiers in Psychology, 4*, 270. https://dx.doi.org/10.3389/fpsyg.2013.00270.

——— (2013c). Are we predictive engines? Perils, prospects, and the puzzle of the porous perceiver. *Behavioral and Brain Sciences, 36* (3), 233–253. https://dx.doi.org/10.1017/S0140525X12002440.

——— (2015). Radical predictive processing. *The Southern Journal of Philosophy*, 53, 3–27. https://dx.doi.org/10.1111/sjp.12120.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

——— (2017). How to knit your own Markov blanket: Resisting the second law with metamorphic minds. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

——— (in press). Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs*. https://dx.doi.org/10.1111/nous.12140.

Clowes, M. B. (1969). Pictorial relationships – A syntactic approach. In B. Meltzer & D. Michie (Eds.) (pp. 361–383). Edinburgh, UK: Edinburgh University Press.

Colombo, M. (2017). Social motivation in computational neuroscience: Or if brains are prediction machines then the Humean theory of motivation is false. In J. Kieverstein (Ed.) *Routledge handbook of philosophy of the social mind*. Abingdon, OX / New York, NY: Routledge.

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking.* New York, N.Y., and London, UK: W.W. Norton & Company.

Dewhurst, J. (2017). Folk psychology and the Bayesian brain. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Downey, A. (2017). Radical sensorimotor enactivism & predictive processing. Providing a conceptual framework for the scientific study of conscious perception. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Dołęga, K. (2017). Moderate predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Drayson, Z. (2017). Modularity and the predictive mind. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Egan, F. (2014). How to think about mental content. *Philosophical Studies, 170* (1), 115-135. https://dx.doi.org/10.1007/s11098-013-0172-0.

Eliasmith, C. (2000). *How neurons mean: A neurocomputational theory of representational content.* PhD dissertation, Washington University in St. Louis. Department of Philosophy.

Engel, A. K., Fries, P. & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top-down processing. *Nat Rev Neurosci, 2* (10), 704–716.

Fabry, R. E. (2017a). Predictive processing and cognitive development. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

——— (2017b). Transcending the evidentiary boundary: Prediction error minimization, embodied interaction, and explanatory pluralism. *Philosophical Psychology*, 1–20. https://dx.doi.org/10.1080/09515089.2016.1272674.

Feldman, H. & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*. https://dx.doi.org/10.3389/fnhum.2010.00215.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks, 16* (9), 1325–1352.

——— (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360* (1456), 815-836. https://dx.doi.org/10.1098/rstb.2005.1622.

——— (2008). Hierarchical models in the brain. *PLoS Computational Biology, 4* (11), e1000211. https://dx.doi.org/10.1371/journal.pcbi.1000211.

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

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

Friston, K. & Buzsáki, G. (2016). The functional anatomy of time: What and when in the brain. *Trends in Cognitive Sciences, 20* (7), 500–511. https://dx.doi.org/10.1016/j.tics.2016.05.001.

Friston, K. & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364* (1521), 1211–1221. https://dx.doi.org/10.1098/rstb.2008.0300.

Friston, K. J. & Stephan, K. E. (2007). Free-energy and the brain. *Synthese, 159* (3), 417-458. https://dx.doi.org/10.1007/s11229-007-9237-y.

Friston, K., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics, 104* (1-2), 137–160. https://dx.doi.org/10.1007/s00422-011-0424-z.

Friston, K., Samothrakis, S. & Montague, R. (2012a). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics, 106* (8), 523-541. https://dx.doi.org/10.1007/s00422-012-0512-8.

Friston, K., Adams, R., Perrinet, L. & Breakspear, M. (2012b). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3* (151). https://dx.doi.org/10.3389/fpsyg.2012.00151.

Friston, K. J., Stephan, K. E., Montague, R. & Dolan, R. J. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry, 1* (2), 148–158. https://dx.doi.org/10.1016/S2215-0366(14)70275-5.

Giordanetti, P., Pozzo, R. & Sgarbi, M. (2012). *Kant's philosophy of the unconscious.* Berlin, Boston: De Gruyter.

Gonzalez-Gadea, M. L., Chennu, S., Bekinschtein, T. A., Rattazzi, A., Beraudi, A., Tripicchio, P., Moyano, B., Soffita, Y., Steinberg, L., Adolfi, F., Sigman, M., Marino, J., Manes, F. & Ibanez, A. (2015). Predictive coding in autism spectrum disorder and attention deficit hyperactivity disorder. *Journal of Neurophysiology, 114* (5), 2625-2636. https://dx.doi.org/10.1152/jn.00543.2015.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 290* (1038), 181–197.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences, 27* (3), 377–396.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, 559–582. https://dx.doi.org/10.1007/s11229-015-0762-9.

Harkness, D. L. & Keshava, A. (2017). Moving from the what to the how and where – Bayesian models and predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Herbart, J. F. (1825). *Psychologie als Wissenschaft neu gegründet auf Erfahrung, Metaphysik und Mathematik. Zweiter, analytischer Teil.* Königsberg: Unzer.

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. https://dx.doi.org/10.5096/ASCS200922.

———(2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3*. https://dx.doi.org/10.3389/fpsyg.2012.00096.

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

———(2016). The self-evidencing brain. *Noûs, 50* (2), 259–285. https://dx.doi.org/10.1111/nous.12062.

——— (2017). How to entrain your evil demon. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Hommel, B. (2015). The theory of event coding (TEC) as embodied-cognition framework. *Frontiers in Psychology, 6*. https://dx.doi.org/10.3389/fpsyg.2015.01318.

Hommel, B., Müsseler, J., Aschersleben, G. & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849–878. https://dx.doi.org/10.1017/S0140525X01000103.

Horn, B. K. P. (1980). *Derivation of invariant scene characteristics from images* (pp. 371–376). https://dx.doi.org/10.1145/1500518.1500579.

James, W. (1890). *The principles of psychology.* New York: Henry Holt.

Kant, I. (1998). *Critique of pure reason.* Cambridge, MA: Cambridge University Press.

——— (1998[1781/87]). *Kritik der reinen Vernunft.* Hamburg: Meiner.

Kiefer, A. (2017). Literal perceptual inference. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350* (6266), 1332-1338. https://dx.doi.org/10.1126/science.aab3050.

Lee, T. S. & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A, 20* (7), 1434–1448. https://dx.doi.org/10.1364/JOSAA.20.001434.

Lenoir, T. (2006). Operationalizing Kant: Manifolds, models, and mathematics in Helmholtz's theories of perception. In M. Friedman & A. Nordmann (Eds.) *The Kantian legacy in nineteenth-century science* (pp. 141–210). Cambridge, MA: MIT Press.

Limanowski, J. (2017). (Dis-)attending to the body. Action and self-experience in the active inference framework. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Lotze, R. H. (1852). *Medicinische Psychologie oder Physiologie der Seele.* Leipzig: Weidmann'sche Buchhandlung.

Metzinger, T. (2004[2003]). *Being no one: The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

———(2017). The problem of mental action. Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Palmer, C. J., Paton, B., Kirkovski, M., Enticott, P. G. & Hohwy, J. (2015). Context sensitivity in action decreases along the autism spectrum: A predictive processing perspective. *Proceedings of the Royal Society of London B: Biological Sciences, 282* (1802). https://dx.doi.org/10.1098/rspb.2014.1557.

Prinz, W. (1990). A common coding approach to perception and action. In O. Neumann & W. Prinz (Eds.) *Relationships between perception and action* (pp. 167–201). Berlin; Heidelberg: Springer.

Quadt, L. (2017). Action-oriented predictive processing and social cognition. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Seth, A. K. (2015). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570108.

Shi, Y. Q. & Sun, H. (1999). *Image and video compression for multimedia engineering: fundamentals, algorithms, and standards.* Boca Raton, FL: CRC Press.

Sloman, A. (1984). Experiencing omputation: A tribute to Max Clowes. In M. Yazdani (Ed.) *New horizons in educational computing* (pp. 207–219). Chichester: John Wiley & Sons.

Snowdon, P. (1992). How to interpret 'direct perception'. In T. Crane (Ed.) *The contents of experience* (pp. 48–78). Cambridge: Cambridge University Press.

Spratling, M. W. (2016). A review of predictive coding algorithms. *Brain and Cognition*. https://dx.doi.org/10.1016/j.bandc.2015.11.003.

Stock, A. & Stock, C. (2004). A short history of ideo-motor action. *Psychological Research, 68*, 176–188. https://dx.doi.org/10.1007/s00426-003-0154-5.

Swanson, L. R. (2016). The predictive processing paradigm has roots in Kant. *Frontiers in Systems Neuroscience, 10*, 79. https://dx.doi.org/10.3389/fnsys.2016.00079.

Todorov, E. (2009). Parallels between sensory and motor information processing. In M. S. Gazzaniga (Ed.) *The cognitive neurosciences. 4th edition* (pp. 613–623). Cambridge, MA / London, UK: MIT Press.

Van de Cruys, S., Evers, K., Van der Hallen, R., van Eylen, L., Boets, B., de-Wit, L. & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review, 121* (4), 649–675. https://dx.doi.org/10.1037/a0037665.

Van Doorn, G., Hohwy, J. & Symmons, M. (2014). Can you tickle yourself if you swap bodies with someone else? *Consciousness and Cognition, 23*, 1-11. http://dx.doi.org/10.1016/j.concog.2013.10.009.

Van Doorn, G., Paton, B., Howell, J. & Hohwy, J. (2015). Attenuated self-tickle sensation even under trajectory perturbation. *Consciousness and Cognition, 36*, 147–153. https://dx.doi.org/10.1016/j.concog.2015.06.016.

Von Helmholtz, H. (1855). *Ueber das Sehen des Menschen.* Leipzig: Leopold Voss.

——(1867). *Handbuch der physiologischen Optik.* Leipzig: Leopold Voss.

—— (1959[1879/1887]). *Die Tatsachen in der Wahrnehmung. Zählen und Messen.* Darmstadt: Wissenschaftliche Buchgesellschaft.

——(1985[1925]). *Helmholtz's treatise on physiological optics.* Birmingham, AL: Gryphon Editions.

Von Holst, E. & Mittelstaedt, H. (1950). Das Reafferenzprinzip. *Die Naturwissenschaften, 37* (20), 464–476.

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L. & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proc Natl Acad Sci U S A, 108* (51), 20754-9. https://dx.doi.org/10.1073/pnas.1117807108.

Wiese, W. (2016). Action is enabled by systematic misrepresentations. *Erkenntnis*. https://dx.doi.org/10.1007/s10670-016-9867-x.

Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician, 42* (4), 278–280. https://dx.doi.org/10.2307/2685143.

# How to Entrain Your Evil Demon

Jakob Hohwy

The notion that the brain is a prediction error minimizer entails, via the notion of Markov blankets and self-evidencing, a form of global scepticism — an inability to rule out evil demon scenarios. This type of scepticism is viewed by some as a sign of a fatally flawed conception of mind and cognition. Here I discuss whether this scepticism is ameliorated by acknowledging the role of action in the most ambitious approach to prediction error minimization, namely under the free energy principle. I argue that the scepticism remains but that the role of action in the free energy principle constrains the demon's work. This yields new insights about the free energy principle, epistemology, and the place of mind in nature.

**Keywords**

Active inference | Agency | Approximate bayesian inference | Coupled oscillation | Epistemic value | Evil demon | Exact inference | Free energy principle | Functional role semantics | Internal models | Internalism | Interventionism | Markov blanket | Perceptual inference | Perceptual learning | Prediction error minimization | Scepticism | Self-evidencing | Variational bayesian inference

**Acknowledgements**

Thanks to Regina Fabry and Thomas Metzinger for a discussion that lead to this paper; thanks to two anonymous referees and the editors of this volume.

## 1 Prediction Error Minimization and the Free Energy Principle

An emerging, unified theory of brain function seeks to explain all aspects of mind and cognition as the upshots of prediction error minimization (Friston 2003; Friston 2010; Hohwy 2010; Clark 2013; Hohwy 2013; Clark 2016b). The idea is that the brain is a model of its environment, which garners evidence for itself by explaining away sensory input. This happens in a process of approximate Bayesian inference, where hypotheses about sensory input are generated from the model, and the predictions of these hypotheses tested against the actual input.

The extent to which these predictions are correct determines the accuracy of the model. Essentially, the better the model is at minimizing the error in its predictions (the prediction error) the more evidence it accumulates for itself. In the course of maximizing evidence for itself, the model infers the causes of its sensory input. To illustrate, it seems reasonable to say that if I hypothesize that the current sound is caused by an approaching train, predict that soon the sound will therefore get louder and then recede, and then confirm this prediction, then I have correctly inferred something about the causes of my evidence. This is enshrined in perceptual inference and perceptual learning where an internal model of hidden causes is optimized through prediction error minimization.

Exact inference would correspond to following Bayes' rule for updating perceptual beliefs however it is unlikely the brain engages in exact inference since for realistic perceptual settings inverting the model that generates the predictions to infer the hidden causes presents an intractable computational problem. Instead, there is some reason to think that the brain can engage in approximate inference

Hohwy, J. (2017). How to Entrain Your Evil Demon.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 2. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573048

(Friston 2003; Friston 2005). In exact inference, prediction error is minimized over the long term perspective, namely as the model becomes increasingly better. This leads to the idea that a system that minimizes prediction error on average and over the long term is likely to approximate the outcomes of exact Bayesian inference. Approximate inference (especially variational Bayesian inference) does not present intractable problems and importantly can be executed by systematically varying internal model parameters for a system that just has access to its own internal states and the states of its sensory organs. The move to approximate inference is attractive as it overcomes formal obstacles for conceiving the brain as an inferential system. Moreover, there is a good case that it is biologically plausible since it speaks to the overall architecture of the brain in terms of not just relatively sparse and focused forward (bottom-up) connectivity but also hitherto poorly understood, more diffuse and copious backwards (top-down) connectivity; perceptual inference and learning through prediction error minimization in the brain is also arguably consistent with the different types of plasticity operating at different time scales and different hierarchical levels in the brain (Mumford 1992; Friston 2005).

The free energy principle (Friston 2010) sets notions of perception and action in just the long term perspective needed for prediction error minimization. The principle begins with the observation that persisting organisms maintain themselves in a limited set of states, rather than dispersing through all states. This is a statistical notion because it allows a description of the organism in terms of a probability density, or model, which identifies the states in which it is most probable to find it. The principle then states that organisms manage to maintain themselves in their expected states by minimizing their free energy, which (given assumptions about the shape of the probability densities) is the long-term average prediction error, given their model (for an introduction to the free energy principle, see Hohwy 2015).

Crucially, the free energy principle imbues organisms with agency, such that they can act to maintain themselves in their expected states. This happens through prediction error minimization: predictions are held stable rather than revised in the light of immediate prediction error, and action moves the organism around to change the input to the senses until the predictions come true. Since prediction error minimization in the long run approximates Bayesian inference, action can be said to be an inferential process in the same sense as perception is inferential. Hence, action is labeled 'active inference' (Friston and Stephan 2007; Friston 2010). Active inference increases the accuracy of internal models and perceptual inference optimizes these models. Active inference is a powerful addition to the inferential framework because it allows the agent to, firstly, increase the epistemic value of their model (e.g., confirming that it is really a train approaching by opening the window facing the tracks to hear the sound better), and secondly, to occupy one's expected states (e.g., predicting that one is travelling on the train, noting the ensuing prediction error, and moving around in the environment until one is in fact in a train and the prediction error is minimized) (see Hohwy 2013, Clark 2015, Clark 2016a, for extensive discussion of the role of active inference in a prediction error minimization framework).

This chapter explores the free energy principle through a very particular, epistemological lens. It seems that the principle entails global scepticism. This could be viewed as a significant epistemological problem, and I will discuss that towards the end of the paper. But entailment of scepticism is also at times viewed as a symptom of a fatally flawed, old-fashioned model of the mind, at odds with contemporary embodied, enactive and extended (EEE) models of mind. These models make action central to mind and cognition, and in some cases imply that the inclusion of action removes the scepticism. This in turn gives rise to the hypothesis that the free energy principle, with its strong focus on action, could avoid scepticism and be an underlying theoretical framework for EEE models. Here, I argue that, action notwithstanding, scepticism remains for the free energy principle and indeed the addition of action does nothing to remove the inferential, internalist aspect of the prediction error minimization approach. The more constructive part of the chapter explores how active inference nevertheless does change the epistemic landscape and how this is reflected in the representational properties of the

internal prediction-generating model. This seems to retain the inferential and internalist aspects of prediction error minimization but is able to accommodate to some degree several EEE insights.

The chapter begins by introducing the notions of Markov blankets and self-evidencing, which gives a principled and relatively clear understanding of inferentialism and internalism within a prediction error minimization scheme. It then moves on to explain how scepticism is entailed by such schemes and how adding action does not make scepticism go away. The chapter then moves on to a discussion of the interesting way in which action does change our conception of the epistemic status of agents and what this tells us about internal models. Finally, the chapter attempts to assess what this all means for our overall understanding of the free energy principle.

## 2 Markov Blankets and Self-Evidencing.

From the previous section's brief description of prediction error minimization under the free energy principle, a somewhat unusual conception of a biological agent emerges on which the agent simply is a model, which is engaged in prediction error minimization through action and through optimization of model parameters. It may seem incongruous to equate 'agent' with 'model', but, on the free energy principle, models are the things that do the acting, based on their representation of the world, and which (therefore) persist through time. It seems reasonable to label as an 'agent', at least in a rudimentary sense, an acting, representing, persisting thing. As we will see, this equation of agent with model in turn has epistemological and theoretical implications.

One way of describing an agent as a model begins with the causal nets terms of a Markov blanket (Pearl 1988; Friston 2013; Hohwy 2016b), where a Markov blanket for a node in a causal net is the node's parents, children and parents of its children (Fig. 1).

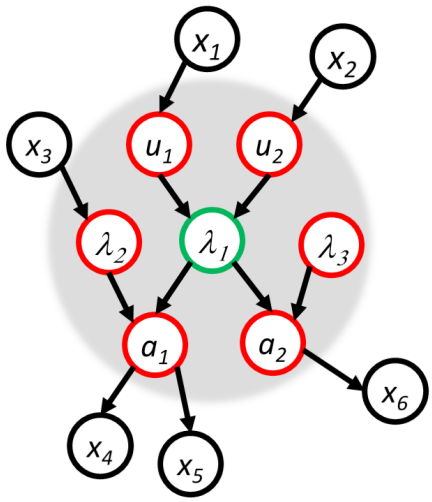


**Figure 1.** Markov blanket. The behavior of the green node $\lambda_1$ is known once the red nodes of the blanket (parents $u_i$, children $a_i$ and parents of children $\lambda_{2,3}$) are observed; states $x$ are external to the blanket and do not need to be observed to know the internal states.

The behavior of the blanketed node (node $\lambda_1$ in Fig 1) is in principle predictable just from observation of the states of the blanket, without knowing anything about the states of the external nodes beyond the blanket (that is, every node in the network is conditionally independent of $\lambda_1$ when conditioned on the nodes of the blanket).

The free energy agent maps onto the Markov blanket in the following way. The internal, blanketed states constitute the model. The children of the model are the active states that drive action through prediction error minimization in active inference, and the sensory states are the parents of the model,

driving inference. If the system minimizes free energy — or the long term average prediction error — then the hidden causes beyond the blanket are inferred.

Delineating the agent in terms of a Markov blanket means we can conceive succinctly of the behaviour that makes something an agent, in terms of a model that is self-evidencing. The notion of self-evidencing comes from Hempel's discussions of scientific explanation and captures the idea, firstly, that a hypothesis is supported to the extent it can explain away evidence and, secondly, that the "information or assumption that [the evidence] occurs forms an indispensable part of the only available evidential support for [the hypothesis]" (Hempel 1965, pp. 372-4). Above we said that the agent garners evidence for its model to the extent it can encounter and explain away sensory input (in much the way a scientist might garner evidence for their theory). But here there is not one thing, the agent, which provides evidence for another, the model: the model changes its sensory and active states to increase evidence for itself. These processes within the Markov blanket happen at the subpersonal level — there is no agent *making* the changes to the model. Hence, the notion of self-evidencing makes better sense of the initially incongruous idea that the model is the agent. The more the internal and active states change to anticipate the changes in sensory states, the more the organism evidences itself and thereby manages to persist (Hohwy 2016b).

Ultimately, this still somewhat challenging equation of model with agency may need to be fleshed out by connecting the notion of self-evidencing to the notion of self-organisation. For present purposes, we note that self-evidencing relative to a Markov blanket, which defines the borders of the agent, is a core ingredient of a prediction error minimization account.

## 3 Introducing the Evil Demon

Crucially, self-evidencing means we can understand the formation of a well-evidenced model, in terms of the existence of its Markov blanket: if the Markov blanket breaks down, the model is destroyed (there literally ceases to be evidence for its existence), and the agent disappears. The model infers the presumably distal hidden causes of its sensory input, and thereby garners evidence for itself, and this entire self-evidencing process can be understood just in terms of the causal behaviour of the Markov blanket and the internal nodes, without giving the model any kind of direct access to the hidden causes (even though naturally no-one but a solipsist would assume there exist no hidden causes whatsoever). The nodes of the internal model have access to the states of the blanket, and can gauge the prediction error but the model only represents the hidden external causes vicariously. The model builds up rich, detailed, and context-sensitive associations amongst the sensory states, and between sensory and active states, and learns which of these many possible associations tend to keep prediction error low in the long run.

This raises a familiar sceptical spectre: the model will not be able to distinguish between possibilities where similar flows of sensory input are caused by two very different causal processes, beyond the blanket. In the first possibility, the causes are as we suppose them to be, the familiar people, houses, trains, trees, fruits, etc. that we perceive in everyday life. In the second possibility, versions of which are familiar from much philosophy going back to Descartes' *Meditations*, the sensory input is caused by an evil demon (or evil scientist) with total control of the sensory states of the agent.

On the demon scenario, it seems the very same states that were assumed to represent people, houses, trains, trees, fruits, etc. are now not representing those things. They are all misrepresentations because really the hidden causes now belong to the cunning machinery and states of the evil demon. The prediction error minimization scheme therefore entails scepticism. There is never any justification for any perceptual belief that $p$ because the evidence for those beliefs cannot exclude the possibility that not-$p$. It is difficult to see how one could both be committed to the prediction error minimization framework and yet prevent this descent into scepticism (for earlier versions of probabilistic schemes and the entailed scepticism, see Eliasmith 2000, Usher 2001, Grush 2003).

Many different kinds of philosophical frameworks entail scepticism, and it is not surprising that such an empiricist kind of framework as the prediction error scheme entails scepticism. A kindred framework, for example, is Thomas Metzinger's phenomenal self-model framework. Metzinger notes how our phenomenal, perceptual experience of the world seems certain to us but then notes that "[e]pistemologically speaking, however, the subjective experience of certainty transported by … phenomenal content is unjustified. [P]henomenal content supervenes on internal and contemporaneous properties of the human brain. A disembodied brain in a vat, therefore, could in principle realize precisely the same degree of experiential certainty going along with our ordinary conscious experience of embodiment" (Metzinger 2004, p. 310). Metzinger is not here primarily interested in the epistemic threat from scepticism but rather in what the advent of scepticism can teach us about a particular conception of mind and consciousness. Though I shall discuss the epistemic ramifications of scepticism later in this paper, the initial focus is the same: skepticism as a testing ground for understanding and assessing competing models of the mind.

In the contemporary debate, entailment of scepticism is at times viewed as symptomatic of a poorly conceived model of the mind that operates with a representation-hungry internalist machine sandwiched between worldly input and bodily action (Hurley 1998). The reasoning is straightforward. Skepticism arises for systems with insuperable sensory veils, such that there is only, and at best, indirect access to the external world, namely by internal representations of external states of affairs. Due to its heavy inferential nature and invocation of Markov blankets, it may appear that the prediction error minimization model is the epitome of scepticism-inducing cognitive science (Anderson and Chemero 2013). Thompson and Cosmelli , likewise, see the evil demon scenario as an important testing ground for the debate whether perception is 'Brainbound' or 'Enactive' (Thompson and Cosmelli 2011). They argue that, if we look at the explanatory aspects of enactive theories, and seriously consider what it would take to maintain an evil demon (or brain in a vat) scenario, then we should accept that the scenario is not possible. Cosmelli and Thompson use this in an argument against brainbound conceptions of the mind, including the type that the prediction error minimization account seems committed to (Thompson 2007, p. 242); I return briefly to their views below.

The alternative is a model that sees mind as non-representational, extended, situated, embodied or enactive (Hurley 1998; Noë 2004; Gallagher 2005; Thompson 2007; Clark 2008; Hutto and Myin 2013). The reasoning with respect to scepticism, on behalf of such EEE views, is that skepticism will not arise if the sensory veil is obliterated and that it is obliterated by having the mind be extended into or embedded into the environment, by not operating with representations, by making the body essential to the operation of the mind, or by having fluid and direct enactive commerce with the world.

The advantage of these alternative models is not merely that they seem to avoid scepticism — though this is not ignored altogether it is rarely the focus of EEE theories — but that they build into the core conception of mind the fact that minds belong to agents with bodies that live and act in environments with other agents. It is seen as misguided to build models of the mind that leave all that matters (body, world, action, others) as mere incidental afterthoughts to the rarified internal, representational workings of the mind.

There are some nice questions about how exactly EEE models of the mind avoid scepticism. These models seem so revisionary about our internal, mental workings that traditional notions of knowledge, belief and justification threaten to become obsolete. Scepticism might be off the table but it may be that basic epistemic conceptions go missing too. For example, if there is no principled distinction between belief and what the belief is about, then it is difficult to see how there can be beliefs at all. It would be a pyrrhic victory if scepticism is conquered but at the price of the very conceptions of knowledge, belief and justification.

Here, I will consider a possible way out of this quandary. Specifically, I will consider if the addition of *action* to the prediction error minimization scheme is sufficient to overcome scepticism while both respecting some insights of the alternative EEE models of the mind and retaining the virtues of

prediction error minimization as a unified explanation of perception, knowledge, and action. There are several attractions to considering this way out of the quandary. First, action in the shape of active inference is a key part of the prediction error minimization scheme conceived under the free energy principle, so it would be odd to mount a defense of prediction error minimization schemes without giving action a central role. The importance of action is recognised clearly by both main philosophical treatments of the scheme, which repeatedly appeal to active inference (Hohwy 2013; Clark 2016b). Second, action is a linchpin of EEE approaches, which overwhelmingly appeal to action in arguments against the staid, passive model of the brain as merely engaged in computational operations over internal representations (see references above). To be clear, in this chapter, I focus on action as a treatment for scepticism within the prediction error minimization scheme. This is an interesting exercise because, as we shall see, it will tell us something about the scheme and its epistemological consequences. This means I am not here offering a full treatment of the connection between EEE accounts and prediction error minimization. That is a task for another occasion (for a first treatment, see Hohwy 2016b, and for others Bruineberg and Rietveld 2014, Clark 2015, Bruineberg 2016, Kirchoff 2016). However, the discussion offered here is, I believe, an important testing ground for some of these debates.

Before moving on to discussing whether the addition of active inference changes the game with respect to scepticism, I will briefly discuss, and then set aside, an argument that denies a role for scepticism for some of these debates. Clark (Clark 2016a, Sec. 3) argues that scepticism and evil demon scenarios are irrelevant to assessing the compatibility of prediction error minimization schemes and (some) EEE approaches. He points out that scepticism is unlikely to disappear even on embodied cognition type views, or, in other words that the emergence of scepticism is a red herring in these debates, something we should not focus on. I have sympathy for this position: it agrees with me that the prediction error minimization scheme entails scepticism though it disagrees concerning the significance of this for assessing embodied cognition. However, as we saw above, Clark's position is at odds with others who are advocating EEE approaches: they see scepticism as a clear symptom of misguided, internalist, non-EEE theories.

The reason Clark can reasonably assert that scepticism is a red herring is that, for him "the claim that lies at the heart of recent work on the embodied mind [is that it] fundamentally rejects […] the richly reconstructive model of perception" (Clark 2016a, p. 12). That is, the embodied mind debate concerns "the question whether apt actions are always and everywhere computed by using sensing to get enough information into the system to allow it to plot its response by exploring an internally represented recapitulation of the distal world" (Clark 2016a, p. 11). I think he is right that the issue of whether internal models are rich and reconstructive is orthogonal to the issue of whether scepticism is entailed by the view or not. What this should tell us, however, is not that the issue of scepticism is irrelevant *tout court* to all kinds of EEE views but rather that it is irrelevant specifically to the issue of rich and reconstructive models.

Notice that Clark invokes active inference in his own conception of the embodied mind and his argument that it is consistent with the prediction error minimization scheme. He says "The appearance of conflict [between the embodied mind and prediction error mimimization schemes] arises from ambiguities in the notions of inference and seclusion themselves. For these notions may seem to imply the presence of a rich inner recapitulation of the distal environment, with a consequent downgrading of the role of action and upgrading of the role of reasoning defined over that inner model" (Clark 2016a, p. 11). The difference between Clark and others, like (Thompson and Cosmelli 2011) and (Anderson and Chemero 2013) is then that the type of action-oriented embodied mind that Clark defends is not threatened by the scheme's entailment of scepticism.

This leads me to a clarificatory point about the notion of internalism. It is tempting to say that any account of perception and cognition that operates with internal models must in some sense be internalist. But the natural next question is what makes internal models internal? I think a natural default has been to answer that internal models are internal because they are housed in the brain. But this

cannot be right. The notion of internal models belongs with machine learning and computational science and as such cannot be necessarily wedded to biological organs. A better answer is provided by the notion of Markov blankets and self-evidencing through approximation to Bayesian inference. Here there is a principled distinction between the internal, known causes as they are inferred by the model and the external, hidden causes on the other side of the Markov blanket. This seems a clear way to define internalism as a view of the mind according to which perceptual and cognitive processing all happen within the internal model, or, equivalently, within the Markov blanket. This is then what non-internalist views must deny (for further discussion of where to place the blanket, and how much it can encompass, see Hohwy 2016b).

Notice that this definition of internalism makes Clark an internalist (as emphasized by Anderson and Chemero 2013) and the main difference between Clark's and my own interpretations of prediction error minimization schemes then seems to be whether internal models are rich and reconstructive, and whether they downgrade the role for action. That discussion is for another occasion. This chapter focuses on the issue of skepticism.

## 4 Active Inference and Sensory Veil Skepticism

Active inference, as we saw in previous sections, builds on the idea that sensory input can become better predictable through making changes to the active states of the Markov blanket. The system is able to learn that certain active states are associated with certain sensory states. For example, there may be a learnt association between my finger actively exerting force on the button of the remote control and a change in sensory states as the TV (television) channel changes. If my expected state is that I am in fact watching the news rather than the soap, then I might prioritise the actually false hypothesis that my finger is pressing the button. The hypothesis generates a prediction of sensory input, which is fleshed out partly in terms of expected bodily states, for example that there will be a certain proprioceptive input related to arm and finger movements. This prediction engenders a prediction error, which is minimized as the arm and finger are entrained via reflex arcs and movement actually occurs (see Hohwy 2016a, for more discussion of active inference).

Adding active inference to the prediction error minimization story does not however make that story any less inferential. Action is inference in the same way as perception is inference, namely by approximating Bayes through prediction error minimization. The processing, from the point of view of the system, is still told entirely in terms of self-evidencing within the Markov blanket: it is mainly a matter of hierarchical statistical associations between patterns of states of the nodes of the blanket and the internal states. In perceptual inference, the internal model might build up expectations for how sensory states unfold over time, for example, between sensory states $u_1$ and $u_2$ (see Fig. 1). In active inference, there will also be learnt associations between states of the blanket, for example between active state $a_1$ and sensory state $u_2$, given internal states $\lambda_1$.

This means that even though action is part of the story we can still in principle understand all that the brain does simply in terms of approximate inference and while 'throwing away the world' beyond the Markov blanket (i.e., all the $x$ nodes in Fig. 1). Naturally, 'throwing away the world' should be taken in an explanatory rather than literal sense. For the purposes of explaining mind and cognition, we just need to know how the Markov blanket system — the brain in this case — constructs these probabilistic mappings among its own states (namely by the message passing underlying perceptual and active inference in the cortical hierarchy). We do not need to know what the worldly states are, even though, obviously, they exist and play a causal role via the interface of the blanket in causing the internal states to be the way they are.

Given this internalist perspective, there is no reason to think that adding action to the prediction error scheme will make any immediate difference to the issue of scepticism. Scepticism arose as a consequence of the fact that prediction error minimization occurs within the sensory veil, and active

inference does not change this fundamental aspect — active inference does not break down the Markov blanket. In fact, the existence of the Markov blanket is evidenced by action.

The free energy principle's marriage of perception with action thus has little prospect of avoiding the traditional internalist picture of mind. This is an intriguing observation because, in other respects, there is much in common between the free energy principle and the EEE approaches to cognition (Friston 2011; Bruineberg and Rietveld 2014; Bruineberg 2016; Clark 2016a; Hohwy 2016b; Kirchoff 2016). However, at the same time, the free energy principle seems at odds with those approaches since it absorbs action within the Markov blanket. The free energy principle is unlikely to be the scheme that will be the perfect partner for existing EEE approaches to the mind. It may be that a new type of understanding is needed which can combine — synthesise — insights from both the EEE approaches and more traditional internalist approaches under the free energy principle.

## 5 Epistemic Advantages of Action

In the setting of the free energy principle, action cannot help evade scepticism — and the sandwich model of the mind — in the way some EEE approaches might have hoped for. This is not to say that there are no epistemic advantages of action. For example, it is a commonplace that we can act to explore the world and find out how it works. There exist in the literature considerations about whether action can deal with scepticism. These considerations happen within a traditional, inferential framework rather than by obliterating the sensory veil through EEE approaches. Consider here Reichenbach's cubical world (Reichenbach 1938, § 14; Fig. 2), which speaks to the kind of Markov blanket scenario we are considering.

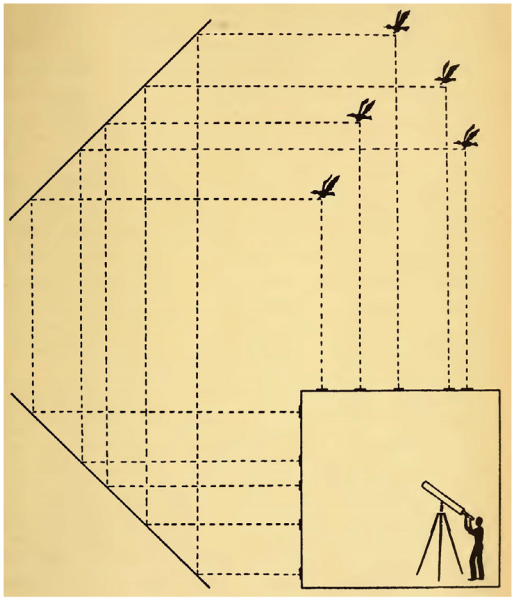

**Figure 2.** Reichenbach's cubical world (Reichenbach 1938, p. 117). The observer in the cubical world must infer the hidden causes (flocks of birds) on the basis of their reflections cast on the semi-transparent walls of the world, some directly, some cast indirectly via mirrors.

The observer is locked in a room with translucent walls (conceive of this as the Markov blanket) on one of which the reflections of five birds are projected. At the same time, a set-up of mirrors projects the reflections of the five birds to another wall, giving a total of ten reflections. The observer must infer the causes of their reflection input without leaving the cubical world. The question is whether the observer can select between the hypotheses that there are five birds and that there are ten birds, and,

further, whether anything in the situation could convince the observer that there are any birds out there at all.

Reichenbach argues that the coincidence between the movement of the two sets of reflections is evidence that there is a common cause (i.e., five birds), and that if there is a common cause it must be beyond the cube. Sober (Sober 2011) argues that this is not efficient, since it presupposes something not given in the evidence, namely that common cause is more likely than coincidence given the observation of coincidence. This prior seems to presuppose too much.

Sober then appeals to interventionist approaches to causal inference. According to interventionism, it is possible to distinguish common causes from coincidence by intervening — acting — on the relevant states of affairs (by 'surgically' varying and holding fixed certain random variables). By acting to hold one of two coinciding variables constant one may discover they are not related as cause and effect, namely if the behavior of the other is not changing after the intervention. This is some evidence that they have a common cause, evidence which is unlikely to be obtained without intervention. Though this is epistemic progress through action, it does not however exclude the possibility that this common cause is somehow internal to the cubical world (or Markov blanket), so it cannot speak to external world scepticism; for example, Sober discusses the possibility that the common cause is a prior intention. Neither can it rule out the evil demon possibility, which is indeed a common cause of sensory input.

Notice here that the intervention is on the hidden causes in the world, conducted via our active states (e.g., brain states entraining our limbs so we move the hand to scare the birds) and detected via our sense organs (e.g., noticing whether there is a change in all the reflections). All the perceptual and active inferential work is done within the Markov blanket.

These kinds of examples show that action can have epistemic value. There are certain questions in causal inference that cannot easily be answered without action. This role of action is part and parcel of active inference and the free energy principle. This corresponds to how active inference was described above; action can minimize prediction error in two ways, to maximize utility (help us occupy expected, low prediction error states) and to maximize epistemic value in the ways discussed in the interventionist framework (Hohwy 2013, Ch. 4, Friston et al. 2015). What we are considering here is active inference for epistemic value about the external causes of sensory input and their relations amongst each other, to our sensory organs, and from our active states.

Even if actions can make such inroads on some epistemic issues for agents considered in terms of Markov blankets, it is no solace to the EEE approaches since it remains a wholly internal, inferential approach. And of course none of this speaks to the evil demon scenario where the issue is not whether the external world exists but whether the external world harbours a demon or not. It is difficult to see how any intervention could increase the probability that the sensory input is not caused by an evil demon since the demon, we must assume, will make sure the sensory input does not reveal its own existence.

## 6 How Action Entrains the Environment

I have argued that action cannot be used to extricate the prediction error minimization scheme from the sceptical scenario: action does not obliterate the sensory veil nor can it be used to favour directly the non-sceptical hypothesis.

Consider however what happens during active inference. For example, the internal states predict that the proprioceptive input is of the type that occurs when the arm raised. Since the arm, let us assume, is not currently raised, this prediction leads to prediction error. The proprioceptive prediction, at the active states of the Markov blanket, triggers reflex arcs to move the limbs around until the prediction is satisfied at the sensory states of the blanket, registering the occurrence of the expected proprioceptive input. Here hidden states external to the Markov blanket (the limbs) are causally af-

fected by states internal to the blanket, and in turn cause changes to the blanket and its internal states. This portrays the central dynamic nature, or circular causality, of the free energy principle, see Fig. 3 (notice that formally speaking causal nets apply to acyclic graphs, not cyclic ones as depicted in Fig. 3.; there are ways around this, for example by appealing to dynamic Bayes nets).

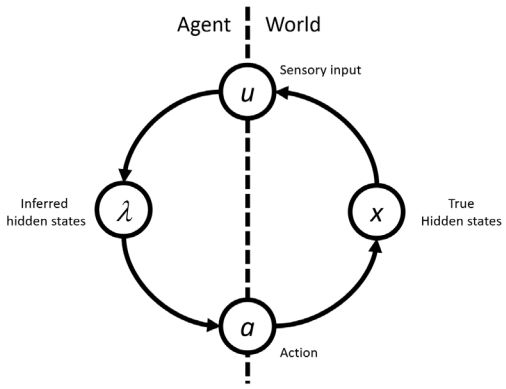


**Figure 3.** A simplified version of the Markov blanket shown in Fig. 1 but with depiction of circular causality induced by active inference. The dotted line marks the mind-world divide or Markov blanket; $u$ are the sensory states of the Markov blanket, $a$ are the active states and $\lambda$ the internal states; $x$ are the true hidden states of the world.

In this scenario, no evidence available to the agent distinguishes between the hypothesis that they really have an arm and the hypothesis that an evil demon is deceiving them to think they have an arm. Note that the internal states of the agent are part of a dynamic causal chain that modulates the states of the hidden causes. This holds whether the external states harbor real arms and other familiar things, or a demon. We are assuming the world contains either of these, and that they causally impact the agent's sensory states. In the demon world, this means the demon's states causally entrain the agent's internal states. But equally, through active inference, the agent's states causally entrain the demon's states. If the demon is not entrained like this, then the agent's prediction errors would not be minimized in active inference. (I am setting aside the 'lucky demon' scenario where the causal link from $a$ to $x$ is cut and yet the demon luckily guesses what sensory input the agent predicts; I am also setting aside the solipsistic scenario where there is a direct causal link from $a$ to $u$; similarly, I am setting aside the case of an 'active Laplacian demon' who manipulates $u$ on the basis of perfect knowledge of the laws of nature and power to set the initial conditions of the world).

Put slightly differently and assuming our demon is an agent with its own Markov blanket, just as the demon's input to the agent works like a learning signal for the agent's internal models (through prediction error) so the agent's output to the demon works like a learning signal for the demon's internal states. The demon thus infers the internal states of the envatted agent, and it acts to minimize its own prediction error. The demon's internal states must in turn have causal power in the world so that they can affect the agent's sensory states in the predicted ways. In very basic terms, here the agent and the demon begin to 'oscillate' together — they are essentially locked together in a system of coupled oscillation. As Friston points out, this is basically a version of the pair of pendulum clocks described in the 1600s by Huygens' (Huygens 1967, p. 185), which when hung from a beam in the right kind of set-up will eventually begin to swing synchronously even if separated by each their own Markov blanket (Friston 2013; Fig. 4).

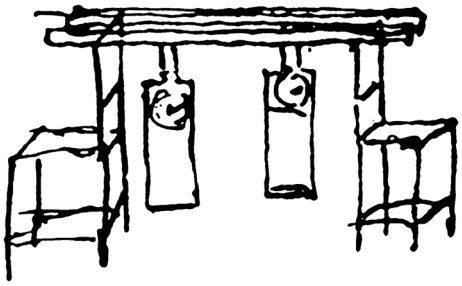

**Figure 4.** Huygens' sketch of coupled oscillation between two pendulum clocks.

Coupling the agent and the demon like this levels the playing field between them. They are tied in a causal dance of give and take. It may be that the demon started things off by giving the agent its basic expectations (as evolution arguably does in the non-demon case) but this is not tantamount to knowing a priori what active inference the agent will engage in since the agent is a self-organising, noisy system with some autonomy, individual learning history, and ability to vary parameters in approximate Bayesian inference. For example, the agent may experiment with the balance between perceptual and active inference in unanticipated ways; sometimes it may optimize the prediction error bound on surprise in a sustained manner before engaging action, other times it may jump to action before perceptual inference has been fully optimized; similarly, at times the agent may act on the expectation that exploration of the free energy landscape is the best way to keep long term error low (Hohwy 2013, Chs. 4, 7). Hence the demon can do nothing but follow suit to the agent's attempts to change the evidence through action to fit its predictions. The only alternative avenue for a recalcitrant demon is to let the prediction error of the agent increase. In the short run, this will just cause the agent to optimize its internal model before it acts, rather than jump to active inference. In the long run, this choice is however only open if the demon wants the agent to perish since, as the free energy principle sets out, the existence of the agent depends on continually acquiring evidence for itself.

Above we encountered the discussion of scepticism by the enactivists (Thompson and Cosmelli 2011). Their point was that for a demon to actually operate a brain-in-a-vat it seems nomologically necessary that the brain is coupled with a body, or something that in some respect is functionally equivalent to a normal body, in an environment conducive to the brain's autonomous homeostatic functioning. There is something right about this, in the sense that external causes, harboured in the demon and its world, must exist for the agent to exist. More, Cosmelli and Thompson appeal to the dynamics of bodily existence, which makes it imperative for the demon to provide a body-like environment. In this sense, their argument provides a more rudimentary, body-bound version of the more general appeal to active inference that I am rehearsing here. Cosmelli and Thompson argue that their approach entails that the mind is enactive and embodied rather than brain-bound, even on the demon scenario; they use this to propose that scepticism in some sense is self-undermining. However, as argued above, if there is a Markov blanket, then this appeal to the necessary existence of body-like external causes does not make the sceptical scenario self-undermining. It is a given that there are external causes including those that keep the body and brain of the agent alive, but the evil-demon scepticism remains regardless.

Consider now what this kind of causal coupling, which I have argued follows from the free energy principle, entails about the epistemic and representational state of the agent. The predictions of the agent are based on its internal hierarchical model of the external world. This model carries information about statistical relations and causal interactions amongst the modeled causes at several interlocked time-scales. For example, it will represent that arms can be raised to greet friends or hail cabs except when one is very tired or something heavy is placed on the arm, and so on. This representation is

specified in the hierarchy by having several levels of inferred, interacting causes that when convolved in the right way best minimize long term prediction error and thus explain away the sensory input.

If we consider this model as a large 'theory' specifying all these statistical and causal relations, then nothing much is lost by stripping away the names of properties (i.e., the property names 'arms', 'friends', 'cabs' etc.) from the statistical and causal information and instead operating with existential quantification over the properties. This leaves behind a statistical/causal functional role that can be used to implicitly define what those properties are. For example, for there to be an 'arm' is for there to be something that has the role of probably being raised when 'friends' are nearby, and which tends to interact causally with 'tiredness' and 'heavy objects', where 'friends', 'tiredness' and 'heavy object' are in turn interdefined by their own partly overlapping statistical/causal roles (see Hohwy 2013, Ch. 8, and references therein for this view, which is inspired by standard functional role semantics). Conceiving the internal model like this reflects that what makes prediction error minimization succeed is only information about the causal-statistical properties of the input, rather than about the intrinsic aspects of the objects themselves.

Conceived like this, there is no difference between the internal representations of the agent in the demon world and in the non-demon world. The statistical/causal roles harboured in the internal model of the agent are they same, and both are blind to the difference between real arms, friends and cabs and the simulacra of these in the demon scenario. These statistical/causal roles are however what gives the internal states their content, and it seems their satisfaction conditions are the same in the two worlds. The demon world will have the action-induced statistical/causal properties, and the definite descriptions defining those properties will quantify over some of the demon's hidden causes.

From the point of view of the free energy principle, there is a sense in which the agent need not care about the fact that their internal model fails to distinguish the demon and non-demon possibility. All the agent cares about is self-evidencing — maintaining itself in its expected states — and as long as prediction error is kept low it succeeds in doing this. Since what the agent needs to help keep prediction error low is statistical/causal information, the functional role conception of their internal model is apt. Ontologically, that is, there may be an evidence-transcendent difference, for example in terms of arms vs. non-arm demon properties that can have no effect on statistical/causal relations (this echoes discussion of Matrix-type scenarios in Chalmers 2005, but set in the context of functional role semantics and active inference).

Epistemically, not only can the agent keep the prediction error low, they must also be getting something right about the causal/statistical structure of the hidden causes, even those states that partly originate in a demon. This epistemic success follows from the entraining of the hidden causes through active inference. If prediction error is minimized through action, then there must be something out there such that (modulo the overall level of prediction error minimization and irreducible noise) it stands in the modelled causal/statistical relations to each other and the agent.

If the story is told without appeal to active inference, then the world need not be entrained to the agent's internal states. An entirely passive perceptual system may be able to minimize prediction error over some time scale but must more promiscuously change its internal model in whatever way will explain away the sensory data. In such a system, which does not bend the world to its expectations, organisms (if any should endure at all) are more short-lived and there is more scope for false, hallucinatory models.

Of course, the causal/statistical epistemic success of an agent engaging in active inference is consistent with considerable accompanying ignorance. The agent's unintentional social demon cognition may fall short; for example, in the demon world there is a deeply hidden cause that is not modelled successful, namely some of the demon's own mental states (such as the demon's desire to keep the agent alive). Similarly, in the non-demon world, there are probably levels of natural law governing deep seated regularities, which are nevertheless as yet undiscovered or even undiscoverable for us.

## 7 Concluding Remarks

Even if an evil demon is causing an agent's sensory input, the agent's actions entrain the demon to cause less surprising sensory input. This leaves the skeptical scenario in place but dulls some of its epistemic sting. The demon world and the non-demon world must have overlapping statistical/causal structure, given the agent's internal model, and the agent can know this structure even if the intrinsic structure and some more deeply hidden causes remain unknown.

As far as competing models of the mind goes, we saw that skeptical scenarios are viewed as a symptom of undesirable internalism that ignores the role of action, embodiment, and extension of mental states into the external world. Some embodied, extended and enactive (EEE) approaches on the other hand avoid the sceptical scenario but arguably at the price of undermining familiar notions of knowledge, belief and justification.

A prediction error minimization scheme is wedded to self-evidencing in the context of Markov blankets and thereby to a principled way of distinguishing mind from world in terms of internal inference of the hidden causes located externally to the Markov blanket. Such a scheme can accommodate action, through the free energy principle. As argued in this chapter, however, including action cannot change the scheme's staunchly internalist nature.

Accommodating action does have an interesting consequence, of relevance to the demon scenario. The actions of a creature who manages to persist over time must entrain the causes in its world (if any such causes exist), whether these are the familiar causes we expect to exist in the world, or whether these are causes harboured in an evil demon. We also saw that since active inference must entrain the causes of the external world, a free energy minimizing system can know at least some of the statistical/causal structure of the world, whether the demon scenario is true or not. This all adds up to an advantage for the free energy approach to mind. It is an account of the mind that can make reasonable and interesting, albeit limited, inroads on the evil demon scenario and, all the while, provide a unified if internalist and inferentialist explanation of perception and action.

This advantage for the free energy approach is achieved by going in what seems to be the opposite direction from EEE approaches. Many of those approaches seek to obliterate the Markov blanket, and attempt to make perception less inferential, more like action and more connected to the world. The free energy approach instead makes action inferential and conceives it as just a matter of internal processing within the Markov blanket.

It would be a mistake to view this as entirely anathema to the EEE approaches. The internalist and inferentialist view does not conceive of the mind or the brain as causally insulated from the world around it. Indeed, this view must conceive of the mind and the world as causally linked, through the causal interface of the Markov blanket, since perception is inference on the causes of sensory input and active inference entrains these causes to fit the internal model, leading to coupled oscillation and reciprocal mirroring of mind and world.

Similarly, under the free energy principle, agents are conceived in terms of the states of the world that they tend to occupy, under the assumption that their Markov blanket will begin to disintegrate, and the agent begin to disperse, if action ceases to minimize prediction error. The mind of an agent is thus just another set of causes in the overall causal nexus, albeit a self-organising type of cause. So the free energy principle delivers a dual perspective that seems to supersede some of the existing debates: epistemic insulation, exemplified with the sceptical demon scenario, goes hand in hand with more embodied, situated causal integration or oscillation (Hohwy 2013, p. 228). In this light, the free energy principle seems to offer progress — a sort of synthesis of internalism and EEE — in our debate about mind and world: it provides a unified perspective on the epistemic and causal status of the mind.

## References


Anderson, M. & Chemero, A. (2013). The problem with brain GUTs: Conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral & Brain Sciences, 36*, 204–205.

Bruineberg, J., Kiverstein, J. & Rietveld, E. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological enactive perspective. *Synthese*. https://dx.doi.org/10.1007/s11229-016-1239-1.

Bruineberg, J. & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience, 8*. https://dx.doi.org/10.3389/fnhum.2014.00599.

Chalmers, D. J. (2005). The matrix as metaphysics. In C. Grau (Ed.) *Philosophers Explore the Matrix*, Oxford: Oxford University Press.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension.* Oxford University Press, USA.

——— (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral & Brain Sciences, 36* (3), 181–204.

——— (2015). In T. K. Metzinger & J. M. Windt (Eds.) *Embodied prediction. Open MIND:* 7(T). Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570115.

——— (2016a). Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs*. https://dx.doi.org/10.1111/nous.12140.

———(2016b). *Surfing uncertainty.* New York: Oxford University Press.

Eliasmith, C. (2000). *How neurons mean: A neurocomputational theory of representational content.* Ph.D., Washington University in St. Louis.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks, 16* (9), 1325-1352.

Friston, K. J. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society London: Biological Sciences, 369* (1456), 815–836.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews. Neuroscience, 11* (2), 127–138. https://dx.doi.org/10.1038/nrn2787.

———(2011). Embodied inference: Or "I think therefore I am, if I am what I think". *The Implications of Embodiment.* W. Wolfgang Tschacher and C. Bergomi. Sussex, Imprint Academic.

———(2013). Life as we know it. *Journal of The Royal Society: Interface, 10* (86). https://dx.doi.org/10.1098/rsif.2013.0475.

Friston, K. & Stephan, K. (2007). Free energy and the brain. *Synthese, 159* (3), 417-458.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T. & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, 1-28. https://dx.doi.org/10.1080/17588928.2015.1020053.

Gallagher, S. (2005). *How the body shapes the mind.* Oxford: Oxford University Press.

Grush, R. (2003). In defense of some 'Cartesian' assumptions concerning the brain and its operation. *Biology and Philosophy, 18* (1), 53–93. https://dx.doi.org/10.1023/A:1023344808741.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science.* New York: Free Press.

Hohwy, J. (2010). In W. Christensen, E. Schier & J. Sutton (Eds.) *The hypothesis testing brain: Some philosophical applications* (pp. 135–144). Macquarie Centre for Cognitive Science. https://dx.doi.org/10.5096/ASCS200922.

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

——— (2015). In T. Metzinger & J. M. Windt (Eds.) *The neural organ explains the mind* (pp. 1-23). Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570016.

———(2016a). Prediction, agency, and body ownership. In: *The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science,* ed. A. K. Engel, K. J. Friston, and D. Kragic. Strüngmann Forum Reports, vol. 18, J. Lupp, series editor. Cambridge, MA: MIT Press.

———(2016b). The self-evidencing brain. *Noûs, 50* (2), 259-285. https://dx.doi.org/10.1111/nous.12062.

Hurley, S.L. (1998). *Consciousness in action.* Harvard University Press.

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

Huygens, C. (1967). *Œuvres complétes.* Amsterdam: Swets & Zeitlinger.

Kirchoff, M. (2016). Autopoiesis, free energy, and the life–Mind continuity thesis. *Synthese*. https://dx.doi.org/10.1007/s11229-016-1100-6.

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity.* MIT Press (MA).

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics, 66*, 241–251. https://dx.doi.org/10.1007/BF00198477.

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

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufmann Publishers.

Reichenbach, H. (1938). *Experience and prediction—An analysis of the foundations and structure of knowledge*. Chicago: University of Chicago Press.

Sober, E. (2011). Reichenbach's cubical universe and the problem of the external world. *Synthese, 181* (1), 3–21. https://dx.doi.org/10.1007/s11229-009-9593-x.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard: Harvard University Press.

Thompson, E. & Cosmelli, D. (2011). Brain in a vat or body in a world? Brainbound versus enactive views of experience. *Philosophical Topics, 39*, 163-180.

Usher, M. (2001). A statistical referential theory of content: Using information theory to account for misrepresentation. *Mind & Language, 16* (3), 311–334. https://dx.doi.org/10.1111/1468-0017.00172.

# How to Knit Your Own Markov Blanket:

## Resisting the Second Law with Metamorphic Minds

Andy Clark

Hohwy (Hohwy 2016, Hohwy 2017) argues there is a tension between the free energy principle and leading depictions of mind as embodied, enactive, and extended (so-called 'EEE[1] cognition'). The tension is traced to the importance, in free energy formulations, of a conception of mind and agency that depends upon the presence of a 'Markov blanket' demarcating the agent from the surrounding world. In what follows I show that the Markov blanket considerations do not, in fact, lead to the kinds of tension that Hohwy depicts. On the contrary, they actively favour the EEE story. This is because the Markov property, as exemplified in biological agents, picks out neither a unique nor a stationary boundary. It is this multiplicity and mutability– rather than the absence of agent-environment boundaries as such- that EEE cognition celebrates.

**Keywords**

Active inference | Autopoiesis | EEE cognition | Embodied cognition | Evil demon | Extended mind | Free energy minimization | Markov blanket | Prediction error minimization | Process ontology

**Acknowledgements**

Thanks to Karl Friston and two anonymous referees for extremely helpful comments on an earlier draft, to John Dupré for invaluable discussions concerning the extended mind and process ontologies, and to Giovanna Colombetti and Joel Krueger for participating in and making possible the 'Breaking Boundaries' symposium at the University of Exeter, where some of these ideas started to take shape. This paper was drafted during a period of sabbatical leave (Autumn 2016) kindly granted by the University of Edinburgh, and completed as part of ERC Advanced Grant XSPECT - DLV-692739.

"My cousin has great changes coming – one day he'll wake with wings"
Cousin Caterpillar (The Incredible String Band)

## 1 The Markov Blanket Conception of Mind

Markov blankets are named after Andrey Markov (1856-1922), a mathematician whose seminal work explored abstract systems that remember their past trajectories only insofar as they store a single (current) value. In such systems (Markov chains – see Norris 1998) the next state depends only on the value of the current state. This is the so-called Markov *property*. For this reason, such systems are sometimes said to be 'memoryless'.

Now consider a complex system composed of many interacting nodes (variables). Pearl (Pearl 1988) introduced the term 'Markov blanket' to describe the set of nodes such that, for some given node X, the behavior of X could be fully predicted just by knowing the states of those other nodes. The states of those neighbouring nodes thus fix (statistically, not causally) the state of the target node conditionally independently of all the other states of the system, forming a 'Markov blanket' that shields

1 Many of the core ideas of EEE cognition are lately referred to using the even larger grouping '4E Cognition' (see e.g. Newen et al. in press) rather than EEE cognition. This adds 'embedded', in the sense defended by (Rupert 2009), to the E-pantheon. I have used the 'EEE' branding partly because that is the label used by Hohwy (Hohwy 2016, Hohwy 2017) in the 'evil demon' papers to which the current work responds. 'EEE Cognition' also provides the most apt label for the contested issues, as it is claims associated with embodied, enactive, and extended (rather than merely 'embedded') cognition that are the locus of our residual disagreements concerning the conceptual implications of predictive processing.

Clark, A. (2017). How to Knit Your Own Markov Blanket: Resisting the Second Law with Metamorphic Minds.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 3. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573031

the target node from the rest of the activity in the system. The practical upshot is that, to predict the state of the target node, all you ever need to know are the states of the nodes that form its Markov blanket[2]. The Markov blanket comprises the so-called 'parents' and 'children' of the blanketed node or nodes, corresponding to the most proximal actors upon the node (the parents) and the most proximal 'acted-upons' (the children), along with whatever else acts upon the children (the parents of the children). Markov blankets can be redundant, in that a target node may be enclosed within many Markov blankets. Markov blanketed organizations may nest within larger Markov blanketed organization (see figure 1), a property that will be important for the treatment that follows. Finally, a Markov *boundary* (Pearl 1988) exists when a Markov blanket for a node has no proper subset that is also a Markov blanket for that node – it is thus the most minimal (or 'non-redundant') Markov blanket for the node.

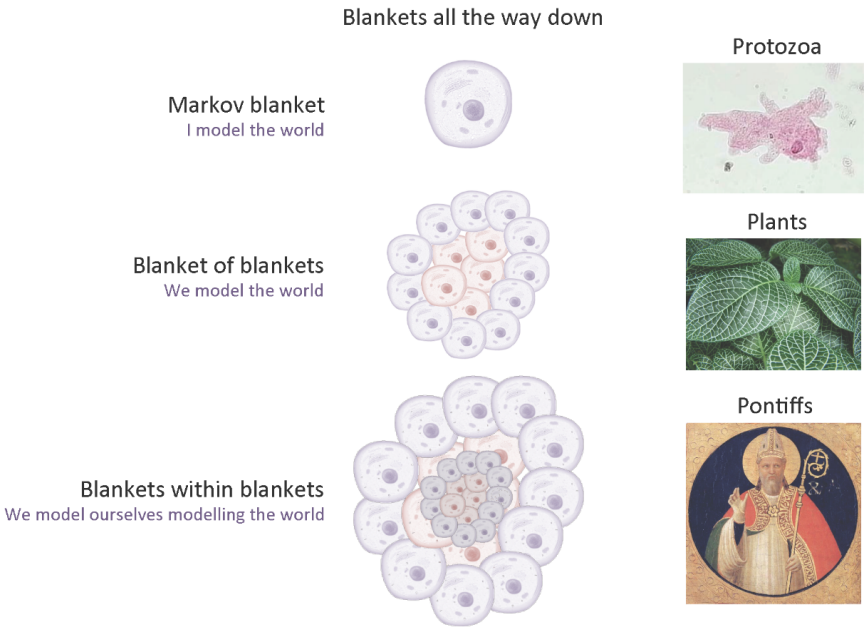


**Figure 1:** Blankets of blankets of blankets. Image by Karl Friston (by permission).

Hohwy's (Hohwy 2016, Hohwy 2017) arguments for a form of 'neurocentric seclusion' depict the states of our biological sensory systems as determining a Markov blanket that (he claims) defines the boundaries of the mind. Thus we read that:

> the mind begins where sensory input is delivered through exteroceptive, proprioceptive and interoceptive receptors and ends where proprioceptive predictions are delivered, mainly in the spinal cord. (Hohwy 2016, p. 276)

Adding in a footnote (footnote 14) that:

> In more technical terms (see Friston 2013), the sensory input and active output at this boundary forms a so-called Markov blanket (Pearl 1988) such that observation of the states of these parts of the system, together with observation of the prior expectations of the system in principle will allow prediction of the behavior of the system as such. Causes beyond this blanket, such as bodily states

2 This does not imply that, just by knowing the states of the blanket, a theorist could perfectly predict the evolution of the target (contained) nodes. This is because the blanketed system may have its own intrinsic dynamics. Thanks to Wanja Wiese for suggesting this clarification.

> or external states, are rendered uninformative once the states of the blanket are known. (Hohwy 2016, p. 283; citation style adapted)

The mind, if this is correct, is firmly bounded by the sensory flows associated with exteroceptive, interoceptive, and proprioceptive signaling. Step beyond those sensory flows and (Hohwy argues) the Markov property bites, rendering all that led up to those sensory stimulations statistically otiose as a predictor of systemic response. That also means that the mind, thus conceived, has no access to states of the world beyond that provided by the evidence available at the Markov blanket. Such minds "will not be able to distinguish between possibilities where similar flows of sensory input are caused by two very different causal processes, beyond the blanket" (Hohwy 2017, sect. 3). Hence they might be fooled by a clever Cartesian demon, or by the keepers of the Matrix.

## 2 The Markov Blanket Conception of Agents

Hohwy (Hohwy 2016, Hohwy 2017) also offers the Markov blanket construct as a formalization of some important claims concerning the nature and definition of agents. The considerations here weave together the Markov blanket considerations and the free energy principle (FEP) described in (Friston 2010), (Friston and Stephan 2007) and elsewhere.

FEP (more on which shortly) leverages the simple truth that agents that exist do so because they are able to persist, appearing to resist – for periods of time - the second law of thermodynamics that states that entropy (disorder) increases over time[3]. Biological agents are able to resist because the second law applies only to isolated (or closed) systems. By exchanging matter and energy with the environment, such systems are able to preserve their own integrity and order. They do so, of course, only by increasing disorder elsewhere (thus 'obeying' the second law). We thus enter the realm of living or adaptive systems – systems that actively seek out, and work to bring about, the conditions that are necessary for their own survival[4].

Such agents are, in an important sense, defined by the particular way they resist disorder. A specific type of living agent simply *is* a set of states that maintain themselves within certain bounds – the bounds that describe the conditions necessary for their own survival. Such bounds include, for human agents, acceptable ranges of body temperature, the production of glucose to power the metabolism that enables foraging, and so on. Only while a whole host of such states remain within set or tolerable ranges does the agent (qua that very living being) exist. This adaptive tautology implies that, in a very broad sense, every living creature must visit and revisit the set of states that define it as the creature it is. In this sense, a living organization is said to be (at least locally) 'ergodic' – see (Friston 2013, sect. 4).

This is where FEP gets invited onto the stage. FEP states that living organisms that persist must minimize free energy in their exchanges with the environment. The 'free energy' in question here is an information-theoretic isomorph of thermodynamic free energy, which is a measure of the energy available to do useful work. Useful work, in the information-theoretic story, involves fitting a model to a domain, so reducing information-theoretic free energy is improving the model. The free energy is then a bound on the long-term average 'surprisal' (Tribus 1961) associated with environmental engagements, where this names the implausibility of some sensory state given a model of the world. Entropy, in this information-theoretic rendition, is the long-term average of surprisal. Reducing information-theoretic free energy thus amounts to improving the model so as to reduce (long-term average) surprisal. Organisms that minimize long-term average surprisal will, by definition, appear to resist the second law of thermodynamics. They will take steps to avoid environmental states and encounters that would cause them to undergo catastrophic ('surprisal-ing') phase-transitions. They

3 More accurately, the second law tells us that entropy in an isolated system will either stay the same or increase over time. It can remain constant only in certain idealized (e.g. 'steady-state') cases. But whenever energy is transferred or transformed, entropy increases.

4 This formulation suggests – correctly in my view – strong links with the notion of 'autopoesis' (Varela et al. 1974). See also section 3 following.

achieve this by being structured in ways that in effect 'exchange entropy' with the environment, allowing them to self-organize so as to avoid, for a while at least, the kinds of catastrophic encounters that would cause them to cease to exist.

Hohwy links the FEP to the Markov blanket considerations by identifying the free energy minimizing agent or model with the internal states screened off by the Markov blanket. It is in this sense that, as Hohwy (Hohwy 2017, sect. 2) puts it "the internal, blanketed states constitute the model". It is important to note that the free energy minimizing 'model' here is that which is bounded by sensory and active (action-causing) states. This model is then said to be self-evidencing in the sense of (Hempel 1965). Self-evidencing is typically exemplified by cases such as the following:

> the velocity of recession of a galaxy explains the redshift of its characteristic spectrum, even if the observation of that shift is an essential part of the scientist's evidence that the galaxy is indeed receding at that the specified velocity. (Lipton 2001, pp. 44–5)

But in the case at hand, the very existence of a whole living system (a plant or an animal) provides evidence for itself considered as a surprisal-minimizing model. In other words, the goodness of the system as a means of exchanging entropy with the environment so as to persist in the face of the second law[5] is (self) evidenced by its own existence.

Notice that nothing in this Markov blanket conception of biological agents requires those agents' brains or control systems to engage in online prediction error minimization at all[6]. Hohwy's treatment (Hohwy 2017) thus covers two issues that – though deeply related - are also importantly distinct. The first concerns the status of living systems as self-evidencing free energy minimizing systems. The second concerns the specific vision of the human brain as implementing a process of prediction error minimization.

Thus consider a very simple free-energy minimizing life-form, such as a single-celled organism capable of survival-enhancing chemotaxis. Such a life-form may respond to environmental perturbations using a variety of tricks and ploys, none of which require it to engage in a process in which incoming sensory stimulations are met with attempts to generate the incoming signal 'from the top down' using stored knowledge about the world. Such a being, though living and perfectly able to resist the second law by exchanging entropy with its environment, could be operating in a purely 'feed-forward' manner, responding to detected chemical gradients in ways not nuanced by any form of top-down predictive flow. Talk of such a being 'predicting' such-and-such, or 'minimizing prediction error with respect to such and such' is either simply false or merely short-hand for what is really a rather different claim – the claim that the creature is structured so as to favour the kinds of environment necessary for its own persistence.

To describe this whole simple (reactive, feed-forward) creature as a 'model' of its world, though common in this literature, can also seem somewhat strained. An intelligent agent might harbour an explicit model of the layout of her own house, and use it to drive various kinds of 'offline' reasoning (such as counting the windows while thousands of miles away). An embodied agent might also use parts of her own body as a model, for example by counting the windows using her fingers. These are core and familiar uses of the notion of a model[7].

5 Or more precisely, in the face of the 'fluctuation theorem'. This applies to far-from-equilibrium systems (such as living beings), and has the second law as a special case. See (Friston and Stephan 2007).

6 Notice that the mere fact that some creature (a simple feed-forward robot, for example) is not engaging in active online prediction error minimization in no way renders the appeal to a Markov blanket unexplanatory with respect to that creature. The discovery of a Markov blanket indicates the presence of some kind of boundary responsible for those statistical independencies. The crucial thing to notice, however, is that those boundaries are often both malleable (over time) and multiple (at a given time), as we shall see.

7 For example, in everyday use the word 'model' might be used to mean a scaled down replica such as a toy model of a boat - but not the boat itself. If I said 'the boat is a model of the sea' most people would take me to be speaking metaphorically. Yet in the technical sense used by Friston and others, the boat is quite literally a model of the dynamics of the sea. It is wise to keep these differences in mind when assessing the claim that the free energy perspective makes agents into surprise minimizing 'models' of their worlds.

The matter is complicated, however, by the self-evidencing that is inherent in free energy or surprisal minimization itself. To see why, notice that negative surprisal is also the 'Bayesian model evidence' (see Friston 2013) associated with some specific Markov blanket and its internal states. Surprisal is the negative log probability of sensory states and action given that Markov blanket. In turn, this is the Bayesian model evidence for the Markov blanket itself. In this (mathematically quite deep) sense minimizing free energy or surprisal is the same as maximizing model evidence for the existence of the Markov blanket. The free energy principle, considered as an imperative for biological self-organization, thus necessarily entails some form of self-evidencing[8].

The upshot is that any adaptive system (any system that persists in the face of a changing environment) must display self-evidencing, and might properly be described as self-evidencing a certain 'model of the world'. This is because any system that successfully 'deals with' its environment, so as temporarily to resist the second law, counts as modelling its environment in this somewhat technical sense. Nevertheless, models (organisms or systems) that are self-evidencing in this specific sense need not rely upon top-down predictions to structure and inform their exchanges with the wider world. Predictive processing thus constitutes a biologically plausible process theory that may or may not be implemented in any given biological system. For example, there may be simple systems (such as bacteria and viruses) that minimize their surprisal through a genetically pre-configured response to sensory perturbations. By the same token, there may be other more complicated systems that resist increases in free energy or entropy in part by relying upon some form of predictive processing[9].

Hohwy (Hohwy 2017, sect. 3) also asks "what makes internal models internal?" and once again answers by appealing to Markov blankets and self-evidencing. The suggestion seems to be that the real internalist commitment is not to cognitive processing being in some important sense 'brain-bound' (in the sense of Clark 2008). Rather, since the organism itself is now (at least in the specific mathematical sense just described) the model, the notion of the 'inner model' collapses into the claim that there exists a Markov blanket organization separating the whole acting organism from the wider world. But whatever its other merits, this is not a plausible reconstruction of the notion of internalism at issue in the literature on EEE cognition. For EEE theorists do not seek to deny the existence or importance of systemic boundaries blanketing the organism from the wider world. Instead, such theorists (see Clark 2003, Clark 2008) stress the multiplicity, flexibility, and transformability of those boundaries, and the way the choice of what boundaries to stress reflects the explanatory interests and projects of the theorist. This is the main issue to be pursued in the rest of this treatment.

Putting all this together, every biological system is treated (by Hohwy) as a self-evidencing 'model' of the survival-relevant web of interacting hidden causes partitioned on the other side of a Markov blanket constituted by a set of sensory and active states – a blanket that also defines it as the very system (creature) it is. The system tracks the wider world only via the changing states of the Markov blanket, so that "[t]he nodes of the internal model have access to the states of the blanket […] but the model only represents the external causes vicariously" (Hohwy 2017, sect. 3). This also provides Hohwy with the opening for his interestingly twisted form of 'evil demon' scepticism, as we shall shortly see.

## 3 EEE Cognition Revisited

Summarizing his own recommended approach, Hohwy comments that:

> Many of those [EEE – embodied, extended, enactive] approaches seek to obliterate the Markov blanket, and attempt to make perception less inferential, more like action and more connected to

8 Thanks to Karl Friston (personal communication) for extremely useful discussion of this issue.

9 Among those systems, those equipped with generative models that enable them actively and systematically to anticipate how the world will alter in response to their own possible future actions plausibly constitute the sub-class most demanding of the full predictive processing interpretation – see (Pezzulo et al. 2015).

> the world. The free energy approach instead makes action inferential and conceives it as just a matter of internal processing within the Markov blanket. (Hohwy 2017, sect. 7)

What is right about the EEE stories, Hohwy immediately suggests, is just the strong emphasis on causal interactions between mind and world – interactions mediated via "the causal interface of the Markov blanket" and resulting in some kind of "reciprocal mirroring of mind and world".

I have argued elsewhere (Clark in press) that EEE approaches are best seen as opposed to substantial notions of mind-world mirroring – the kinds of mirroring that depicts minds as harbouring rich enough internal recapitulations of reality to enable us to do most of our cognitive work using traditional brain-bound inner models rather than by making the most of the opportunities provided by body, world, and action. But suppose the free-energy minimizing 'model' that does the mirroring is in fact the whole embodied, active organism. We can, if we wish, describe this as an organism's having "transcribed physical laws governing their environment into their structure" or speak of systems that "embed [the laws that govern environmental unfoldings] into their anatomy" (both quotes from Friston and Stephan 2007, p. 422). But the notion of mirroring itself is now hugely emaciated, effectively reduced to that of (non-accidentally) doing whatever it takes to ensure persistence in the face of the second law of thermodynamics. Every simple trick and ploy that has been celebrated, in the EEE literature, as a means of securing adaptive success thus counts (relative to this undemanding conception) as the organisms 'mirroring' their environment[10]. Now, the organism 'mirrors' the environment much as the shape of the boat (recall note 6 above) 'mirrors' the dynamics of the ocean. At this point, some theorists may prefer to drop the appeal to mirroring *tout court*.

Moreover, the emphasis on causal interactions, once all that is taken into account, seems entirely of a piece with even the most radical versions of EEE. For example, citing Wiener's Cybernetics (Wiener 1961), Maturana and Varela's autopoesis (Varela et al. 1974), Chiel and Beer's neuroethology (Chiel and Beer 1997), and Clark's situatedness (Clark 1997), a leading group of situated roboticists write that:

> In this view, adaptive behavior can best be understood within the context of the (biomechanics of the) body, the (structure of the organism's) environment, and the continuous exchange of signals/energy between the nervous system, the body, and the environment. Hence the appropriate question to ask is not what the neural basis of adaptive behavior is, but what the contributions of all components of the coupled system to adaptive behavior and their mutual interactions are (Mohan et al. 2013, p. 17)

The free energy minimizing story, as we briefly saw, delivers a perfect and principled fit with just these kinds of consideration. It depicts the whole embodied organism, appropriately coupled with the larger environment, as the system relative to which free energy (and surprise/prediction error) is minimized. To that extent, it strikes me as a story that counts in favour of core tenets of the EEE conception: not merely one that merely preserves the best of EEE while avoiding mistakes and excesses.

But rather than dwell on these issues, it may be helpful to look a little harder at Hohwy's conception of the EEE project itself. Many EEE approaches, Hohwy commented (op cit, sect. 7) "seek to obliterate the Markov blanket". This, it seems to me, is Hohwy's core reservation. It marks the crucial spot at which EEE approaches, in Hohwy's view, go wrong. But even granted the wide diversity of work that falls under the EEE umbrella, I am not persuaded that, on the whole, they seek to obliterate the Markov blanket at all. Instead, they aim only to reveal something that is highly compatible with the larger story on offer viz that the Markov property, as exemplified in biological agents, picks out neither a unique nor a stationary boundary. It is this multiplicity and mutability – rather than the absence of

10 For useful surveys of many of those tricks and ploys, see (Clark 1997, Pfeifer and Bongard 2006).

agent-environment boundaries as such - that EEE cognition celebrates, as we'll see in more detail in section 4 following.

Indeed, a functional analogue of the Markov blanket construct already plays a key role in seminal work on the 'enactive' approach that stresses the importance of 'autopoiesis' (Varela et al. 1974) – a form of organization in which the constituent processes actively produce the components needed to maintain themselves, and hence maintain the organization itself. Illustrating this, Thompson (Thompson 2007) notes that:

> A cell stands out of a molecular soup by creating the boundaries that set it apart from that which it is not. Metabolic processes within the cell determine these boundaries. In this way the cell emerges as a figure out of a chemical background. Should this process of self-production be interrupted, the cellular components…gradually diffuse back into a molecular soup. (Varela et al. 1974, p. 99)

It would be harder to write a more elegant and compelling description of the importance, for living forms that resist the second law, of a Markov blanket organization than this. Any autopoietic system will, by definition "embody a circular process of self-generation [that] continually re-creates the difference between itself and everything else" (Thompson 2007, p. 99). Compare Friston:

> For example, the surface of a cell may constitute a Markov blanket separating intracellular and extracellular states. On the other hand, a candle flame cannot possess a Markov blanket, because any pattern of molecular interactions is destroyed almost instantaneously by the flux of gas molecules from its surface. (Friston 2013, sect. 2)

The Markov blanket considerations offer a formal and statistical (but, importantly, not intrinsically causal[11]) window onto the space of autopoietic organizational forms. The enactive 'E' is thus potentially in perfect harmony with the FEP.

Perhaps there is something in the embodied 'E' that will cause a problem? Here, it helps to consider a concrete example. In a series of experiments, Havas and colleagues (Havas et al. 2010) used Botox to induce facial rigidity during an 'emotion language processing' task. Subjects were shown sentences such as "your closest friend has just been hospitalized" and asked to push a button when they had finished reading the sentence. The study found that Botox-induced rigidity hindered the speed of processing of emotion language, concluding that "involuntary facial expressions may play a causal role in the processing of emotional language". In a follow-up study, Neal and Chartrand (Neal and Chartrand 2011) showed the reverse effect, demonstrating that improving facial feedback[12] enhanced the speed of processing. Botox-induced facial rigidity thus slowed reading, while gel-induced enhancement speeded it up. At a minimum, such results suggest that the processing of emotion language involves causal-functional loops that run through our own involuntary facial expressions. Loops running through the actual production of these facial expressions (not merely the issuing of neural commands that would normally result in those expressions) thus look to be both functional and integral to the normal processing of such stimuli.

This is a very typical piece of research in 'embodied cognition' (for many more such examples, see Clark 2008). It is also highly compatible with the predictive processing (PP) story itself. It seems very plausible that the role (in emotion processing) of the facial movements is to influence the flow of top-down prediction that attempts to 'explain away' the facial sensory signal. Our own facial expressions (like our own visceral states – see Seth 2013, Pezzulo 2014) simply constitute further sensory evidence

11 The Markov property is defined by statistical, rather than causal, relations between activity in the various inner and outer nodes. It is this fact, ultimately, that opens up a space of interesting and important questions concerning the active creation of new Markov blanket organizations as part of a life-cycle or as the result of lifetime learning and contingent external influences. We turn to these matters in sections 5-7 below.

12 Feedback signals are enhanced when muscle contractions meet resistance, so they used a gel to amplify that effect.

that helps drive the Bayesian machine more rapidly towards conclusions – here, conclusions concerning the meanings of the words on the screen.

Someone might still ask, of course, if this functionally important loop though facial expression is part of the cognitive process itself, or is merely input to that process? Otherwise put, is the facial expression partly *constitutive* of the cognitive process or is its role merely causal – just input to the true cognitive process? Here, the appeal to Markov blankets seems of little help. For we could just as easily ask 'does the actual facial expression lie within or beyond the Markov blanket that marks the boundaries of the mind and the mental?' The point to notice is that translating the question into these terms goes no way at all towards providing an answer! This is because (as we shall see in more detail below) a single adaptive system will typically comprise multiple blanketed organizations – blankets of blankets of blankets. Which ones (if any) should count as tracking, at a given moment, the machinery of mind is precisely the question at issue.

Such debates have (for better or for worse) been central to the philosophical and scientific debate concerning the 'extended mind'[13] – see, for example, a similar debate (Clark 2007) concerning the role of actual physical gestures in the processes of thinking and reasoning. Notice that the 'causal versus constitutive' question makes sense even if all parties are agreed that the inner neural regime is treating the facial states as additional evidence in a Bayesian inference. The question stands – and all the old arguments, pro and con apply - even if the neural contributions are all understood through the lens of Markov blankets, predictions and prediction error minimization.

## 4 Extended Predictive Minds

Hohwy (Hohwy 2016) does not agree. There, he offers three reasons to reject the 'extended mind' claim concerning processes and operations supported by familiar bio-external hardware such as notebooks and smart-phones. The first is that:

> it is far from clear that notebooks and smartphones actually play any part of the functional role set out by PEM. (Hohwy 2016, p. 270)

But what exactly is required for an item to play *part of* that functional role? It seems clear enough that a reliable non-biological systemic element could play a role in enabling a system to minimize free-energy and resist the second law – examples might include cochlear implants and spectacles. The prediction-error minimization (PEM) story is, however, more specific and refers to the specific message-passing scheme depicted in work on hierarchical predictive coding. Deploying that rather specific kind of process-schema is not required, however (as we saw earlier) in order to count as part of an integrated system that resists the second law. A simple free-energy or surprisal minimizing system need not necessarily engage in anything that resembles the top-down use of stored information to predict the current or evolving sensory flux. By extension, not every *proper part* of an integrated free-energy minimizing system (e.g. a cognitive agent) that *does* implement such an online prediction error minimizing process need itself be directly involved in that process. It is plausible, for example, that gross morphology (e.g. the shape of my hand) serves to minimize free energy in my embodied exchanges with the world, while not implementing any part of such an online process. By the same token, a system that minimizes free energy using online prediction error minimizing techniques (a 'PEM system') could be part of a larger free energy minimizing whole that includes multiple sub-systems that do not work that way.

Moreover, even if we decided - for whatever reason - that any genuinely cognitive process must be in some way entangled with prediction error minimization (a very strong claim for which no justification has been provided) that would still fall short of ruling out the notebook. It would fall short

13 See e.g. (Adams and Aizawa 2001, Clark 2008), and many of the exchanges in (Menary 2010).

because the timescale of entanglement would remain to be determined. Thus recall that Clark and Chalmers (Clark and Chalmers 1998), in their flagship treatment of these issues, argued only for the extension of long-term standing (so-called 'dispositional') beliefs – for example, your standing belief that Missouri is one of the United States of America. Most likely, you do not go around rehearsing this sentence even though you count, at all times, as believing it. Our claim was that for this kind of belief, long-term bio-external storage could be on a par with long-term bio-internal storage. The entries in Otto's infamous notebook play (we argued) this very role, under-writing patterns in potential behaviour (e.g. how Otto would answer certain questions if asked) apt for one who holds that belief. But for the notebook entry to actually drive a piece of behaviour in the here-and-now, that information needs to be activated or accessed, potentially as part of a process that involves prediction error minimization along the way. That means that the making use of the bio-external encoding may call upon whatever processes we deem (for whatever reason) essential to here-and-now cognizing.

To illustrate this, let's adapt an example treated at greater length in (Clark 2005). Imagine that one option for long-term bio-storage, in some alien life-form, is to create a stable 'bit-mapped image' – a kind of 'pixel-level' retinal screenshot, tagged with some information to enable later retrieval. Let's further suppose that this bit-mapped image is stored in some relatively inert fashion, e.g. using an alien-biological equivalent of flash memory. When these alien creatures want to recall that event, they may choose (if they wish – we may assume they command a fluid reconstructive bio-memory also) to retrieve the full bit-mapped image. When they do so, the information in the stored image still needs to be interpreted, and hence is now dealt with in much the same way as incoming sensory information.

The bit-mapped 'memory' is thus every bit as stable and 'inert' as the notebook used by Otto in Clark's and Chalmers' (Clark and Chalmers 1998) infamous thought-experiment. But in the aliens, the process of encoding and accessing the stored trace is fully biologically supported. Notice that this is not to deny that memory in us humans is reconstructive and dynamic. The point here is just that it was not conceptually necessary for that to be the case. Had cognitive science found that a few elements of the human bio-memory system were not reconstructive, we would not have concluded that those elements failed to form part of human cognizing. The imaginary aliens are meant to help reveal this fact. The extended mind theorist argues that, in the alien case, we would have no hesitation in counting the bit-mapped encodings as part and parcel of the alien life form's cognition. But if so, then the fact that something very similar happens in other beings, such as ourselves, using a combination of biological and bio-external (e.g. notebook-based) ought not – she argues - be treated in any different way. For the functional upshot is relevantly similar. So it can only be skin-and skull prejudice that stops us extending the same courtesy to many of our own bio-external resources.

Might the relative inertness of a resource itself be a problem for the claim that it forms part of a cognitive economy that enables a creature temporarily to resist the second law? I don't think so. For a relative inert systemic element may nonetheless help a larger system resist the second law. As a clear example, consider birds that swallow stones and grit to help them digest their food. These so-called 'gizzard stones' (Solomon et al. 2002, p. 664) are relatively inert but function to help grind the food, rendering it digestable and thus helping the birds stay alive. Likewise for bit-mapped images and trusty notebooks – they may be relatively inert when not in use yet function, within the larger system, so as to help enable apt adaptive response.

Such, in a nutshell, is the core argument for the 'extended mind'. I will not here rehearse the layers upon layers of further argument that might ensue (again, see Clark 2005, for the full story). Instead, let's also imagine that the alien life-form's brain is, in all other respects, a predictive processing device. Moreover, let's assume that at the time of encoding and at the time of use, the use of the bit-mapping faculty is selected and orchestrated by the core PP principles of precision-weighted prediction error minimization. Thus the choice of when to use the bit-mapping faculty is itself responsive to systemic best-guessing about what means of storage (and what kind of current retrieval) will reduce the greatest amount of sensory uncertainty in the long term. In other words, the prediction-free strategy is deli-

cately embedded in a prediction-rich economy. It would be wrong (I submit) to argue that the bit-map storage faculty is not part of the overall cognitive economy of that creature. The moral is that not every *part* of the full cognitive economy needs itself to display the full PEM profile. And once this is allowed (as it surely must be) for the imaginary, bio-internal case of the alien life-form, I see no principled reason to disallow the bio-external analogues available to creatures such as ourselves. Thus we should reject Hohwy's assertion that:

> The challenge is to specify the role of notebooks or smartphones, or any other thing, such that it clearly plays an appropriate prediction error minimization role. (Hohwy 2016, p. 270)

We should reject this because the very most that could reasonably be required is that such a resource be appropriately embedded within some larger prediction error minimizing economy. Alternatively, it might be suggested that the 'embedding' scenario I described, since it appeals to the long-term error-minimizing effects of the use of the notebook or bit-mapping, shows how very easy it is to achieve the right degree of integration to count as 'playing part of the PEM functional role'. Either way, the prima facie case against the notebook is defeated.

Hohwy's second worry is that cases such as the notebook raise questions concerning the evidentiary boundary for the self-evidencing model. The thought is that:

> This boundary should make it clear that prediction error is minimized for a system including the external object to which cognition is extended, and with respect to hidden causes outside this extended boundary. (Hohwy 2016, p. 270)

In the case of the notebook, let's assume it includes an entry with an address that matters for doing my job, hence getting paid, hence tending to eat and persist. This, in any civilized society, ought to be a caricature - but you get the idea. The address picks out a place beyond the boundary of the bio-Andy-plus-notebook system. So the relevant 'hidden cause' lies beyond the boundary, just as Hohwy (op cit, p. 270) insists. Is the notebook 'part of the model providing evidence for itself'? I suggest that it is – or rather, I suggest that this is one place we might draw a defensible (though non-unique – more on that soon) boundary. Again, we can motivate this by considering a different, more purely biological, case. Consider the spider and the web. The web, once created, is an organism-external structure that forms a proper part of the free-energy minimizing system. This is the system responsible for adaptive fitness, and it is in exactly this sense that the web is depicted as part of an 'extended phenotype' (see Dawkins 1982) or even as an 'external physiological organ' (Turner 2009).

In just the same way, it seems to me, we should count the notebook as an organism-external structure that is part of a larger free energy minimizing system that, by exchanging entropy with the even larger environment, provides evidence for itself. Hohwy (op cit) also insists that evidentiary boundaries be defined so as to demarcate systems that minimize average prediction error 'in the long run'. This raises complex issues concerning change within the life-cycle, and we return to these in section 5.

Is there, within this larger free energy minimizing whole, a smaller whole relative to which the notebook is external? Surely there is. Similarly, my biological body is heavily reliant on the activity of various colonies of microbes. Indeed, up to 90% of the cells comprising my body are said to be 'microbial symbionts'. Each individual microbe is a free energy minimizing whole, complete with its own Markov blanket, and I would easily survive the loss of one or many of these cells. The lesson (which we return to below) is that complex creatures are composed of layer upon layer of Markov blankets - which layers we choose to emphasize can only be fixed by our local explanatory interests and purposes.

How might we apply this lesson to the matter of locating the mind? Here, the key observation (going back at least to Haugeland 1998) is that we should not simply assume that metabolic boundaries and cognitive boundaries always and everywhere coincide. Just as the boundaries of the liver cell are

not those of the liver, and the boundaries of the digestive system are not those of the immune system, so the bounds of metabolism and bio-sensing need not be those most relevant for sensing (more generally understood) or thinking. Which boundaries we choose (hence which capacities we identify as those of the agent herself) will depend on our wider explanatory purposes.

Hohwy's third and final worry is that:

> There is something unattractive about both acknowledging that an external object (such as a notebook) is represented in the mind's model of the world and insisting at the same time that that object is itself part of some of the mind's mental states. It is unattractive because it means the object is both beyond one evidentiary boundary and within a further evidentiary boundary. (Hohwy 2016, p. 270)

To see why this is not a problem, compare the case of an agent who uses a sensory aid such as spectacles or a blind-person's cane. When in use, the spectacles or cane mark clear (though non-unique) evidentiary limits. My brain minimizes prediction error relative to a flow of evidence that would not be available without the spectacles/cane. But suppose I lose the spectacles or the cane. Now they lie outside the evidentiary limits, and I must minimize prediction relative to an impoverished flow of evidence.

Hohwy goes on to argue that:

> This is not an inconceivable state of affairs but it […] requires that we posit two overlapping yet intimately linked EE [Explanatory-Evidentiary] -circles with different evidentiary boundaries. If there are two EE-circles, then the input to each of the circles will be evidence for the existence of two distinct yet overlapping agents. This may be considered an argument for extended cognition but at the unattractive cost of proliferating the number of agents centered on a particular organism. (Hohwy 2016, p. 270)

But what this really shows, I suggest, is something subtly but importantly different. What it really shows is that the notion of a single persisting agent should not be identified with a stationary set of Explanatory-Evidentiary (EE) boundaries at all. Instead, we should adopt a *process ontology* relative to which a persisting agent can be identified with a rolling process that builds and re-builds its own evidentiary boundaries on the fly. Such agents 'knit their own Markov blankets' in ways that can change over time, without the agent thereby ceasing to exist. It is to this – admittedly challenging – project that I next turn.

## 5 The Multiplicity and Malleability of Markov Blankets

Consider the metamorphic insects[14]. These insects undergo a dramatic change of form as part of their standard lifecycle. Where metamorphosis occurs, the young life-stages do not look or behave in anything like the same way as the adult or mature life-stages. They may eat radically different foods, and locomote in totally different ways. A familiar example is the transformation of a caterpillar into a butterfly. Caterpillars crawl, eat leaves and do not mate. Butterflies fly, seek out nectar, and mate with other butterflies. In the amphibian world, the small swimming tadpole becomes a large jumping frog. Such examples can seem exotic. But in fact, metamorphic insects alone account for at least 40% of the world's total animal populations. As an evolutionary strategy, metamorphosis works - it is not a rare or exceptional solution to the problem of adaptive success. Moreover, even non-metamorphic animals such as cats, dogs, and humans, exhibit striking differences between mature and immature forms, with the very young looking and behaving quite differently to the older forms. While at the extreme end

14 The brief sketch that follows is based on (Jabr 2012), and the Encyclopædia Britannica entry at https://www.britannica.com/science/metamorphosis.

of the spectrum lie the so-called hypermetamorphic insects that exhibit a whole series of dramatic changes across the lifespan.

Metamorphosis poses an interesting puzzle[15] for Hohwy's picture of the links between persisting agents, free energy, and the 'self-evidencing model'. The phase-transitions that occur as part of the developmental trajectories characteristic of metamorphic animals are dramatic enough to count as failures of the earlier stage life-form to continue to harvest evidence for its own existence. Yet it seems wrong to think that the transitions constitute breakdowns or failures in the organism's war against the second law of thermodynamics. The caterpillar does not lose the war when it transforms into a butterfly. On the contrary, the act of transformation is itself an essential part of the on-going project of exchanging entropy with the environment so as to persist in the face of the second law. For example, it is conjectured that metamorphosis has adaptive value because it allows younger and older forms to share the same territory without consuming the same resources or being exposed to the same predators.

A natural response to this prima facie puzzle is to point to the genetically determined nature of the phase-transitions (hence the succession of differently blanketed organizational forms) themselves. The genes that control metamorphosis are reasonably well-understood, and can be selectively blocked so that (for example) the caterpillar never turns into a pupa and hence never undergoes the caterpillar-to-butterfly phase transition – see (Bayer et al. 2003). It is reasonable, then, to think of the overall life-cycle as an evolved, self-evidencing, free-energy minimizing strategy. The life-cycle is self-evidencing insofar as the very existence of the linked stages (caterpillar, pupa, butterfly) provides evidence for the 'model' that is the metamorphic agent, where that agent is not identified with a specific morphology (which would correspond merely to one stage of the lifecycle) but with the temporally extended whole[16].

This, I submit, is the correct way to think about metamorphic beings in the free energy framework. It is also a revealing platform from which to re-consider Hohwy's worries concerning embodied cognition and the extended mind. Thus consider the wide range of bodily and sensory augmentations that already characterize many human lifespans. These include the use of spectacles to improve vision, and the wearing of clothes to help tolerate heat and cold. For some of us, they include the use of pacemakers, cochlear implants, and prosthetic limbs. In the near-future they may include the use of techniques such as refractive lens exchange to deliver not just restored but augmented vision – for example, by providing infra-red (IR) sensitivity. If my bio-typical lens is replaced with an artificial one that augments my visual repertoire, and its outputs properly integrated in downstream neural processing, it would seem strangely unmotivated to insist that 'my' true evidentiary boundaries remain those of the bare (IR-insensitive) biological system. Closer to home, we can easily imagine IR sensory evidence being made available via a mediating wearable technology such as Google Glass.

To be sure, we could identify a Markov blanket wherever the new technologies interface with the old bio-systems. In the case of ordinary spectacles or the IR-enabling Google-glass, this would line up with a known long-term boundary of interest. In the case of the lens replacement, perhaps only surgeons would consider the interface (between the new lens and the post-retinal wiring) a boundary of interest. But the lesson is that complex living beings are composed of layer upon layer of Markov blankets, reaching at least all the way down to cellular organelles and macro-molecules like DNA (deoxyribonucleic acid). Different explanatory purposes drive us to highlight some of these blankets (of blankets) at the expense of others. But no blanket or set of blankets is privileged in and of itself. Nor does the temporality of any specific Markov blanket organization seem intrinsically privileged. In the

15 I was led to consider the intriguing case of the metamorphic insects by (Friston and Stephan 2007, p. 435), who note the prima facie puzzle posed by their dramatic phase-transitions, that nonetheless occur within well-defined within developmental trajectories.

16 This again raises interesting question concerning ergodicity – here, the tendency of the blanketed systems to visit and re-visit the same sets of states over time. For we now confront a time-series in which there are profound changes in the states that are re-visited during different 'chunks' of the life-cycle. This is reminiscent, in the context of work on the 'extended mind', of the arrival of a new technology whose operations become so deeply integrated as to be called upon again and again in (some part of the) temporally-subsequent lifespan.

case of the metamorphic insects there is dramatic change across time, so that no stable Markov blanket organization characterizes the biological agent across the life-span. Instead, what seems to have been selected is an ongoing process that delivers different organizational forms (different blankets of blankets) at different times. As one of these forms gives way to another, there is some kind of preservation of systemic integrity – the biological being thus resists dissipation and death. But it would be a mistake to look for a single form or blanket of blankets that persists throughout the lifespan. Instead, we should see the shifting forms (and the shifting mosaics of Markov blankets) as themselves the means – the process – by which long-term surprisal is minimized.

The extended mind claim is that, considered as cognitive agents, human beings are 'mentally metamorphic' – as we move through life, the sets of tools, strategies, and devices (neural, bodily, and bio-external) that constitute us as the mindful beings we are undergoes dramatic alteration. Very young human life-forms do not count, as part of their cognitive apparatus, the use of pens, papers, and notebooks. But that can change over time, as the capacities they make available become more and more deeply integrated with bio-native cognitive operations. As we move through life, different bits of the encountered and humanly designed world become repeatedly and deeply incorporated into our individual cognitive routines, persisting or decaying according to need, use and the vagaries of our enabling socio-technological cocoon. Importantly, the predictive processing architecture itself provides a powerful mechanism enabling the flexible, repeated integration of capacities and operations made available by the use of reliable bio-external resources (see Clark in press, sect. 8, and Clark 2016, ch. 8). I won't rehearse those considerations here, since it is in any case evident – merely from observing our human capacity to become fluent users of a lifetime succession of new tools and technologies – that such fluidity is a crucial part of our heritage[17].

## 6 Blanket-Weaving Blankets of Blankets

Can we reconcile Hohwy's vision of agents as self-evidencing, free-energy-minimizing systems with the EEE emphasis on the multiplicity and malleability of the Markov blanket organizations themselves? The best way to do so, it seems to me, would be to embrace a process (rather than a state-based) ontology, perhaps of the form described and defended in (Dupré 2012, Dupré 2014)[18]. State-based ontologies, Dupré notes, view entities as things and often ask what changes those 'things' can and cannot tolerate. Process–based ontologies for biology, by contrast, take change as given and focus on the way some processes of change (in the case of living systems) constitute a powerful means of temporarily resisting the second law.

Dupré notes that the life-cycles of many organisms – including humans – includes multiple very different forms, and asks (Dupré 2014, p. 33) "why assume there must be anything common to every stage beyond their participation in a continuous process?" The metamorphic insects (section 5 above) raise this issue in an especially dramatic way. But it applies to every animal that undergoes developmental change. From such a perspective "a cat is a pathway from zygote to kitten to mature animal to death" (Dupré 2014, p. 33). As this pathway unfolds, nothing need be common to all the temporal parts except for their participation in the process. Nonetheless there seems no reason to think that a temporally extended process cannot itself be a free-energy minimizing system that might even be identified with some self-evidencing agent of interest. For it is the process-based succession of processes that must, ultimately, account for the temporary resistance of living organisms to the second law.

17 But notice also that the fluidity in question is itself plausibly at least partially determined by cultural innovations such as reading and writing, since these enable the easy transmission of the staged training regimes required to master new tools and techniques. See (Heyes 2012).

18 See also (Dupré and O'Malley 2009), and for a general introduction, (Seibt 2016).

At this point, I suspect that Hohwy will wish to emphasize the difference between an organism with a genetically (or at any rate, epigenetically[19]) determined pathway from one form to the next, and the case of a human agent who comes across a new tool or technology such as a smartphone and slowly incorporates it into their cognitive routines. Should we say that the caterpillar-butterfly process counts as a unity that resists entropy and maximizes evidence for itself over the long run, whereas the human-smartphone unity is too fleeting and too arbitrary to count? Such a response is perhaps suggested by the comment that:

> Whereas prediction error can be minimized transiently by systems with all sorts of objects included (*e.g.* shooting the tiger with a gun), on average and over the long run, it is most likely that the model providing evidence for itself is just the traditional, un-extended biological organism. (Hohwy 2016, p. 270)

But notice, first, that we do not have to settle for a single self-evidencing model. Within the human agent, there will be many self-evidencing 'models' including single cells and, as Dupré (Dupré 2014) nicely notes, every member of the successive colonies of bacteria that inhabit the human gut and help us digest our food. Should we nonetheless assume that where there is a persisting human person there is a unique, unchanging self-evidencing system bounded by a unique, unchanging Markov blanket? There is no reason to think so. The sets of unfolding processes that constitute the infant human are not the same as the sets of processes that constitute the adult form, and the infant brain itself is very different from the adult. So even once we fix on some Markov blanket organization of interest, that organization will itself be realized as a process undergoing constant change. Any Markov blanket bounded organizational form can be changed, re-deployed, or re-configured through interaction with the inner and outer environment.

Most importantly of all, we need to recognize learning itself as a key transformative process. It is a process (or set of processes) that forms part of both our biological and cultural heritage, and that – I would argue – delivers a succession of differently constituted mental and sensory organizations as part of its normal operation. That same process (or set of processes) can, we saw, re-configure effective systemic boundaries in many ways. The person equipped with the IR-lens will soon learn to use that information fluidly for the control of perceptuo-motor routines, as will the person who uses an add-on device such as Google Glass. In each case, the effect of learning is to generate a 'motor-informational weave' (Clark in press) that alters the effective bounds of sensing and action. When the well-woven equipment (the notebook, a software package) plays a role in the storage and transformation of information, we may speak of 'extended cognizing' rather than merely 'extended sensing' – but in each case, the key features are the same. The fact that we can still, if we wish, define a set of boundaries (a Markov blanket) that falls within these larger wholes is unsurprising. For even within the basic biological whole, there are many other boundaries we could choose. And whatever set of boundaries we happen to focus upon, they will themselves be subject to change in virtue of their realization by temporally-evolving processes rather than stable states.

Creatures like us, I conclude, are Nature's experts at knitting their own Markov blankets. Courtesy of biology, culture and learning we are 'natural-born Cyborgs' (Clark 2003) – self-organizing processes that constantly re-invent themselves, repeatedly re-defining their own cognitive, bodily, and sensory forms.

19 For example, (Dupré 2014, p. 34) notes that "the growing meristem of a plant is typically an opportunistic growth process capable of producing a variety of structures – leaves, flowers, roots – in response to the environment it encounters."

## 7 Demonic Couplings

I have tried to show that the Markov blanket organizations most relevant to the study of mind and life are multiple and malleable. But multiple, malleable organizations are organizations nonetheless. That means that for a given explanatory interest, at a given moment in time (a window within the evolving process), there will indeed be a Markov blanket on one side of which lie the interacting worldly causes that are (in Hohwy's very broad sense) 'modelled' by the temporarily persisting system. Suppose in addition that the system in question is one that constructs some kind of conscious experience of its world. This awareness, Hohwy argues, must involve a form of inference conditioned upon the sensory evidence, concerning the world beyond the blanket. This, opens the door to scepticism, since as long as the energetic exchanges across the sensory boundaries are held constant the agent cannot know whether the world beyond is one way or another. Given her priors and the sensory evidence, she will construct the same world regardless. This is the 'evil demon' (or Matrix) scenario pursued in (Hohwy 2016) and revealingly refined in (Hohwy 2017). Is there something about this sceptical scenario that puts pressure on EEE cognition?

In (Clark in press) I argue that there is not. I claim that what EEE cognition rejects is not scepticism so much as the 'richly re-constructive' vision of mind. A richly reconstructive vision depicts the mind as an inner arena populated by representational forms sufficiently rich and stable to enable us to do all the 'real' cognitive work in an effectively 'offline' manner, relegating sensing and action to simply inputting a problem specification and outputting a solution - one that action merely implements. The alternative is to recognize that action, as well as gross bodily form and reliable environmental features, can form part of the solution itself – as we see in many well-studied examples (see Clark 2008) ranging from the use of fingers in counting, to gestures while speaking, to yellow sticky-notes posted on the wall. The core EEE insight, I believe, is that circular (perceptuo-motor) causality in a richly structured and often repeatedly self-structured environment is itself the core adaptive strategy, and that the neural contribution is best seen as just one part of this wider body-and-world-involving mosaic.

As far as I can tell, Hohwy (Hohwy 2017) accepts that broad vision. But he remains concerned to stress the apparent insulation of conscious experience behind a veil of sensing and acting. If all that this means is that, in principle, as long as my priors and the flow of energies across my sensory surfaces remain fixed, I cannot know whether or not I am a Matrix-world 'brain-in-a-highly-intelligent-vat', I am happy to agree. Notice that being thus en-vatted would in no way alter the fact that the problem-solving strategies I deploy make great use of extra-neural opportunities. The vat-being, thus construed, can robustly rely on the use of her fingers for counting, and she can post yellow sticky notes aplenty on the surfaces of her sensed office. Mildly cognitively impaired Otto, in vat-world, likewise makes use of a trusty notebook that allows him to behave in many of the ways characteristic of having a multitude of standing beliefs that are not enshrined in his neural apparatus in the bio-typical manner. As we move deeper and deeper into the 21st century, vat-world too may be increasingly home to enhanced and augmented agents, who have extended their bio-typical sensory boundaries using new devices such as IR-sensitive replacement lenses. All this is true, it seems to me, regardless of the fact that, in vat-world, these cognitive extensions and boundary-changing alterations are implemented in vat-controlled software rather than whatever hardware ultimately - or so we presume - comprises the physical universe we inhabit

In a revealing twist on his original (2016) demon-description, Hohwy (Hohwy 2017) further shows that to the extent that all this is true, the prediction-error minimizing agent is in fact manipulating the demon (the intelligent vat) by giving action commands whose fulfilment requires the demon to compliantly alter the apparent flows of energy across her (apparent) sensory surfaces.

What shall we make of this? In Hohwy's well-chosen words:

> [N]o evidence available to the agent distinguishes between the hypothesis that they really have an arm and the hypothesis that an evil demon is deceiving them to think they have an arm. Note that the internal states of the agent are part of a dynamic causal chain that modulates the states of the hidden causes. This holds whether the external states harbor real arms and other familiar things, or a demon. We are assuming the world contains either of these, and that they causally impact the agent's sensory states. In the demon world, this means the demon's states causally entrain the agent's internal states. But equally, through active inference, the agent's states causally entrain the demon's states. If the demon is not entrained like this, then the agent's prediction errors would not be minimized in active inference. (Hohwy 2017, sect. 6)

This is a lovely observation. The inevitable entrainment of the demon/vat by the agent shows (I would argue) that the shape and structure of the experienced world – the world as presented in experience – is nothing over and above the shape and structure of an open-ended set of possible encounters, shaped by human needs and the affordances of body and world. Whether all that is ultimately implemented in the standard or some non-standard (e.g. envatted) way is unimportant. This is just another way of showing that the demon thought experiment puts no pressure on the core claims of EEE cognition, at least as I would construct them. Within vat-world, all the interesting and unexpected patterns of reliance upon bodily and extra-neural sources of stability and structure celebrated by EEE cognition remain in force. More radically still, Chalmers (Chalmers 2005) and Clark (Clark 2005) argue that that the very reality of arms, legs, notebooks, and sticky-notes is unaffected by these variant implementations, since 'realness' should never have been predicated upon the correctness of some specific deep physics in the first place. If we go this route, the premise of the thought experiment is itself undermined – 'all' the demon has done is altered some implementation details while preserving our perfectly real world. But I mention this only in passing since (as we just saw) it is not necessary to endorse this more radical view in order to move beyond the image of neurocentric seclusion, even for the envatted brain.

Finally – more for fun than as an attempt at further argument – suppose we step outside the realm of constructed experience and ask how the vat-agent exchanges entropy with the environment so as to persist (as some kind of evolving process) in the face of the second law? The answer, from the agent's perspective, implicates multiple perception-action loops that exploit extra-neural resources of varying kinds. But those extra-neural resources are now held steady only – I presume – by the expenditure of a great deal of demonic energy. If we deny this, then the truth of the demon scenario would simultaneously imply the falsity of the free-energy principle itself. Assuming this is not the intention, we must imagine that the demon is busily exchanging entropy with some larger demon-world, so as to persist long enough to be suitably entrained by the vat-agent. Such a demon would herself be well-placed to conform to the core principles of EEE cognition, increasing the efficiency of her own demonic stratagems by making maximal use of sources of order and structure beyond the bounds of the demon-brain itself.

## 8 Conclusions: From States to Processes

I have tried to show that EEE approaches, even those that embrace 'extended minds', do not thereby seek to obliterate Markov blankets from their depictions of mind and agency. To do so would, as Hohwy and Friston correctly stress, disable the controlled exchange of entropy and expose the agent or cognitive system to immediate fatal dissipation – bad news even for an extended mind. In (Clark in press) I argue that EEE approaches are best seen as rejecting 'richly reconstructive' visions of mind. But that is really only half the story. The other half of the story (pursued today) concerns the way

embodied minds and embodied agents repeatedly re-configure their own boundaries in ways that promote adaptive success. Such repeated reconfiguration does not always result in the destruction of the older blanketed organizations, which may simply become further nested. But it may sometimes – as in the case of the metamorphic insects – do so. Either way, it results in the creation of new systems of explanatory interest. Such systems remain locally ergodic. They visit and re-visit states that their predecessors did not, while remaining stages of a single life-cycle.

Dramatic change in the metamorphic insects is genetically controlled, and thus counts as 'expected' in some very broad evolutionary sense. But many biological agents are, in the same broad sense, 'expected' to engage in lifetime learning, altering their effective environments and (in our case) creating tools and technologies that may shift the sensory boundaries themselves. That same process can usher into being new individualistic organizational forms, such as the bio-being+smartphone, or new collective forms, such as a nation-state. Learning and neural plasticity thus open the doors for new or successor blanketed systems consequent upon cultural and technological (including bio-engineered) innovations.

The perspective I am recommending may seem challenging insofar as it invites us to contemplate agents whose sensory boundaries are not fixed and whose cognitive architectures may extend, in temporally varying ways, beyond the biological brain. But perhaps this should not surprise us unduly. For there already exists a potent inner equivalent in the 'predictive processing' account of attention and variable precision-weighting itself. Attentional mechanisms, that story suggests, alter patterns of inner (neural) effective connectivity so as to enforce information flows that are highly specialized for the task at hand (Clark 2016, Clark in press). Attention, if this is correct, itself imposes a kind of transient organizational form, with its own distinctive Markov blanket organization (marked by temporary conditional statistical independencies), upon the brain. Attentional mechanisms may thus be seen as driving the formation and dissolution of a short-lived Markov partitioning within the neural economy itself[20], temporarily insulating some aspects of on-board processing from others according to the changing demands of task and context. These transient neuronal ensembles then recruit (and may also be recruited by) shifting coalitions of bodily and worldly elements, resulting in the repeated construction of temporary task-specific devices that span brain, body, and world. In this way, the ebb and flow of neural influence is matched by an ebb and flow of bodily and worldly influence. It is the progressive generation and maintenance of these nested transient partitionings, swept up in the circular causal dynamics that bind perception and action, that enables living beings to persist and minimize free energy across their lifespan.

Hohwy's insightful probing has thus revealed something deeply important. It has forced us to recognize that the picture of biological agents as free-energy-minimizing systems requires something closer to a process-based (rather than a static or state-based) ontology. If the free-energy minimizing system is really a free-energy minimizing process, much that is otherwise puzzling falls into place. For processes are by their very nature on-going, and can repeatedly generate new forms, boundaries, and constituent structures as they continue to exchange entropy with whatever counts (at a given stage) as the wider world.

20 A more familiar case may be the wake-sleep cycle itself, which regularly creates a new set of partitions (a new transient Markov blanket) between sensory systems and the world. This possibility is noted by Karl Friston in comments reported on the Frith blog at: http://frithmind.org/social-minds/2014/05/12/under-the-markov-blanket/

## References


Adams, F. & Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology, 14* (1), 43–64.

Bayer, C., Zhou, X., Zhou, B., Riddiford, L. M. & von Kalm, L. (2003). Evolution of the Drosophila broad locus: The Manduca sexta broad Z4 isoform has biological activity in Drosophila. *Development Genes and Evolution, 213* (10), 471–476.

Chalmers, D. (2005). The Matrix as metaphysics. In C. Grau (Ed.) *Philosophers explore The Matrix*. New York: Oxford University Press.

Chiel, H. J. & Beer, R. D. (1997). The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neurosciences, 20* (12), 553–557.

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

Clark, A. (2003). *Natural born cyborgs: Minds, technologies, and the future of human intelligence.* New York: Oxford University Press.

Clark, A. (2005). Intrinsic content, active memory and the extended mind. *Analysis, 65* (285), 1–11. https://dx.doi.org/10.1111/j.1467-8284.2005.00514.x.

——— (2007). Curing cognitive hiccups: A defense of the extended mind. *The Journal of Philosophy, 104* (4), 163–192.

——— (2008). *Supersizing the mind: Embodiment, action, and cognitive extension.* New York: Oxford University Press.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

——— (in press). Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs*. http://dx.doi.org/10.1111/nous.12140.

Clark, A. & Chalmers, D. (1998). The extended mind. *Analysis, 58* (1), 7–19.

Dawkins, R. (1982). *The extended phenotype.* Oxford: Oxford University Press.

Dupré, J. (2012). *Processes of life: Essays in the philosophy of biology.* New York: Oxford University Press.

——— (2014). A process ontology for biology. *Physiology News, 100*, 33–34.

Dupré, J. & O'Malley, M. A. (2009). Varieties of living things: Life at the intersection of lineage and metabolism. *Philosophy & Theory in Biology, 1*, e003.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11* (2), 127–138. https://dx.doi.org/10.1038/nrn2787.

——— (2013). Life as we know it. *Journal of The Royal Society Interface, 10* (86). https://dx.doi.org/10.1098/rsif.2013.0475.

Friston, K. & Stephan, K. (2007). Free-energy and the brain. *Synthese, 159* (3), 417–458.

Haugeland, J. (1998). Having thought: Essays in the metaphysics of mind. In J. Haugeland (Eds.) *Mind embodied and embedded* (pp. 207–240). Cambridge, MA: Harvard University Press.

Havas, D. A., Glenberg, A. M., Gutowski, K. A., Lucarelli, M. J. & Davidson, R. J. (2010). Cosmetic use of botulinum toxin-A affects processing of emotional language. *Psychological Science, 21* (7), 895–900.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science.* New York: Free Press.

Heyes, C. (2012). Grist and mills: On the cultural origins of cultural learning. *Philosophical Transactions of the Royal Society B, Biological Sciences, 367* (1599), 2181–2191. https://dx.doi.org/10.1098/rstb.2012.0120.

Hohwy, J. (2016). The self-evidencing brain. *Noûs, 50* (2), 259–285. https://dx.doi.org/10.1111/nous.12062.

——— (2017). How to entrain your evil demon. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Jabr, F. (2012). How did insect metamorphosis evolve? *Scientific American Online*. https://www.scientificamerican.com/article/insect-metamorphosis-evolution/.

Lipton, P. (2001). What good is an explanation? In G. Hon & S. S. Rakover (Eds.) *Explanation: Theoretical approaches and applications* (pp. 43–59). Dordrecht: Springer Netherlands.

Menary, R. (Ed.) (2010). *The extended mind.* Cambridge, MA: MIT Press.

Mohan, V., Morasso, P., Sandini, G. & Kasderidis, S. (2013). Inference through embodied simulation in cognitive robots. *Cognitive Computation, 5* (3), 355–382.

Neal, D. T. & Chartrand, T. L. (2011). Embodied emotion perception: Amplifying and dampening facial feedback modulates emotion perception accuracy. *Social Psychological and Personality Science, 2* (6), 673–678.

Newen, A., de Bruin, L. & Gallagher, S. (in press). *Oxford handbook of 4E cognition.* New York: Oxford University Press.

Norris, J. R. (1998). *Markov chains.* Cambridge: Cambridge University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Francisco: Morgan Kaufmann.

Pezzulo, G. (2014). Why do you fear the bogeyman? An embodied predictive coding model of perceptual inference. *Cognitive, Affective, & Behavioral Neuroscience, 14* (3), 902–911.

Pfeifer, R. & Bongard, J. (2006). *How the body shapes the way we think: A new view of intelligence.* Cambridge, MA: MIT Press.

Rupert, R. D. (2009). *Cognitive systems and the extended mind.* New York: Oxford University Press.

Seibt, J. (2016). Process philosophy. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy* Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/process-philosophy/.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17* (11), 565–573.

Solomon, E. P., Berg, L. R. & Martin, D. W. (2002). *Biology. Sixth edition.* Stamford, CT: Thompson Learning.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind.* Cambridge, MA: Harvard University Press.

Tribus, M. (1961). *Thermostatics and thermodynamics: An introduction to energy, information and states of matter, with engineering applications.* Van Nostrand.

Turner, J. S. (2009). *The extended organism: The physiology of animal-built structures.* Cambridge, MA: Harvard University Press.

Varela, F. G., Maturana, H. R. & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems, 5* (4), 187–196.

Wiener, N. (1961). *Cybernetics: Or control and communication in the animal and the machine.* Cambridge, MA: MIT Press.

# Of Bayes and Bullets:

## An Embodied, Situated, Targeting-Based Account of Predictive Processing

Michael L. Anderson

Here I argue that Jakob Hohwy's (Hohwy 2013) cognitivist interpretation of predictive processing (a) does not necessarily follow from the evidence for the importance of Bayesian processing in the brain; (b) is rooted in a misunderstanding of our epistemic position in the world; and (c) is undesirable in that it leads to epistemic internalism or idealism. My claim is that the internalist/idealist conclusions do not follow from predictive processing itself, but instead from the model of perception Hohwy's adopts, and that there are alternate models of perception that do not lend themselves to idealist conclusions. The position I advocate is similar to Andy Clark's embodied/embedded interpretation of Bayesian processing (Clark 2015); however, I argue that Clark's position, as currently stated, also potentially leads to idealist conclusions. I offer a specific emendation to Clark's view that I believe avoids this pitfall.

**Keywords**

Bayesian brain | Ecological psychology | Embodied cognition

### 1 Mechanical Applications of Bayes' Theorem

As it is generally presented, Bayes' theorem is a way of updating confidence in a given belief or hypothesis given accumulating evidence. In its modern mathematical form, it looks like this:

$$P(B|E) = [P(E|B)P_{prior}(B)] / [\Sigma P(E|B')P_{prior}(B')]$$

The probability $P$ of belief $B$ given evidence $E$ is equal to the probability of $E$ given $B$, times the previous estimate of the probability of $B$, divided by the sum of the probabilities of the evidence given all possible $B$s. Note that this formulation is fundamentally *representational*: it posits contentful beliefs about things in the world that are deemed more or less likely to be true in light of ongoing observations (Wiese and Metzinger 2017).

It is probably the case that this is the right interpretation to give those symbols when the equation is being used by human beings to determine things like the probability of a terrorist attack given intercepted phone calls, or a medical diagnosis in light of test results. But I think there is an important question to be asked about what interpretation to provide in cases where this mathematical relationship is being implemented by impersonal (or sub-personal) *mechanisms*. One of the earliest practical successes of Bayes' Theorem was in artillery targeting, and a search of the military research literature suggests it remains in use today. The point for now is simple: automated targeting computers don't believe anything, at least not yet, and still Bayes' Theorem is effective at setting the parameters of the mechanism—getting the machine into the right configuration—to hit its target.[1]

To signal my intentions at the outset, I believe that words like "targeting", "(re)configuring" and "guiding" should serve as replacements for the prevailing metaphors of "belief", "inference" and "hypothesis" at play in the majority of the predictive processing literature. Yes, we scientists can use Bayes'

1 The details of the targeting mechanism are, it seems, classified. But roughly speaking the procedure involves feeding observations of munition landing positions (degree and direction of deviation from the target, along with several other variables) into a fire direction computer with a database of targeting settings/adjustments that were pre-computed using Bayesian methods. The output of the computer is relayed to the gunners who change settings including barrel angle and orientation accordingly. Note this *could* be done in a fully automated way, but for cultural reasons, the U.S. military prefers "human-in-the-loop" systems. See U.S. Army field manual 3-09.22 for the general outlines. Theoretical background can be found in Kolmogorov 1942, and Kolmogorov and Hewitt 1948. An interesting blog post on using Bayes to solve the related problem of artillery fire safety can be found here. https://linesoftangency.wordpress.com/2012/01/11/check-one-two/

Anderson, M. L. (2017). Of Bayes and Bullets: An Embodied, Situated, Targeting-Based Account of Predictive Processing.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 4. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573055

rule to update our beliefs and hypotheses; it can also be used to appropriately change the parameters in a control system. I think it is an open question what Bayesian processing might be up to in our brains. It could be that it is building sub-personal representations including hypotheses, models and beliefs about a mind-independent world, and is thus an important part of the process of *perceptual reconstruction* of the shape of that world. But a more minimalist, mechanical interpretation of Bayesian updating is available to us, and we should take seriously the possibility that it offers a truer account of the Bayesian brain.

In what follows, I will argue that Hohwy's (Hohwy 2013) cognitivist interpretation of predictive processing, although consistent with the views of its most prominent proponents, (a) does not necessarily follow from the evidence for Bayesian processing; (b) is rooted in a misunderstanding of our epistemic position in the world; and (c) is undesirable in that it leads to epistemic internalism or idealism.

I will then turn to Clark's (Clark 2016) reading of the same literature, which purports to avoid the idealist pitfalls of Hohwy's account by situating predictive processing in the action-oriented framework offered by the embodied cognition movement (Anderson 2003; Anderson 2014; Clark 1997). Although I agree entirely with Clark's final analysis, I will suggest that in trying to combine aspects of both the cognitivist and the embodied accounts of predictive processing, Clark has generated a dilemma of perceptual content that ultimately lands him back in Hohwy's foundering epistemic boat.

I send Susan Oyama (Oyama 2000/1985) to the rescue. I argue that—at least at critical moments in his exposition—Clark (and Hohwy, too, although I won't make this case in detail) appears to be in the grip of an untenable, ultimately dualist account of the nature and origins of information, and this underlies his inability to escape epistemic internalism. I suggest that a more thoroughgoing dynamic systems account than the one Clark defends, along with a monist account of information according to which content emerges only in organism-environment interactions, offers both a way out of the internalist hole, and also a truer picture of the role of Bayesian updating in our cognitive economy. According to this picture, perceptual content is action-oriented, rooted in the fundamental intentionality of action[2] (Anderson and Chemero 2009; Anderson and Rosenberg 2008; Clark 2015), and Bayesian updating helps set the parameters that allow our dynamic, transiently assembled local neural sub-systems (TALoNS; Anderson 2014) to appropriately target adaptive behavior.

## 2 The "Problem" of Perception

It is commonly thought that the senses are inadequate for perception. It sounds odd, I hope, to put it that way, but this is indeed the majority position. For those who hold it, there's no point puzzling about how some three billion years of evolution could have failed to furnish us with better systems, because it is not just true but *obvious* that such is the case. After all, the world is 3-dimensional, and the retinal image not just flat, but inverted. The world varies along uncountable dimensions, and our senses respond to only a small fraction of them. On this view, the information delivered by our senses is at best an *impoverished reflection* of the world out there, and reconstructing that world from such materials takes both intelligence and wisdom. It takes intelligence, for instance, in the form of clever algorithms for extracting shape and texture from luminance cues, and wisdom in the form of knowledge that illumination generally comes from above. The brain that manages to solve the problem of perception does so just in so far as it is able to both *manipulate* and *add to* the information delivered by the senses.

2 Classically, intentional states are directed at things in virtue of a subject's perception of those things, and the intentional relation is grounded in the causal relation that gave rise to the perception (e.g. Devitt 1981; Grice and White 1961; Kripke 1980). On such views, the intentionaliy of action is derivative from the intentionality of perception and/or the intentional states guiding the action. In my view, this gets things exactly backward. Part of what it is to be an action is to be directed toward objects and ends; this intentionality is not derived from intentional mental states, but is inherent to behavior itself. Intentional mental states are directed at things *because* the actions they guide are, and in this sense the intentionality of mental states is derivative from the basic intentionality of action. Defending this view would of course require its own paper.

On the specifically Bayesian version of how the brain solves this problem of reconstruction, the trick is managed by using the deliverances of the sense organs as the basis for *inferences* about the ultimate causes of those stimulations. Beliefs about the world (explicit as well as tacit) both result from and constrain these inferences. As Hohwy expresses the point:

> The problem of perception is a *problem* because it is not easy to reason from only the known effects back to their hidden causes. This is because the same cause can give rise to very different effects on our sense organs. […] Likewise, different causes can give rise to very similar effects. (Hohwy 2013, p. 13)

This is a classic poverty-of-the-stimulus argument. Because the senses are incapable of specifying a unique situation, the reverse-inference problem posed by perception is too unconstrained to solve with sensory resources alone. Hohwy again: "[i]f the only constraint on the brain's causal inference is the immediate sensory input, then, from the point of view of the brain, any causal inference is as good as any other." (Hohwy 2013, p. 14) But of course, immediate sensory input is *not* the only constraint; there are, in addition, general beliefs about the world, specific hypotheses about the current state of the world, and *ongoing* sensory input. All this together, combined in the right way, is what on this view is needed to solve the problem of perception.

One other feature of the problem of perception, an unnecessary but oft included part of the story, takes the form of an additional metaphysical posit couched as a simple corollary. As Helmholtz put it: "[w]e always in fact only have direct access to the events at the nerves, that is, we sense effects, never the external objects." (1867, p. 430; quoted in Hohwy 2013, p. 17) What this means is that in solving the problem of perception, in using the senses to answer the question of what's out there, we are not permitted to "help ourselves to the answer by going beyond the perspective of the skull-bound brain." (Hohwy 2013, p. 15) Thus, one of the things we need inference for is to "overcome the brain's encapsulation in the skull" (Hohwy 2013, p. 41) and give us knowledge of what we cannot directly access.

One immediate point, worth making for our purposes here, is that predictive processing is only one proposed solution to what is supposed to be a very general problem. Very few of the apparent consequences for perception that follow from Hohwy's treatment—that it is knowledge-rich, representational, and indirect, for instance—result from features specific to the predictive processing solution. Rather, these are consequences that would follow from *any* adequate solution to the problem as posed.

What is perhaps even more striking are the metaphors of containment that abound—that are apparently irresistible—when treating perception in this manner. We are "trapped inside our skulls" (Hohwy 2013, p. 224), "skull-bound" (pp. 15; 75), screened off from the world by a "veil of sensory input" (p. 50). Given this predicament, it's the best we can do to infer the "hidden" (pp. 50; 81) causes; the only alternative would be to "impossibly jump outside the skull and compare percepts to the causal states of affairs in the real world." (p. 50)

It's worth asking exactly *who* is thus trapped? Who is the "we" inside the skull with "direct access [only] to the events at the nerves" and longing to "jump outside"? The answer has to be the knowing subject, the *person*, and although Hohwy is at pains to deny any hint of homuncularism, it's hard not to notice the little man peeking out from behind that veil. It is no defense to point out—however correctly—that these are just metaphors. For as Lakoff and Johnson have been telling us for some time (e.g. Lakoff and Johnson 1980; Lakoff and Johnson 1999) metaphors *matter*. They change how we think, what we attend to, what questions we ask, and what answers we will accept.

The metaphors are certainly doing their work here. Hohwy's initial statement of the problem of perception asks us to imagine being alone in a dark, sealed house and trying to determine what is making that tapping noise outside. It's an effective image, in that it's hard to conceive of resources *other* than reason that could be brought to bear in such restricted circumstances. It's also wildly misleading. Sensory systems are nothing like that, passive, isolated, and unimodal (not to mention objective and reconstructive (Akins 1996; Anderson 2014)). Hohwy later admits the inadequacy of the metaphor

(p. 77) but by then the damage has been done. Although the house gets windows and legs, the walls remain intact, the doors are locked, and the mind (in this metaphor identical to the person) is trapped inside.

I confess, I don't recognize myself in this image. For a start, I'm pretty sure *I* encompass my skull; it is inside me, and not the other way around. It's a minor point, perhaps, but telling, for it forces us to question this notion of "access". Given that I am in the world (and not in my skull) in what sense am I cut off from that world? Do I not have access to air? To food? To sunlight and the solid surfaces that support me? If I have physical access to these things, what prevents having perceptual access?[3] What is meant by the claim we have direct perceptual access only to events at the nerves? It's certainly the case that nerves lie along some important causal paths that run between organisms and the world. But what licenses the claim that our perceptual access to the world ends where the nerves do?

An obvious candidate answer would be that I have access only to the last link in the causal chain; the links prior are increasingly distal. But I do not believe that identifying our access with the cause most proximal to the brain can be made to work, here, because I don't see a way to avoid the path that leads to our access being restricted to the chemicals at the nearest synapse, or the ions at the last gate. There is always a cause even "closer" to the brain than the world next to the retina or fingertip. Moreover, causes aren't really linear chains, but complex webs of conditions. The causes of the shattering glass include the tension in my muscles, the energy they put into the descending hammer, the rigidity of the hammer's head, the brittleness of the glass, the firmness of the supporting surface, and more (see, e.g. Mackie 1965). In light of such complexity, it seems to me that identifying "the" causal link we access in perception is likely a fool's errand to start.

We are dangerously close to confusing epistemic and causal mediators here. Hohwy doesn't make this mistake, but he doesn't avoid it, either, for he slides between the *causal* claim that we have access only to events at the nerves (e.g. pp. 17; 50) and the *epistemic* claim that we have access only to sense data (e.g. p. 75). This slide introduces psychological and informational content where before there was only cause. It's not clear what warrants this, and, if it is warranted in general, what justifies the restriction of epistemic access to "sense data". If we have access to content and information, why conceptualize that access in so narrow a way? And if that content is *about the world* it is not clear what blocks the conclusion that we thereby access the world. On the other hand, if we have access only to causes, we are still owed a story about how those causes give rise to experience, and it's the details of this story that will tell us whether and how we have epistemic access to the world; epistemic encapsulation is not something to be declared by fiat. Hohwy knows (as Clark 2016, insists) that it is only in virtue of being embedded in the causal nexus that we have perceptual access to the world at all; far from isolating us from the world, our nervous system fully embeds us in it. What we stand behind in this picture is not a causal barrier, but an epistemic veil, and the image is arresting enough that we can fail to notice that its existence flows only from the metaphor of the house, and not from the facts of our situation.

As noted above, most of the conclusions I find objectionable here stem not from predictive processing per se, but from the account of what predictive processing must do to solve the problem of perception so formulated. In the next section, I'll briefly describe an alternative statement of the "problem" of perception that avoids these issues, and opens the path to a different treatment of predictive processing. But first we need to fill out Hohwy's account. For what *is* specific to Hohwy's interpretation of predictive processing is the story about exactly *how* beliefs, hypotheses, and sensory input get combined to enable perception. Arguably, that story exacerbates, or at least highlights, the apparent indirect nature of our perceptual—and hence epistemic—access to the world.

3 A critic might counter that the assumption that I am in the world, breathing air, is itself a mere seeming, compatible with being a brain-in-a-vat (which is indeed what Hohwy is arguing we are). Here I can only say that if I am not allowed the assumption that I am physically in the actual world, then it is the skeptic who is assuming what must be shown. More pointedly: it is the skeptic's way of setting up the problem of perception that opens up the gap between mind and world that (s)he fills with barriers and doubt. But this way is not the only way of understanding the job of perception. I sketch an alternative, below, from which these conclusions do not in fact follow (see also Anderson 2014; Anderson 2006).

The predictive processing story, neutrally specified, goes like this: our brains implement a massive, complex, hierarchical Bayesian network structured such that higher levels dynamically predict the states of lower levels, all the way down to the changing states of our sensory receptors and physical actuators. On this story, what flows "down" the hierarchy are predictions of state, and what flows "up" the hierarchy are the differences between the predicted and actual values. These error signals cause adjustments to the settings of the higher level(s) so as to bring the predictions into closer alignment with the actual states of the lower levels of the hierarchy.

It is universally acknowledged that this picture of the mechanisms of perception differs significantly from the classical model, which posits a feed-forward cascade of feature detection. There is, however, significant disagreement over what this difference in mechanism implies for the nature of perception itself (Butz 2017; Clark 2015; Clark 2016; Gallagher and Bower 2014; Hohwy 2013; Seth 2014). In Hohwy's case, because he accepts the assumptions driving the classical account that the purpose of perception is to build objective models (representations) of the mind-independent world, his model of the cortical hierarchy is essentially identical to the classical model: perception results in the representation of properties of the environment at ever-higher levels of complexity and organization. As he puts the matter:

> The brain responds to … causal, hierarchical structure [in the world] in a very comprehensive manner: it recapitulates the interconnected hierarchy in a model maintained in the cortical hierarchy of the brain. (Hohwy 2013, p. 28)

I have no intention to mount a sustained argument against this position. I'll simply make two points: First, there is nothing in the nature of the mechanism itself that warrants this interpretation. What a given layer of the cortical hierarchy is doing (and what kind of representations it trucks in) is an *empirical* question. It is not inferable from knowledge that the layer is performing Bayesian updating, any more than it would be inferable from knowledge that it is engaged in Hebbian plasticity. The brain is composed of functionally differentiated parts; no general mechanisms can tell us about these differences. Moreover, Hohwy's interpretation is hardly a natural one if the basic idea is that the brain represents what it predicts; given that each layer is predicting the states of lower levels of the hierarchy, the hypothesis that sticks closest to the facts of the mechanism is that each layer represents the one below it (O'Regan and Degenaar 2014). More generally, I think there's an interesting and unexamined commitment to two things, here, both of which are highly questionable. The first is structuralism about experience: the classic thesis that conscious experience is composed of simple *elements* (such as sense data), combined into different complex structures (Titchener 1929), and the second is an expectation that there must be an isomorphism between the structure of experience and the structure of the mechanism that gives rise to it. Clark (Clark 2016) is not entirely unaffected by these tacit commitments.

Second, once one combines a classical representational account of perception with a predictive processing mechanism, I believe that, although there may be defenses that could be mounted to resist the conclusion[4], it is at least plausible that epistemic internalism naturally follows. Because the information from the world, flowing up the hierarchy, encodes only the mismatch between the prediction and reality, one's sense of that reality can only be encoded in the prediction. As Hohwy succinctly states: "Perceptual content is determined by the hypothesis that best suppresses prediction error." (Hohwy 2013, p. 117) On this model, perception is only indirectly caused by the world.

I most certainly do *not* want to recapitulate the realism vs. idealism and epistemic externalism vs. internalism debates that raged around the end of the 20th century (but see O'Donovan-Anderson 1997). I'll simply assert without argument that the realists and externalists won, and that any episte-

4 I gestured at some possible options above.

mology that implies internalism and its attendant skepticism has made a mistake *somewhere*. I believe this conclusion is reliable enough that it can be used as a *reductio ad falsum*, and that is how I will employ it here: if internalism, then not (predictive coding & classical representationalism)[5]. Because we are taking predictive coding as true by hypothesis, it is classical representationalism that must go.

In this section, I have outlined Hohwy's understanding that the task of perception is to internally reconstruct the state of the world from impoverished sensory stimuli. For him, perception is a process of reverse inference, wherein the causes of sensory stimulation are inferred from their effects on sensory receptors. Because there is no unique solution to this problem—multiple different models could, in theory, account for the effects equally well—we are in the position of generating uncertain *hypotheses* about the true state of the world. According to Hohwy, predictive processing is the main neural-psychological mechanism we have for continually updating our uncertain models in light of ongoing stimulation. Our models/hypotheses make predictions about incoming states, and are updated in light of any mis-matches between the models and the states.

Hohwy argues that one important philosophical consequence of having such a perceptual system is that we have only indirect epistemic access to the actual world; he thus embraces a version of epistemic internalism that, under at least some circumstances, can lead to various forms of skepticism. Although I take issue with some of the steps in this argument, in the end I find the conclusion both plausible and unacceptable. To resist this conclusion, then, means denying one or more of its antecedents. For me, the obvious candidate is Hohwy's characterization of the problem of perception. I argue (1) that the internalist/idealist conclusions follow only from that model of perception, and not from predictive processing per se and (2) that there are alternate models of perception that do not lend themselves to idealist conclusions. It is to one such alternative model that I turn in the next section.

## 3 What Perception Is for[6]

One key aspect of the problem of perception as sketched above is that there is insufficient information delivered by the senses to specify the world. Another is the notion that the immediate function of perception is the veridical, objective representation of the external world. These two suppositions work together to support an *inferential, reconstructive* account of perception that centrally features the maintenance of world models. The core of Hohwy's account of predictive processing is an explanation of how Bayesian updating enables the construction and maintenance of such models. But what I want to suggest here is that both of these suppositions are questionable, if not demonstrably false, and therefore we don't need such an explanation. Rather, what needs explanation is how organisms appropriately attune to their environments to support adaptive behavior.

I'm not sure I know exactly what sense data are, but if a sense datum is meant to be something like the information about hue and intensity delivered by a single pixel of a digital camera, then I can say with confidence that neither the retina nor indeed any sensory system delivers sense data. Visual input to the brain is not like the snapshot of a camera, but involves multiple, mutually informing structured flows coming from the eyes (and not just the retina!), ears, head, limbs, and the rest of the active, sensing body. The raw materials of perception are not the momentary impacts of light on the retina, or chemicals on the olfactory receptors, but rather the relationships between *changes* across multiple modalities as one's position and posture change. As Gibson (Gibson 1966) writes:

5 In fact, this formulation is too simple, for clearly there are forms of representationalism that do not imply idealism/epistemic internalism. So in fact there are at least two choices for resisting Hohwy's conclusions: (1) reanalyze the origins of perceptual content in predictive processing while denying that epistemic access is restricted to sense-data, but otherwise accepting a representationalist framework; or (2) rejecting the representationalist gloss on predictive processing. I think the former is a live possibility that deserves sustained exploration, but I will follow out the latter here, in part because I think Andy Clark can be understood as offering a proposal of the former sort, and, as I will argue below, that proposal suffers from its own epistemic issues.

6 Portions of this section are adapted from chapter 5 of Anderson 2014.

> The active observer gets invariant perceptions despite varying sensations. He perceives a constant object by vision despite changing sensations of light; he perceives objects by feel despite changing sensations of pressure; he perceives the same source of sound despite changing sensations of loudness in his ears. The hypothesis is that constant perception depends on the ability of the individual to detect the invariants, and that he ordinarily pays no attention whatever to the flux of changing sensations. (Gibson 1966, p. 3)

When we attend to the fact that perception is an activity, that part of seeing, and smelling, and feeling is *moving*, we see that the actual deliverances of perception are extremely rich, multimodal, and perfectly capable of revealing the higher-order invariants in our environment and uniquely specifying the shape of the world. More to the point, perception doesn't start with sensory stimulation, for each and every "stimulation" was itself preceded (and is generally accompanied) by an action—action and perception are constant, ongoing and intertwined. As Gibson writes, "The active senses cannot simply be the initiators of signals in nerve fibers or messages to the brain; instead they are analogous to tentacles or feelers" (Gibson 1966, p. 5).

On this alternate view, then, the problem of perception is *not* how organisms get from stimulus to model, for perception and action are deeply linked. It is because one's view changes as eyes, head, and body move around the world that it is possible to know the world. This is why Hohwy's darkened house metaphor rigs the game: the perceptual system is an exploratory and not an inferential system. In active perception, the presumed poverty of the stimulus disappears. There is no infinity of possible worlds to sort through, each equally consistent with incoming information, for the deliverances of perception are not limited to the momentary 2-dimensional image on the retina (or the surface of the skin), but consist rather in the much richer set of distinctive transformations of that input given changes in posture and position. These transformations are sufficient to reveal the shape of the world.

Indeed, as Warren (Warren 2005) points out, the ability to pick up on such environmental invariances is likely a condition of perceptual systems evolving at all:

> Perceptual systems become attuned to informational regularities in the same manner that other systems adapt to other sorts of environmental regularities (such as a food source): possessing the relevant bit of physiological plumbing (whether an enzyme or a neural circuit) to exploit a regularity confers a selective advantage upon the organism. Since the water beetle larva's prey floats on the surface of the pond and illumination regularly comes from above, possession of an eye spot and a phototropic circuit can enhance survival and reproductive success. But if illumination were ambiguous and prior knowledge were required to infer the direction of the prey, it is not clear how such a visual mechanism would get off the ground. Natural selection converges on specific information that supports efficacious action.
>
> What the [traditional] view treats as assumptions imputed to the perceiver can thus be understood as ecological constraints under which the perceptual system evolved. The perceptual system need not internally represent an assumption that natural surfaces are regularly textured, that terrestrial objects obey the law of gravitation, or that light comes from above. Rather, these are facts of nature that are responsible for the informational regularities to which perceptual systems adapt, such as texture gradients, declination angles, and illumination gradients. They need not be internally represented as assumptions because the perceptual system need not perform the inverse inferences that require them as premises. The perceptual system simply becomes attuned to information that, within its niche, reliably specifies the environmental situation and enables the organism to act effectively. (Warren 2005, pp. 357-8)

This brings us to a second supposition that underlies the traditional approach. Insofar as it is the fundamental job of the perceptual system to build and maintain a model of the world, then it becomes

natural to think the fundamental content of perception is written in a perceiver-independent physics-influenced vocabulary of edges, colors, velocities and weights (Gibson 1979). But if the purpose of perception is to guide action, then the more parsimonious hypothesis is that the organism will be primarily sensitive not to these sorts of properties but rather to action-relevant relationships between organism and environment (Anderson and Rosenberg 2008; Anderson and Chemero 2009).

> The frog's visual system, for example, is tuned to particular patterns of motion that, in the restricted context of its niche, specify small edible prey and large looming threats. The fish's lateral line organ is tuned to pressure waves that specify obstacles, the movements of predators and prey, and the positions of neighbors in the school. Even the narwhal's tusk turns out to be a sense organ tuned to salinity differentials that specify the freezing of the water's surface overhead. The narwhal is thereby in perceptual contact with a property of its niche—the penetrability of the surface—that is critical to its survival. (Warren 2005, pp. 340-1)

It is the overall job of perceptual systems to keep organisms in contact with the values of relevant organism-environment relationships (the closeness of the obstacle; the penetrability of the surface). Put differently, the world properties it is important to pick out for the purpose of *reconstruction* are not the same as those that best support *interaction*, and psychology has tended to (mistakenly) focus on the former class of properties to the exclusion of the latter.

This, of course, is the thought behind the Gibsonian affordance-based theories of perception that have been widely influential in embodied cognition (Gibson 1979; Orlandi 2014). Affordances are relationships between things in the world and an organism's abilities (Chemero 2009): for the average human the chair (but not the twig) affords sitting, the path (but not the wall) affords walking, and the rock (but not a tiny urticating bristle) affords throwing, but things are different for (and look different to) the bird and the spider. According to the view I am advocating here, perception *is* primarily perception *of* such affordances; the world is seen as a changing set of opportunities for action and interaction. Perception is not for building *models* of the world; it is for building *control systems* for the organism.

Cisek (Cisek 1999) has developed this line of thought in a deeply interesting way:

> As evolution produced increasingly more complex organisms, the mechanisms of control developed more sophisticated and more convoluted solutions to their respective tasks. Mechanisms controlling internal variables such as body temperature or osmolarity evolved by exploiting consistent properties of chemistry, physics, fluid dynamics, etc. Today we call these "physiology". Mechanisms whose control extends out through the environment had to exploit consistent properties of the environment. These properties include statistics of nutrient distributions, Euclidean geometry, Newtonian mechanics, etc. Today we call such mechanisms "behavior". In both cases the functional architecture takes the form of a negative feedback loop, central to which is the measurement of some vital variable. Fluctuations in the measured value of the variable outside some "desired range" initiate mechanisms whose purpose is to bring the variable back into the desired range... The alternative "control metaphor" being developed here may now be stated explicitly: *the function of the brain is to exert control over the organism's state within its environment.* (Cisek 1999, pp. 8-9. emphasis in original)

The beauty of Cisek's analysis here is that he shows how the idea of (negative) feedback loops, which are known to be of crucial importance to the regulatory systems of living things at multiple spatial scales and levels of analysis, can be generalized to cover the case of overt behavior. If that's a valid reconceptualization, and I think it is very, very promising (Anderson 2014), one effect should be to shift our focus from how brain mechanisms like Bayesian predictive coding implement and maintain models of the world, to how such mechanisms enable the feedback loops that maintain attunement to the environment and support adaptive behavior.

## 4 Clark and Coding

It would appear that Clark (Clark 2016) supports this shift of focus. There is evidence for that thought:

> By the end of our story, the predictive brain will stand revealed not as an isolated inner 'inference engine' but an action-oriented engagement machine—an enabling … node in a pattern of dense reciprocal exchange binding brain, body and world. (Clark 2016, p. xvi)

Consistent with the research program I advocated for at length in Anderson (Anderson 2014), Clark suggests that we should understand the search of possibility space that Bayesian predictive processing enables is not the search for the best hypothesis or world model, but rather the search for the best sensorimotor machine:

> Integration (of the rather profound kind exhibited by the neural economy) means that those functionally differentiated areas [of the brain] interact dynamically in ways that allow transient task-specific processing regimes (including transient coalitions of neural resources) to emerge as contextual effects [that] repeatedly reconfigure the flow of information and influence. […] It is the guidance of world-engaging action, and not the production of 'accurate' internal representations, that is the real purpose of the prediction error minimizing routine. […] [It] must find the set of neuronal states that best *accommodate* (as I will now put it) the current sensory barrage. (Clark 2016, pp. 142; 168; 192. Emphasis in original)

And yet, he also endorses positions that would seem perfectly at home in Hohwy's book:

> Perception is controlled hallucination. […] active agents get to structure their sensory flow […] [b]ut it remains correct to say that that all the system has direct access to is its own sensory states (patterns of stimulations across its sensory receptors. […]The task is […] to infer the nature of the signal source (the world) from just the varying input signal itself. […] The ongoing process of perceiving, if such models are correct, is a matter of the brain using stored knowledge to predict, in a progressively more refined manner, the patterns of multilayer neuronal response elicited by the current sensory stimulation. This in turn underlines the surprising extent to which the structure of our expectations (both conscious and non-conscious) may be determining what we see, hear, and feel. (Clark 2016, pp. 14; 16; 27)

I find myself confronted by deep tensions in Clark's account. On the one hand we have a thoroughgoing action-oriented, affordance-laden, sensorimotor coupling account of an agent's ongoing engagement with the world. Bayesian predictive processing emerges here as a crucial mechanism for modulating the agent-environment coupling by dynamically adjusting the parameters of the neural mechanisms that support the agent's capacity to target its behaviors. According to this story, "prediction-driven learning delivers a grip on affordances" (Clark 2016, p. 171) and thereby reveals—allows the perception of—a specifically human world (Clark 2016, p. xv). Here we are thoroughly, actively *in* the world, and it would appear to be impossible (or at least extremely unnatural) to conclude that we are epistemically isolated from it.

And yet we also have, on the other hand, the persistence of the model of perception not as attunement, but as inference from effects to causes, which leads to metaphors of cognitive confinement, of perception as hallucination (p. 14), as an *Ender's Game*-style simulation in which our access to reality is mediated by a virtualization of it (p. 135). As nice as it would be for our expository task here if the inferential model of perception could be understood as one expository step on the way to understanding, to be discarded as enlightenment is attained, in fact the two models of perception remain in tension quite late in the book. Consider the following:

> [P]rediction errors help select among (while simultaneously responding to) competing higher-level hypotheses, each of which implies a whole swathe of sensory and motor predictions. Such high-level hypotheses are intrinsically affordance-laden. (Clark 2016, p. 187)

Elsewhere on that same page, Clark suggests that neural mechanisms also select "salient representations that have affordance [i.e.] sensorimotor representations that predict both perceptual and behavioral consequences." (p. 187) Note the slide from getting a "grip" on affordances (in the world) to *representing* them (in the head). Although Clark is at pains to deny that such representations act as epistemic mediators (p. 195), it is pretty clear we've taken at least a half step back inside the mind. Why? Because Clark cannot quite shake the influence of the inferential model of perception. Discussing the epistemic internalism adopted by many advocates of the predictive processing framework, he writes:

> There is something right about all this and something (or so I shall argue) profoundly wrong. What is right is that accounts on offer depict perception as *in some sense* an inferential process: one that cannot help interpose *something* (the inference) between causes (such as sensory stimulations or distal objects) and effects (percepts, experiences). (Clark 2016, p. 170, emphasis in original)

I'm not quite sure how to analyze the tension in Clark's account, but I suspect it is driven by his quasi-neo-Kantian account of perception, according to which experience is not the immediate apprehension of the material world, but sensation transformed by the subject. Hohwy unabashedly accepts that perceptual content is driven by top-down expectations, but Clark is trying to avoid the internalist consequences of that view by providing a rather greater epistemic role for the incoming sensory signal. Like the Kantian percept that draws concepts into operation, Clark's sensory signals "select" the representations that determine the content of perceptual experience. Moreover, he suggests that increasing the gain on the sensory signal is epistemically equivalent to changing the amount or degree of content that sensory information drives:

> This weighting determines the balance between top-down expectation and bottom-up sensory evidence. That same balance, if the class of models we have been pursuing is on track, determines what is perceived and how we act. (Clark 2016, p. 221)

Just as Kant attempted to avoid Humean skepticism by insisting on the necessity to experience of both percept *and* concept, so Clark tries to avoid Howhian idealism by insiting on the importance to perception of both sensation and prediction. It is because of the balancing act between bottom-up and top-down signals that Clark suggests that we should say we have "not-indirect perception" (Clark 2016, p. 195). The double negative speaks volumes about Clark's conceptual struggle, here. Now, neither Clark nor Hohwy actually offer a detailed theory of content (and, indeed, there has been little work along these lines, but see Gładziejewski 2016, and Wiese 2016), so my analysis is an admittedly speculative attempt to make sense of the tension in Clark's account. But where Hohwy can probably avail himself of a fairly standard account of narrow content (e.g. conceptual role semantics like Loar 1988; Chalmers 2002; or a more radical internalism such as Segal 2000), it is very unclear what theoretical resources Clark has to work with. This is because his externalism would seem to require the sensory signal to carry content (perhaps to be analyzed along the lines of a causal theory of content, e.g. Dretske 1981), but he at the same time accepts Hohwy and Friston's contention that top-down expectations carry (and at least partly determine) perceptual content. From this emerges the notion that these two sources of content must be "balanced" and together determine the content of experience.

I do not know whether or how this can be made to work. If we admit, as Clark appears to do, that the top-down prediction gives us only *indirect* knowledge of the world, then the direct access would have to be provided by the bottom up signal, perhaps by having content determined by the external cause of the signal. But by hypothesis the bottom-up sensory input is an *error* signal, specifying not

the world itself, but its deviation from predictions. Because the epistemic role played by the error signal is updating predictions/expectations, it appears that its epistemic effects can *only* ever be indirect (whatever its causal origins), leaving the top-down expectations to directly determine the content of perceptual experience (more on this in the next section). Although I am in complete agreement with Clark's *goals* in offering an externalist, embodied, action-oriented account of predictive coding, I must reluctantly conclude that he has not in fact fully succeeded in avoiding some undesirable epistemic consequences of the view. At the very least, building an externalist theory of content that is consistent with the whole of Clark's account faces significant challenges.

In the next section, I will suggest that the way forward involves denying that *either* the top-down *or* the bottom-up signal carry information, and adopting not a causal but a guidance theory of content, according to which content is determined by the intentional directedness of action, and not by the causal origins of perceptions.

## 5 Structure, Information, and Bayesian Mechanism.

There is perhaps no term in the cognitive sciences that is more abused than "information". There is information in the world, in the brain, in the genes, in the sensory signal…information is ubiquitous, and so are scientific references to it. "Information" does an immense amount of work for us. However, if Oyama (Oyama 2000/1985) is right, and I think she is, this is a serious problem. I do not have space here to offer a complete account of her argument, but by way of illuminating my motivations for adopting her fairly radical solution of simply *denying* that there is information in any of these places, I believe she convincingly demonstrates that what she labels "preformationism", the notion that information exists before its utilization or expression, lies behind some of our more persistent and pernicious dualisms, including mind-body, nature-nurture, person-situation, and nativism-empiricism, just to name a few. I would add internalism-externalism to that list, insofar as elements of that debate can seem to turn on determining just where content-specifying information comes from: the mind or the world?

Her argument is complex and subtle, but we can get most of the way to her conclusions by reflecting on the fact that a message has a meaning only on the assumption of a specific receiver or decoder. If different consumers extract different meanings from the same message, then the notion that it carries "information" is, if not false, then essentially useless for a theory of perceptual content. Insofar as a signal can carry the same e.g. Shannon-Weaver information relative to some specific situation, but trigger the instantiatiation of different semantic properties[7] at different times or in different people, it makes little sense to think of those properties as being "in" the signal or information. If information is not inherently meaningful, supposing it specifies content is a mistake. It isn't even clear that one can substitute "structure" for "information" here, for although signals and messages are certainly and necessarily structured, the notion that it has "a" structure that thereby fixes content falls to the same considerations that deny it can be said to have "a" meaning: for different decoders, different aspects of its structure may be relevant.

What I want to argue here is that such considerations, along with my analysis of the purpose of perception offered above, should motivate us to give up on information-processing accounts of mind in favor of a developmental systems account. Central to such an account is the notion of *mutual constraint* between multiple interacting influences, at multiple spatial and temporal scales of analysis. Oyama develops the idea most fully in an account of gene-x-environment interactions in development, but I believe we can apply it to mind-world interactions as well. The root of Oyama's investigative problem is the question of whether organism-specifying information is in the genome, the environ-

7 Thanks to Thomas Metziger for this way of putting the matter. Note that when Oyama denies there is information, she is really denying that there is content-specifying or outcome-specifying or otherwise inherently significant information. I do not believe she would deny (and I myself certainly do not) that any given signal can be said to have Shannon-Weaver information, but she would rightly point out that such information is specified *relative to a sender, receiver, and situation*.

ment, or both. Consider a gene that leads to an eye when expressed in its normal milieu, but that leads to a limb when expressed in a bodily location that normally produces neurons. What sense can be made of this? Surely the conclusion must be that neither the gene nor the environment specified what we thought when we observed normal development. Multiply this example by the myriad observations of such variability of phylogenetic outcome and it is hard to resist the conclusion that neither specifies any phylogenetic trait at all.

Similarly, one can ask where the content of experience comes from, the mind, the world, or both? Presumably, everyone will agree that the answer must be both, for we are not the passive recipients of the word's imprint, but active epistemic agents, shaping, categorizing and otherwise selecting incoming percepts. Yet, as in the case of gene-x-environment interactions, the 2-source solution is inherently unstable. Consider: if the error signal specifies a world-state, there isn't a need for the predictive model to specify that state (and we have epistemic support for a kind of passive externalism). If the error signal does *not* specify a world-state, then it must be the predictive model that does, and we oscillate back to internalism. But if content is somehow irreducibly determined by both, such that it is different from what either specifies on its own (or, to put it somewhat differently, if what each specifies or contributes to the whole depends on the state of the other) then it makes little sense to say that either intrinsically specifies anything at all.

On this sort of view, perceptual content is determined not by top-down expectations, nor by the incoming sensory stream, nor, Clark's solution, by both in varying degree. All these solutions have in common the preformationist fallacy that it is possible to specify the contribution of each interacting element in light of the information it brings to the interaction. This is what Oyama is at pains to deny. In systems marked by mutual causal constraint, what Oyama sometimes calls reciprocal selectivity, it is not generally possible to parse this out. Instead, we should say that perceptual experience is determined by the mutual constraint between the incoming sensory signal and ongoing neural and bodily processes, and no aspect of that content can be definitively attributed to either influence. As Oyama expresses the point:

> A structured system selects its stimulus—indeed, defines it and sometimes produces it (the state of the system determines the kind and magnitude of stimulus that will be effective, and intrasystemic interactions may trigger further change)—and the stimulus selects the outcome (the system responds in one way rather than another, depending on the impinging influence). Nativists have generally focused on the former, while empiricists have stressed the latter. In doing so, they have perpetuated and further polarized the opposition between fated internal structure and fortuitous outside circumstance. The mutual selectivity of stimulus and system applies to causal systems of all sorts, and illustrates the impossibility of distinguishing definitively between internal and external control, the inherent and the imposed, selection and instruction. (Oyama 2000/1985 p. 68)

What an ongoing interaction *means* to an organism depends on the state of that organism, and is expressed in what the stimulus (and the organism) *does*. The sensory stimulation does not carry information independent of its causal effects, and these effects irreducibly depend on the state of the system within which the stimulation is occurring. From the standpoint of the project of this essay, what that means is that we should offer *only* a causal, but not an epistemic, information-processing, or representational gloss on the different roles of the top-down and bottom-up signal in predictive coding. Perceptual content is determined by the mutual constraint imposed between the interacting elements, which itself depends (and here I am in 100% agreement with Clark) on "transient task-specific processing regimes (including transient coalitions of neural resources) [that] emerge as contextual effects repeatedly reconfigure the flow […] of influence."(Clark 2016, p. 142) Put differently (and here I offer the seed of a new analysis of perception), *structure* (in the sensory inputs, inner states, and neural and bodily processes) may be intrinsic, but content is not. Perception is an event characterized by

the co-determination of an ongoing behavioral process by both inner and outer conditions. Structure becomes information via such events.

And how should we analyze the content of experience that dynamically emerges in these interactions? I believe that the guidance theory (Anderson and Chemero 2009; Anderson and Rosenberg 2008) is well suited to the job. According to the guidance theory, the intentionality of content rests on the fundamental intentionality of action. Put differently, content *is* what content *does*, and what it does is provide guidance for action. A full formalization of the theory is offered in (Anderson and Rosenberg 2008), but roughly speaking, a percept P (or representation R) is *of* an entity E just in case P (or R) is used to guide an agent's action with respect to E.[8] On the view I am developing here, then, Bayesian updating should be understood as one crucially important neural process that gets the parameters of the mechanism properly set to guide behaviors to their targets.

## 6 Conclusion

In this essay I have attempted to trace the underlying causes of Hohwy's internalism to a faulty conception of the nature of our epistemic situation. I outline an alternative, embodied, situated and action-oriented perspective that should allow us to take predictive coding on board, while avoiding its internalist consequences. I then puzzle over the fact that, although Clark is fully on board with the action-oriented perspective I advocate, he hasn't quite entirely avoided internalism. I suggest it may be because he has replaced Hohwy's one-content-source account of perception with a two-content-source view, when he should have instead rethought the epistemic interpretation of predictive coding entirely. I then briefly offer a *merely causal* account of predictive coding, and gesture at a theory of perceptual content that fits nicely with the action-oriented perspective developed here.

8 The guidance theory is neutral and pluralist on the question of *how* guidance representations serve their representational function of guiding action. Guidance representations can be map-like, or emulators, or pictures, or something else entirely, and we expect different kinds of representations to be used in different circumstances and by different mental systems/subsystems.

## References


Akins, K. (1996). Of sensory systems and the aboutness of mental states. *The Journal of Philosophy, 93*, 337–372.

Anderson, M. L. (2003). Embodied cognition: A field guide. *Artificial Intelligence, 149* (1), 91–130.

——— (2006). Cognitive science and epistemic openness. *Phenomenology and the Cognitive Sciences, 5* (2), 125–154.

———(2014). *After phrenology: Neural reuse and the interactive brain.* Cambridge, MA: MIT Press.

Anderson, M. L. & Chemero, A. (2009). Affordances and intentionality: Reply to Roberts. *Journal of Mind and Behavior, 30* (4), 301.

Anderson, M. L. & Rosenberg, G. (2008). Content and action: The guidance theory of representation. *Journal of Mind and Behavior, 29* (1–2), 55–86.

Burr, C. & Jones, M. (2016). The body as laboratory: Prediction-error minimization, embodiment, and representation. *Philosophical Psychology, 29* (4), 586–600. https://dx.doi.org/10.1080/09515089.2015.1135238.

Chalmers, D. (2002). The components of content. In D. Chalmers (Ed.) *The philosophy of mind: Classic and contemporary readings* (pp. 607–633). Oxford, Oxford University Press.

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

Cisek, P. (1999). Beyond the computer metaphor: Behaviour as interaction. *Journal of Consciousness Studies, 6* (11–12), 125–142.

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

———(2015). Predicting peace: The end of the representation wars. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. https://dx.doi.org/10.15502/9783958570979. http://open-mind.net/papers/predicting-peace-the-end-of-the-representation-wars.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

Devitt, M. (1981). *Designation.* New York: Columbia University Press.

Dretske, F. (1981). *Knowledge and flow of information.* Cambridge, MA: MIT/Bradford Press.

Gallagher, S. & Bower, M. (2014). Making enactivism even more embodied. *AVANT, 5* (2), 232–247.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton-Mifflin.

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

Grice, H. & White, A. (1961). Symposium: The causal theory of perception. *Proceedings of the Aristotelian Society, Supplementary Volumes, 35*, 121–168. http://www.jstor.org/stable/4106682.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, 559–582. https://dx.doi.org/10.1007/s11229-015-0762-9.

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

Kolmogorov, A. N. (1942). Determination of the center of scattering and the measure of accuracy by a limited number of observations. *Izvestiia Akademii Nauk SSSR. Series Mathematics, 6*, 3–32.

Kolmogorov, A. N. & Hewitt, E. (1948). *Collection of articles on the theory of firing.* Santa Monica, CA: RAND Corporation.

Kripke, S. (1980). *Naming and necessity.* Cambridge, MA: Harvard University Press.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

——— (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought.* New York: Basic Books.

Loar, B. (1988). A new kind of content. In R. H. Grim & D. D. Merrill (Eds.) *Contents of thought: Proceedings of the Oberlin Colloquium in Philosophy.* (pp. 117–139). Tucson, Arizona, University of Arizona Press.

Mackie, J. (1965). Causes and conditions. *American Philosophical Quarterly, 2* (4), 245–264.

O'Donovan-Anderson, M. (1997). *Content and comportment: On embodiment and the epistemic availability of the world.* Lanham, MD: Rowman & Littlefield.

O'Regan, J. K. & Degenaar, J. (2014). Predictive processing, perceptual presence, and sensorimotor theory. *Cognitive Neuroscience, 5* (2), 130–131. https://dx.doi.org/10.1080/17588928.2014.907256.

Orlandi, N. (2014). *The innocent eye: Why vision is not a cognitive process.* New York: Oxford University Press.

Oyama, S. (2000/1985). *The ontogeny of information: Developmental systems and evolution.* Durham, NC: Duke University Press.

Segal, G. (2000). *A slim book about narrow content.* Cambridge, MA: MIT 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. https://dx.doi.org/10.1080/17588928.2013.877880.

Titchener, E. B. (1929). *Systematic psychology: Prolegomena.* New York: MacMillan.

Warren, W. (2005). Direct perception: The view from here. *Philosophical Topics, 33* (1), 335-361.

Wiese, W. (2016). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 1-22. https://dx.doi.org/10.1007/s11097-016-9472-0.

Wiese, W. & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing.* Frankfurt am Main: MIND Group.

# Active Inference and the Primacy of the 'I Can'

Jelle Bruineberg

This paper deals with the question of agency and intentionality in the context of the free-energy principle. The free-energy principle is a system-theoretic framework for understanding living self-organizing systems and how they relate to their environments. I will first sketch the main philosophical positions in the literature: a rationalist Helmholtzian interpretation (Hohwy 2013; Clark 2013), a cybernetic interpretation (Seth 2015b) and the enactive affordance-based interpretation (Bruineberg and Rietveld 2014; Bruineberg et al. 2016) and will then show how agency and intentionality are construed differently on these different philosophical interpretations. I will then argue that a purely Helmholtzian is limited, in that it can account only account for agency in the context of perceptual inference. The cybernetic account cannot give a full account of action, since purposiveness is accounted for only to the extent that it pertains to the control of homeostatic essential variables. I will then argue that the enactive affordance-based account attempts to provide broader account of purposive action without presupposing goals and intentions coming from outside of the theory. In the second part of the paper, I will discuss how each of these three interpretations conceives of the sense agency and intentionality in different ways.

**Keywords**

Active inference | Affordances | Cybernetics | Free energy principle | Helmholtz | Phenomenology | Sense of agency | Skilled intentionality

**Acknowledgements**

Thanks to Joel Krueger for discussing a very early draft of this paper in 2013. Thanks to the participants of MIND23 for inspiring discussions that shaped the paper in its current form and thanks to Julian Kiverstein, Thomas Metzinger, Erik Rietveld, Martin Stokhof, Wanja Wiese and two anonymous reviewers for critical feedback on the paper. This research was financially supported in the form of VIDI-grant by the Netherlands Organisation for Scientific Research (NWO) awarded to Erik Rietveld.

## 1 Introduction

After computationalism, connectionism, and (embodied) dynamicism, cognitive science has over the last few years seen the resurgence of a paradigm that might be dubbed "predictivism": the idea that brains are fundamentally in the business of predicting sensory input. This paradigm is based on older ideas in psychology and physiology (Von Helmholtz 1860/1962), and has been revived by parallels that have been discovered between machine learning algorithms and the anatomy of the brain (Dayan and Hinton 1996; Friston et al. 2006). The emergence of the paradigm of "predictivism" has sparked great interest in philosophy of mind and philosophy of cognitive science, mainly through the work of Clark (Clark 2013; Clark 2016) and that of Hohwy (Hohwy 2013; Hohwy 2016). This interest has led to a vast number of papers attempting to ground concepts from phenomenology, philosophy of mind and psychopathology in predictive architectures (see for example Hohwy 2007; Limanowski and Blankenburg 2013; Apps and Tsakiris 2014; Hohwy et al. 2016).

Predictivism might be better off than these earlier paradigms in cognitive science, exactly because most of its core ideas are *not* very new. As Clark writes in the introduction to his book:

> [W]hat emerges is really just a meeting point for the best of many previous approaches, combining elements from work in connectionism and artificial neural networks, contemporary cognitive and computational neuroscience, Bayesian approaches to dealing with evidence and uncertainty, robotics, self-organization, and the study of the embodied environmentally situated mind. (Clark 2016, p.10)

Bruineberg, J. (2017). Active Inference and the Primacy of the 'I Can'.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 5. Frankfurt am Main: MIND Group. doi: 10.155027/9783958573062

To put it in Kuhnian (Kuhn 1962) terms, for Clark we might currently see the transition of cognitive science from a pre-paradigmatic stage, with competing paradigms developed by incompatible schools of thought, to *normal science* in which one dominant paradigm provides the concepts and questions to be solved. Whether or not this is true is for Kuhn a question that can only be answered in hindsight. In any case, by providing a meeting point for these different approaches, "predictivism" simultaneously also provides a new battleground for competing schools of thought in philosophy of mind concerning internalism and externalism, embodiment, and computationalism.

Currently, it is unclear whether "predictivism" entails a particular philosophical position, and whether "predictivism" tells us much about the nature of cognition without these philosophical assumptions frontloaded. Different scientists and philosophers working on predictive-coding take different, supposedly mutually incompatible starting points: a Helmholtzian theory of perception (Hohwy 2013; Clark 2013), Ashbyian cybernetics (Seth 2015b) and an enactive affordance-based account borrowing from Merleau-Ponty and Gibson (Bruineberg and Rietveld 2014; Bruineberg et al. 2016; Rietveld et al. forthcoming)[1]. To me, there seems to be little hope to settle philosophical issues concerning embodiment and the mind-world relationship deriving from a theory-neutral presentation of predictive processing (PP). In fact, as mentioned in the introductory chapter (Wiese and Metzinger 2017) a theory-neutral presentation of PP seems itself unfeasible. Rather, much of the literature poses a problem in which a philosophical worldview is presupposed and then shows the compatibility of PP with this view, be it about using sensory input to represent a distal world (Hohwy 2016, p. 1), tending towards grip on a field of affordances (Bruineberg and Rietveld 2014, p. 7) or the problem of homeostatic regulation and interoceptive inference (Seth 2015b).

In this paper, I will focus on how to conceive of agency and the sense of agency under the free-energy principle (FEP). The free-energy principle is the most theoretical and all-encompassing version of the "predictivist" approach, being compatible with, but not limited to, predictive-coding accounts of the brain. In itself, the free-energy principle is a system-theoretic framework for understanding living self-organizing systems and how they relate to their environments. I will first present the main tenets of the free-energy principle and consequently present three different philosophical approaches to the free-energy principle: a rationalist approach (based on Helmholtz), a cybernetic approach (based on Ashby) and an enactive affordance-based approach (based on Merleau-Ponty and Gibson). I will argue that whereas the rationalist and cybernetic approaches face a number of conceptual problems in construing agency under the free-energy principle, these conceptual problems can be resolved by the enactive affordance-based approach.

## 2 Main Tenets of the Free-Energy Principle

In this section I will give a non-mathematical treatment of the basic tenets of the free-energy principle, introducing the main assumptions and reasoning steps that lead to its formulation. (For an introduction to predictive processing and the free-energy priniciple more generally, see Wiese and Metzinger 2017, and references therein.)

As mentioned in the introduction of this paper, the free-energy principle is a proposal for understanding living self-organizing systems (Friston and Stephan 2007; Friston 2011). Based on a descriptive statement (living systems survive over prolonged periods of time), the free-energy principle provides a prescriptive statement (a living system *must* minimize its free-energy) to provide the necessary and sufficient conditions for this descriptive statement to be true. The major premises underlying this move are the following:

1 I do not wish to say that these positions are *a priori* mutually exclusive. However, they do have very different philosophical starting points and it therefore remains to be seen to what extent they are (in)compatible.

1. The embodiment of an animal implies a set of viable states of the animal-environment system.

One can formalize this in information-theoretic terms by assigning a probability distribution to the viable states of the organism. For example, human body temperature has a high probability of being around 37°C and a low probability of being elsewhere. Information theoretically, this means that the event 'measuring a body temperature of 37°C' has low surprisal, while measuring a body temperature of 10°C has a very high surprisal. Remaining within viable bounds can then be understood in terms of minimizing surprisal. For ectothermic (cold-blooded) animals, this directly puts constraints on the places in its environment that it may seek out (i.e. a lizard seeking out a sunny rock in the morning). For endothermic (warm-blooded) animals, this means it needs to find energy sources to sustain its metabolism and, in some cases, seek shelter to complement its internal heat regulation. In short, with a particular agent we can identify a probability distribution of the states the agent typical frequents and has to frequent. I will call this distribution the *embodied* distribution and the surprisal of an event relative to this embodied distribution *embodied* suprisal (see Bruineberg et al. 2016, for a more elaborate introduction of this vocabulary and see Wiese and Metzinger 2017, for an informal analysis of how this distribution can be found based on the typical states the animal frequents and the assumption of ergodicity).

2. The animal's regulatory system (for instance the nervous system) does not have access to the viable states of the agent-environment system. Instead it needs to estimate them.

A regulatory system needs to minimize surprisal without being able to evaluate it directly. It cannot evaluate surprisal directly, because the embodied probability distribution of the viable states of the organism is not known to it. This is where free-energy comes in. Free-energy is a function of sensory states and estimated worldly states that generated the sensory states and involves two probability densities:

- A generative density $p(w, s|m)$, specifying the joint probability of sensory state $s$, and worldly states based upon a probabilistic model $m$ embodied by the agent.
- A recognition or variational density $q(w; b)$, encoding the agent's 'beliefs' about the worldly states entailed by its internal state $b$.

Free-energy is defined in terms of these two densities:

$$F(s,b) = -\int_w q(w;b) \ln \frac{p(w,s|m)}{q(w;b)} dw$$

The free-energy formulation can be rearranged so as to show its dependence on perception and action respectively (see Friston and Stephan 2007; McGregor et al. 2015). The basic idea behind the free-energy framework is that whatever shape or form the recognition density takes, free-energy over the long run related to this *estimated* recognition density will be equal to or greater than the surprisal I receive at any point in time related to the *embodied* distribution. The long-term average of free energy (obtained by integrating over the temporal domain) is called free action.

The quantity of free-energy is a function of sensory states and estimated worldly states and priors. Each of these can change in order to minimize free-energy: optimizing estimated worldly states (typically called perceptual inference), optimizing sensory states (brought about through action), and optimizing the generative model (learning).

3. In order to stay alive, it suffices for the animal to stay within the viable states of the animal-environment system. It does so by minimizing free-energy using its estimated conditions of viability as priors.

The assumption here is that the internally estimated conditions of viability and the real (embodied or intrinsic) distribution are similar enough to make adequate regulation possible (i.e. my regulatory

system should not anticipate a body temperature of 10°C). Homeostatic control can then be achieved by predicting particular sensations corresponding to a body temperature of around 37°C and minimizing the discrepancy from, or prediction-error with respect to that implicit hypothesis (the expectation of body temperature of around 37°C). The logic here is that through evolution and development the agent comes to expect itself to be in an optimal state and continually minimizes the discrepancy between its current state and its optimal expected state.

4. To achieve homeostatic control, the animal needs to be able to act on the world. This implies, at least implicitly, a model of how actions lead to changes in interoceptive and exteroceptive sensory input. Since the state of the world mediates how actions change perception, optimal regulation requires taking the state of the world into account. This process of optimal regulation while taking the estimated state of the world into account is called 'active inference'.

As we will see later in this paper, perceptual inference will only minimize free-energy to the extent that it becomes a tighter upper bound on embodied surprisal: the animal becomes more and more certain of it being in a too cold state. Only action will change embodied surprisal (the animal being in a too cold state) itself. For example, the ectothermic lizard needs to be able to compensate its interoceptive prediction-error by moving around in a way that makes it seek out the warm rock in the sunshine.

To summarize: to continue being a living creature is to maintain oneself in a particular type of environment. For example, to be a fish the fish must maintain itself in a fish-like environment. This is possible for the fish through its prediction of the sensory input associated with a fish-like environment (a certain pressure, temperature, light etc.) and through its actions (being able to avoid, accommodate and counteract mismatches between predicted sensory input and actual sensory input). This implies a particular kind of congruence between the dynamics and structure of the environment and of the organism, to which I will return in a later section.

## 2.1 Prediction-Error Minimization and the Free-Energy Principle

The free-energy principle does not provide in and of itself a mechanism for realizing free-energy minimization. However, it often gets paired, or even conflated, with the prediction-error minimization framework (PEM)[2]. PEM can best be introduced as a form of Bayesian model-based statistical inference. The animal possesses an internal model of the possible causal structures of the world and the kind of sensory information associated with these causal structures. Based on its priors and sensory states, weighted by its confidence in both, it can then infer the hidden state of the environment based on a series of sensory states. Adequate inference and adequate prediction are then two sides of the same coin.

Much work in computational neuroscience and machine learning has been carried out in the PEM framework with the aim of understanding how inference through prediction-error minimization is possible in brains. One important feature is that the generative model is hierarchical: each layer of the hierarchy tries to predict the information it is receiving from a lower level (Friston 2008). Another central feature of this work concerns whether the agent's probability distribution is updated so as to approximate Bayes' theorem as the agent is exposed to new sensory input, and if so, which approximation algorithms work best. These developments might make predictive-coding neural architectures a good computational implementation for free-energy minimization.

However, there remains a number of conceptual tensions between machine learning approaches and the free-energy principle that will be the main focus of my discussion in the remainder this paper. They concern the role of action: is action auxiliary in obtaining the most likely hypothesis or is action

2 For example, Hohwy writes: "Since the sum of prediction error over time is also known as free-energy, PEM is also known as the free-energy principle" (Hohwy 2016, p. 2).

goal-directed? If there is no strong distinction between the two, how can we conceive of both the epistemic and the goal-directed function of action simultaneously? In the next section, we will see how different philosophical approaches respond to these questions.

## 3 Philosophical Interpretations of Predictivism

In this section, I will present three philosophical approaches to the free-energy principle: a Helmholtzian approach (Hohwy 2013; Clark 2013), a cybernetic approach (Seth 2015b) and an enactive affordance-based account borrowing from Merleau-Ponty and Gibson (Bruineberg and Rietveld 2014; Bruineberg et al. 2016). I will discuss how they conceive of action under the free-energy principle.

### 3.1 Helmholtz and Hypothesis-Testing

The standard point of departure for "predictivist" approaches to the mind is Helmholtz's (Wiese 2017) notion of perception as unconscious inference (more recently, Gregory (Gregory 1980) articulated the idea of perception as hypothesis-testing). The basic idea is that perception is essentially continuous with the scientific method. That is to say, the perceptual system holds a hypothesis (or a range of hypotheses) with a certain degree of confidence. Incoming data might corroborate the current hypothesis, cause the system to change its hypothesis, or cause it to abandon it altogether (i.e. shift to a new hypothesis). By iterating this process over time, the system comes to infer the true hidden state of the environment. Expected precision (i.e. degree of confidence) of both the hypothesis and the sensory input plays a crucial role in how (and whether) one settles on a particular hypothesis. For example, in a perceptual decision-making trial, I might start out in a very low confidence state. Over time, while sensory information comes in, I develop a hypothesis (say, dots on average moving to the right) that explains away the prediction-error. Over time, the confidence in the hypothesis grows until a threshold is reached. Noise in the system has a high impact in the beginning of the trial, when confidence in the current hypothesis is low and has low impact in the end, when there is high confidence (see Bitzer et al. 2015, for an example of such a Bayesian model of perceptual decision-making).

The Helmholtzian perspective seems to work well for perceptual inference. For Helmholtz, as for Gregory, it is *perception* that is inferential. They implicitly endorse a 'sandwich model of the mind' (Hurley 1998): perception supplies input to the cognitive systems, which figure out what to do next, and action translates decisions into motor commands. This is not to say that Helmholtzians think perception is "dumb" or passive. Contrary to other sandwich models (Fodor 1983), Helmholtzian perception is thought to be active and knowledge based. Modern "predictivist" accounts depart from Helmholtz in the sense that they deny the sandwich model, and attempt to closely intertwine perception and action in what is called "active inference". Regardless of these differences, I take the basic commitments of Helmholtzian cognitive science to be 1.) That the *aim* of perception and action is to disambiguate the hidden causal structure of the environment, and 2.) That the *means* by which this *aim* is achieved is by some process continuous with or analagous to scientific hypothesis-testing

At first glance, active inference might only seem to strengthen the link between the workings of the mind and scientific inferences. The way Friston (Friston et al. 2012) and Hohwy (Hohwy 2013) add action to perceptual inference is by appealing to setting up experiments. Scientific hypothesis testing is not just passively recording results of data coming in, but carefully setting up experiments and actively intervening in the chains of causes and effects in order to disambiguate the hidden causes of sensory input. Hohwy relates this to causal inference (Pearl 1988) in which the system is able to calculate where to intervene in order to disambiguate between causal structures. This is an elegant way of combining perception and action under the umbrella of "predictivism".

However, the only demand on a perceptual system is whether it is adequately able to infer, represent and predict the hidden state of the environment. It lacks an account of motivation, value and reward—on whether a particular environmental state is conducive or detrimental with respect to its

bodily needs, habits or plans. As we have seen in section 2, the free-energy principle *does* aim for such deeper integration of prediction and motivation (Friston et al. 2009). The way in which it does so is by reducing the traditional roles of cost functions, value and reward to prediction. However, I will now argue that, in doing so, it fundamentally changes the very nature of prediction error-minimization.

Consider the following example as an intuition pump: suppose I am standing under a steaming hot shower. This will lead to prediction-errors on the skin. There may be some physiological reactions that might help to reduce temperature (such as vasodilation), but the most obvious reaction would be to get out of the shower, or to manually change the temperature. This requires an implicit generative model of how interoceptive and exteroceptive prediction-errors change with particular actions and not others, while taking into account the peculiarities of the shower I am standing under now.

What is important here is that the most *likely* cause of the sensory input I am receiving is the fact that "I am standing under shower that is too hot" and any "experiment" I set up will corroborate that hypothesis. What the system *needs* to do is to treat burning under a hot shower as *extremely unlikely*. Since it is extremely unlikely, I cannot accept this hypothesis, but rather am forced to change the world so as to reduce prediction-errors with respect to the hypothesis that "I am standing under a comfortably warm shower". To emphasize: although FEP treats the current state as highly unlikely, it *is* the actual state I find myself in and if I do nothing I *will* get burnt. *If* the aim of the Helmholtzian account is solely to figure out what the hidden state of the world is, then "I am standing under a shower that is too hot and will get burned" will be the hypothesis it settles for, but it is not. This gives rise to the crooked scientist argument: if one wishes to compare the activities of the brain to that of a scientist, it *needs* to be a 'crooked scientist'. The brain acts like a scientist invested in ensuring the truth of a particular theory, which is the theory that "I am alive"[3]. As contradictory evidence comes in, it manipulates the world until the perpetual truth of that theory is ensured (or dies trying) (Bruineberg et al. 2016).

I believe there is an important shift here in the conceptualization of active inference. On the Helmholtzian picture, a system does better when it is more accurate and precise in its representations of the causal structure of the environment, i.e. when it delivers true representations that the perceiver has high confidence in. On the ecological enactive picture that I will sketch, a system performs better when it supports the system's movement towards an optimal state, where that optimal state is to be understood relative to the animal's conditions of viability/flourishing. In that sense, active inference requires a thoroughly optimistic generative model of how the animal expects to flourish in its econiche. Only with such an optimistic generative model will active inference lead to adaptive behavior.

Note that within this picture there is ample room for epistemic actions like those the scientist performs in carefully setting up an experiment. Consider the everyday example of standing in a small space under a too hot shower while having shampoo in your eyes. One can imagine the following response to the situation: I first orient myself, for instance by touching the wall, and then reach for the tap to turn down the temperature. The first action can be seen as largely epistemic, the second one as largely goal-directed[4]. However, epistemic actions unfold within the context of the movement towards an optimal state, where the optimality, in this context, is grounded in the system's conditions of viability/flourishing.

Hohwy also seems to want to ground or justify prediction-error minimization by appealing to biological self-organization. He raises the issue in the context of Kripke's (Kripke 1982) interpretation of Wittgenstein's (Wittgenstein 1953) rule following argument. The skeptical question, phrased in predictive-coding terms, is whether it makes sense to say of someone that he or she correctly obeys the imperative of minimizing prediction-error. What is the fact of the matter about that person that

3 Of course, staying alive underdetermines what to do in everyday situations. For such cases, enacting a (more or less coherent) identity, "flourishing" or having grip on the situation might be better suited notions.

4 In a recent paper, Friston et al. (Friston et al. 2015) show that one can mathematically decompose free-energy minimization into epistemic value and extrinsic (goal-directed) value (Seth 2015a calls these epistemic and instrumental, respectively). Epistemic value serves to reduce uncertainty related to hidden states of the world (i.e. the location of the tap), while extrinsic value serves to bring the agent closer to an optimal state.

justifies the assertion that he or she is correctly obeying that rule? Hohwy states: "The answer cannot be something along the lines of: you should minimize prediction-error because it leads to truthful representations. That answer is couched in semantic terms of 'true representations', so it is circular" (Hohwy 2013, p.180). He finds his way out of this circularity by appealing to a non-semantic feature of our existence: self-organization. "I should minimize prediction error because if I do not do so then I fail to insulate myself optimally from entropic disorder and the phase shift (that is, heat death) that will eventually come with entropy" (Hohwy 2013, p.181). I agree with Hohwy that any notion of predictive-processing needs, in order to avoid being circular, ultimately to be grounded in the requirement for biological self-organization. But I disagree on the constraints this requirement places on how to conceptualize PEM. Hohwy continues:

> Perhaps we can put it like this: misrepresentation is the default or equilibrium state, and the fact that I exist is the fact that I follow the rule for perception, but it would be wrong to say that I follow rules in order to exist — I am my own existence proof. (Hohwy 2013, p.181)

If by 'the rule of perception' Hohwy means that our perceptual system is in the business of minimizing prediction-error (and action is at most auxiliary), then we are in disagreement about what grounding PEM in biological self-organization entails for PEM. As the crooked scientist argument shows, accepting the requirements for biological self-organization entails a shift from tending towards a more truthful representation to tending towards a more optimal agent-relative equilibrium. It is exactly this shift, and its implications for how to conceptualize agency, that remain concealed on the Helmholtzian account.

As mentioned above, I take the basic commitments of the Helmholtzian cognitive scientist to be 1.) That the *aim* of perception and action is to disambiguate the hidden causal structure of the environment, and 2.) That the *means* by which this *aim* is achieved is by some process continuous with scientific hypothesis-testing. Based on Friston et al. (Friston et al. 2015) and the crooked scientist argument, I take it that commitment 1 is false in a strict sense. At best, figuring out the hidden structure of the world is *auxiliary* to moving towards a more optimal state. I take commitment 2 to be false in a strict sense as well. In Helmholtzian language, perception and action serve to optimize the likelihood of the animal's theory that it is alive. If certain "experiments" don't give the right answer, the animal will switch to performing new "experiments" that do give the right answer: I change the temperature of the shower and not the hypothesis about the kind of being that I am (i.e. one that survives at 37°C).

The 'crooked scientist argument' is problematic for those who wish to endorse both the free-energy principle and a Helmholtzian theory of cognition. The Helmholtzian metaphor gives you exactly the *wrong* intuitions about some core aspects of the free-energy principle. The intuition in active inference should not be, as Hohwy claims, how the brain can use available sensory input to accurately reconstruct the hidden state of affairs in the world (Hohwy 2016, p.1), but rather how the space of possible 'hypotheses' is always already constrained and crooked in such a way as to make the animal tend to optimal conditions. It is the analysis of (organism-relative) value as prediction error that makes the free-energy principle such a challenging framework to understand. Appealing to Helmholtzian inference does very little to make these conceptual difficulties clearer. As I hope to have shown, particular aspects of the Helmholtzian framework might be retained, but overall it does a poor job. Furthermore, I think that both the "cybernetic Bayesian brain" (Seth 2015b) and Merleau-Pontyian cognitive science (in the form of the "Skilled Intentionality Framework" (Bruineberg and Rietveld 2014, figure 1; Rietveld et al. forthcoming) and "Radical Predictive Processing" (Clark 2015), cf. also Downey 2017) might be better alternatives. I will turn to them next.

## 3.2 Ashby and Cybernetics

Rather than starting from Helmholtz, Seth takes the work of cyberneticist Ashby (Ashby 1952; Ashby 1956) as his starting for theorizing about "predictivism". Cybernetics focuses on control systems. The field is dubbed cybernetics after its prototype: the Watt governor (κυβερνήτης is Greek for governor), a clever device capable of stabilizing the output of a steam engine based on a system of rotating flyballs that controls the throttle valve (see e.g. Van Gelder 1995; Bechtel 1998). The point about the Watt governor is that it is able to suppress perturbations in the system and, in doing so, stabilizes the governor-engine system. Inspired by the governor, cybernetics proposes the more general principle that an adaptive system maintains its own organization by suppressing and responding to environmental perturbations. This often includes the control of so-called *essential variables*. For example, in the case of living systems, body temperature and metabolic needs—when action and perception are coupled with a temperature sensor, a system might move through a space in order to seek out a place where the temperature is optimal.

The basic principles of cybernetics seem to fit well with the basic tenets of the free-energy principle: active homeostatic control to stay within viable bounds. Auletta (Auletta 2013) provides a nice example of how a coupled informational (sensorimotor) and metabolic system can provide a model for bacterial chemotaxis and how this can be understood in terms of free-energy minimization. In simple and stable environments, such as are often used in evolutionary robotics, it is sufficient to train a simple neural network to pick up stable regularities between sensors, aspects of the environment and the availability of heat or food. Clark, in his work on embodied predictivism and embodied cognition more generally, Clark (Clark 1997; Clark 2016) has proposed that we understand the internal workings of the animal as such a "bag of tricks" fit to deal with its niche.

Seth's proposal makes progress in relation to one of the main weaknesses of the purely Helmholtzian account: the exclusion of values. Demanding that particular essential variables are kept constant puts constraints on the interactions the animal has with the environment. The example Seth gives is of that of blood sugar level. When blood sugar level is too low, the following responses arise: interoceptive prediction-errors signals travel upward in the brain, which lead to subjective experiences of hunger or thirst. These prediction-errors then travel further upward in the hierarchy where multimodal integration of interoceptive and exteroceptive inputs take place. These high-level models then instantiate predictions that flow down the hierarchy, leading to an autonomic response (metabolize bodily fat stores), or allostatic actions (eating a banana) (Seth et al. 2011; Seth 2015b). Contextual information, about for instance the availability of food (encoded in the precision of the allostatic response hypothesis), might contribute to the decision as to which response is initiated (or whether both are).

Similarly, the cybernetic account can handle the hot shower example given in the previous section. The hot shower will lead to prediction-errors (perhaps showing in the form of pain and dizziness) that stand in need of being reduced. This then puts constraints on the actions I might undertake, leading to the combination of epistemic and purposeful (extrinsic) actions that make me leave the shower or reduce the heat of the shower. In short, the cybernetic account is better suited to explaining adaptive and ecological action than the Helmholtzian account.

One other aspect of the free-energy principle that comes to the foreground on the cybernetic interpretation is the structure of the generative model. If the function of the generative model shifts from inferring the hidden state of the environment to steering the animal towards an optimal state, then the generative model is not just a model of the environment, but rather of the animal situated in its environment. What counts as the most likely state is not the most likely state of the environment *per se* given current sensory evidence and one's prior beliefs, but rather the optimal state for the animal-environment system to be in (Friston 2011). I will return to this point in the next section.

On a purely cybernetic account, all actions are responses to (or responses of anticipations to) deviations from homeostatic variables. Seth's model of active inference (Seth 2015b) integrates both cy-

bernetic and Helmholtzian elements: action can both serve to confirm, disconfirm and disambiguate hypotheses as on the Helmholtzian account and also account for homeostatic behavior. Although the cybernetic account improves upon the Helmholtzian account, it is still limited in the account it gives of active inference. What is lacking is that the optimality conditions that the animal generates are broader than essential variables related to homeostasis. My metabolic needs underconstrain, for instance, in what way I will finish this sentence. Being an academic philosopher, the practices I participate in and the skills I have acquired in these practices, *do* constrain my writing style. Some of these practices and habits, like working and skipping dinner in order to finish a paper, might actually squarely oppose those metabolic needs. The challenge ahead, as I understand it, is to provide an account of non-metabolic purposes without an appeal to goals as unexplained explainers. The answer lies, I believe, in understanding how our purposive actions are situated in a social setting with which we are familiar. For these reasons, I will turn to a Merleau-Pontyian approach to cognitive science next.

### 3.3 The Enactive Affordance-Based Account

A third philosophical perspective on "predictivism" can be distilled from the work of French phenomenologist Maurice Merleau-Ponty. The great insight put forth in Merleau-Ponty's *The Phenomenology of Perception* (Merleau-Ponty 1945/1962) is that, as skilled humans, we have a pre-reflective bodily engagement with the world, prior to any objectification: we bike home from work, cook dinner and have a conversation. In such cases, we do not continuously decide what to do, but are open to and respond to the demands of the situation. According to Merleau-Ponty, a perceptual scene does not show up as a set of objects but is colored or structured by the demands of the situation:

> For the player in action the football field is not an 'object' [..]. It is pervaded with lines of force [..] and articulated in sectors (for example, the 'openings' between the adversaries) which call for a certain mode of action and which initiate and guide the action as if the player were unaware of it. [..]; the player becomes one with it and feels the direction of the 'goal', for example, just as immediately as the vertical and the horizontal planes of his own body. It would not be sufficient to say that consciousness inhabits this milieu. At this moment consciousness is nothing other than the dialectic of milieu and action. Each maneuver undertaken by the player modifies the character of the field and establishes in it new lines of force in which the action in turn unfolds and is accomplished, again altering the phenomenal field. (Merleau-Ponty 1942/1966, p. 168-169)

What we perceive in skilled action are the relevant action possibilities that the situation provides. We perceive these possibilities not as mere theoretical possibilities, but as what Dreyfus and Kelly (Dreyfus and Kelly 2007) call relevant affordances or solicitations.

**Affordance** $=_{Df}$ A possibility for action provided by the environment to an animal.

**Solicitation** $=_{Df}$ An affordance that stands out as relevant for a particular animal in a specific situation.

**Tendency toward an optimal grip** $=_{Df}$ The tendency of a skilled individual to be moved to improve its grip on the situation by responding to solicitations.

What is perceived as relevant depends on the situation, the skill of the agent and socio-material norms the agent is attuned to. Everything the football player has learned through years of practice feeds back in the way the situation appears. This tight coupling between skilled agent and environment, in which every action modifies the experiential field, is what Merleau-Ponty calls "the motor-intentional arc" (Merleau-Ponty 1945/1962; Dreyfus 2002).

There is a second notion borrowed from Merleau-Ponty that is important for our current purposes, and that is the notion of the "tendency towards an optimal grip". This is a primarily phenomenological

notion that signifies the way a skilled individual relates to its environment. Merleau-Ponty gives the example of perceiving a picture in an art gallery: "There is an optimum distance from which it requires to be seen" (Merleau-Ponty 1945/1962, p.352). The details of the painting get lost when we step further away, and we lose the overview of the painting as a whole when we move too close. In a sense, the painting *demands* a particular perspective, just like the situation on the football field demands an action to be made. Note that, for Merleau-Ponty, absolute grip is never obtained, but it is the *tendency towards* grip that guides our actions.

A third insight from Merleau-Ponty that might be of help in the current context is the manner in which active agents bring forth their own world. Clark (Clark 2016, p. 289), drawing upon the continuity between Varela et al. (Varela et al. 1991) and Merleau-Ponty (Merleau-Ponty 1945/1962) writes:

> In a striking image, Merleau-Ponty then compares the active organism to a keyboard which moves itself around so as to offer different keys to the "in itself monotonous action of an external hammer" (Merleau-Ponty 1945/1962, p.13). The message that the world 'types onto the perceiver' is thus largely created (or so the image suggests) by the nature and action of the perceiver herself: the way she offers herself to the world. The upshot, according to Varela, et al. (Varela et al. 1991, p.174) is that "the organism and environment [are] bound together in reciprocal specification and selection.

Now, as Clark is careful to note, the world is more than just a brute hammer, but the important message here is that the active agent meets the world on its own terms. This phenomenon is labelled differently in different traditions: ecological psychologists speak of perturbations being not *given by*, but *obtained from* the world (Turvey and Carello 2012), autopoietic enactivists speak of an *autonomous* system bringing forth significance (Varela et al. 1991; Thompson 2007; Di Paolo 2005). The demand for self-organization provides, for both the free-energy principle and autopoietic enactivism, specific constraints on the circularity (sometimes called "circular causality"; see Tschacher and Haken 2007) between organism and environment: the environment and skilled agent mutually constrain each other in such a way that the overall dynamic remains within a flourishing regime.

### 3.4 Selective Openness and Active Inference

In Bruineberg and Rietveld (Bruineberg and Rietveld 2014), we attempted to frontload Merleau-Ponty's notions of the intentional arc and the tendency towards an optimal grip within the free-energy principle in what is called the *Skilled Intentionality Framework* (see also Rietveld and Kiverstein 2014, and Rietveld et al. forthcoming). The central tenet of active inference is that perception and action jointly minimize the discrepancy between actual and anticipated sensory input. However, as we have seen, the *goal* of active inference is not, as on the Helmholtzian account, to infer the hidden causes of the environment (at most, this is auxiliary), but rather to steer its interactions with the environment in such a way that a *robust* agent-environment system is maintained in which the agent is flourishing. There is an intricate circularity built in at the heart of the free-energy principle: only when I predict myself to be an agent acting in the world, and flourishing in my environment, does minimizing prediction-errors lead to a flourishing state. Hence, if the agent's model is a generative model of something, it is a model of the agent acting in its niche (see Friston 2011) and of how its own actions will change its exteroceptive and interoceptive sensations. I have suggested extending this circularity to include not just regulation of metabolic needs but also to incorporate attunement to the regular ways of acting (norms) of the patterned practices the agent participates in. For example, the way an agent responds to an outstretched hand has, arguably, no bearing on her viability conditions (as long as physical harm is avoided), but refusing to shake someone's hand might be seen as a violation of a social norm or as a political statement.

To return to the earlier example: the expert football player perceives more and more fine-grained possibilities for action and how they affect the unfolding of the situation. The skilled player perceives the gap between the two defenders as "for-running" in the context of a soccer game in which a teammate is advancing on the left flank. During a defensive corner, the same gap might be perceived as "for-countering" and not solicit any direct action, but might instead ready the agent to make a move.

The central problem of interest for a cognitive science studying skilled action is, I believe, that of context-sensitive selective openness to only the relevant action possibilities. Cognitive science needs to explain how selective sensitivity to relevant affordances is shaped by context and previous experience in a way that realizes grip on the situation. Next, I will continue to argue that active inference, understood in the proper way, is the right kind of framework for such a kind of cognitive science.

I've argued that what the agent is "modeling" in a concrete situation is not so much the causal structure of the environment, but rather the relevant action possibilities that bring the agent closer to a self-generated optimum. Brain dynamics self-organize as to enact an action-oriented relevance-centered perspective on the world. When responded to this action-oriented perspective leads to interactions with the world that in turn lead to a new perspective in which other aspects of the environment stand out as relevant and so forth. This is the circularity at the heart of both the free-energy principle and of skilled action. What the agent needs to be modeling then is not the relation between sensory stimulation and the causal structure of the environment *per se*, but rather the relation between sensory stimulation and its ways of living/flourishing in an ecological niche with a particular action-related structure. The generative model of the agent is thus shaped by previous experience resulting in more and more subtle refinements to the context-sensitive relevance of available affordances.

This interpretation of active inference has a number of distinct features. First, it conceptually blurs the distinction between epistemic and purposive actions. Tending towards an optimal grip includes both running in a gap between defenders as well as looking to whether an anticipated pass is coming or not. There is no clear demarcation between the two. Second, it puts both perception and action in the service of tending towards a (partly) self-generated optimum and provides conceptual grounds for explaining where this optimum comes from. Rather than appealing to the need for truthful representations or the need for homeostasis, I appeal, in the case of humans, to the normative character of the socio-cultural practices in which the agent participates. It takes a skilled and enculturated agent to be sensitive to the relevant affordances of playing football. Last, and perhaps most importantly, it provides an account of intentionality without presupposing goals or intentions as unexplained explainers. Instead it tries to understand intentionality in terms of the agent's history of interactions with the environment based on a concern to improve grip (Bruineberg and Rietveld 2014). What is relevant is not calculated based on our inferred representation of the outside world and a desire or an intention, but rather directly shows up in the way a skilled agent perceives the world.

One might point out here that replacing an appeal to internal goals by the tendency towards an optimal grip merely shifts the problem of purposive action. With the notions of "skill" and "practice" I might just presuppose the goal-directedness that internal goals typically account for. I think this shift is warranted for two reasons. First of all, it breaks the problem of purposive action up in two parts: the purposiveness of the practice and the ability of the individual to more or less adequately take part in that practice, this leads to a different explanandum. Second, and more importantly for this paper, active inference, at least for humans, *requires* intentional practices for the acquisition of priors. The reason for this is intimately related to the 'crooked scientist argument': my priors need to be of an optimal world, not the actual world. As Friston notes:

> One straightforward way to acquire priors—over state transitions—is to marinate an agent in the statistics of an optimal world, as illustrated in (Friston et al. 2009) One might ask where these worlds come from. The answer is that they are created by teachers, parents and conspecifics. In robotics and engineering, the equivalent learning requires the agent to be shown how to perform a task. (Friston et al. 2012, p 524-525)

In other words, the developing infant is engaging with specific practices carefully set up so as to teach the infant the relevant aspects of its environment. This process of "education of attention" (Gibson 1979, p.254) shapes the individual's selective openness to affordances in a way specific to its form of life. Unfortunately, the theme of learning of optimistic priors is currently underdeveloped in the active inference literature, but cultural learning and participating in '*regimes of shared attention*' (Ramstead et al. 2016) seem to hold the key to acquiring the right expectations. At any rate, what will *not* be sufficient is for an individual to learn the statistics of its actual environment, since the actual environment misses the optimality that active inference requires.

First and foremost, I hope to have shown that the Helmholtzian, the Ashbyian and the enactive-affordance based account are each very different interpretations of active inference. Unlike the Helmholtzian and the Ashbyian framework, the Merleau-Pontyian framework is able to frontload the relevance problem in active inference. This is not to say that active inference solves the relevance problem, it should rather be a central problem to those studying active inference. Furthermore, active inference tacitly assumes the agent to be endowed with optimistic priors. This promotes the idea of the developing active inference agent as an apprentice rather than a scientist.

## 4 Sense of Agency and Predictive-Processing

In the previous section, I have introduced different accounts of agency under the free-energy principle. In this section, I will discuss the notion of the *sense of agency*. Phenomenologically, the *sense of agency* is understood as the feeling of being the cause of one's actions; the feeling that accompanies intentional and agentive voluntary actions. This feeling might be present when I take a step forward, but not when I am being pushed forward (Gallagher 2000). My starting point, in this section, will be an early, and interesting, proposal by Hohwy (Hohwy 2007) to map the sense of agency onto predictive processes. Hohwy provides a clear functional role for the self in agency and bodily movement:

> An individual needs to be able to generate and intimately track motor commands in accordance with her desires and beliefs about the world. There must be a distinction available between changes in her body and in the environment that are due to her own agency and those changes that are due to other factors in the environment or her sensorimotor system. (Hohwy 2007, p.2)

In other words, first, the agent needs to track whether an intention to act in the world actually has the desired result, and, second, distinguish between sensations following from its own movement and from other causes. The latter is explained by predicting the sensory consequences of a self-initiated movement and comparing them with actual sensory (reafferent) feedback. In expected situations, the error-signal that will be passed on in the model will be precise, which leads to attenuation of reafferent feedback thereby giving rise to what Hohwy calls a 'sense of mineness' of the movement. In a sense, we are 'at home' in the movement because we can precisely predict the sensory consequences of the movement. In contrast, we can't in the same way precisely predict the sensory consequences of other people's movements as well as our own. Hence the sensory consequences arising from the other's movements are not attenuated and so we don't experience the same feeling of mineness. According to Hohwy, the feeling of mineness *colors* our experiences in such a way as to enable us to perceive "one's body as a locus of mental causation", and to understand "where the mind ends and where the world begins" (Hohwy 2007, p.2). In other words, the feeling of mineness is necessary to make sense of ourselves as agents acting in a world that makes sense.

A similar explanation might be constructed for a sense of self in case of perception. Perceptual inference depends on the disambiguation of self-caused and other-caused sensory stimulation: when moving around, I need to as it were "subtract" the influence of my own movements from percepts to be able to infer the state of the environment. Perceptual mineness is experienced when we are able

to predict what we perceive: when we are able to understand the changes in the persistent, external world. Susan Hurley (Hurley 1998) (based on Gallistel 1980) provides a contrast class in which a man with paralyzed eye muscles tries to look to the right. While the eye does not move (the pattern on the retina stays the same), for the man the world appears to move to the right. The anticipated change in sensory input creates an experience in which the world seems to rotate in the direction of the anticipated glance.

There is an interesting link here between "perceptual mineness" and knowledge of so-called sensorimotor contingencies (O'Regan and Noë 2001). If I am able to anticipate how my percepts are to change if I moved in a particular direction, I will gain both a sense of "perceptual mineness" in which I am "at home" in the situation, but also a sense of "perceptual presence". Hohwy is thus completely right to state that: "as you gain the world you gain a sense of self" (Hohwy 2007, p.7).

What, I have argued, is distinctive of the enactive-affordance based account developed in the previous section, is that it provides a skill-based account of intentionality without presupposing internally represented goals and intentions. That is to say, for a skilled agent the relevant solicitations show up in perception. This is importantly different from accounts of the sense of agency that start from attenuation of reafferent feedback. Using such models, the account of the sense of agency starts with a precise counterfactual hypothesis ("I have a cup of coffee in my hand") and the temporary attenuation of actual sensory input ("my hand is resting on the keyboard"). This then triggers the body to change the world so as to make the sensory input fulfill the counterfactual prediction ("I have a cup of coffee in my hand"). A sense of agency arises when the sensory input changes in the way I anticipate. However, this approach presupposes the adequate generation of counterfactual predictions, which, in the PP framework take over the role of intentions. This is a commonly used strategy in the motor control literature: "[w]ill or intentions are external input parameters similar to task parameters" (Latash 1996, p. 302; quoted from Dotov and Chemero 2014), but it is a problem for any theory that wishes to give an exhaustive and complete account of the workings of the brain and our minds. I take it that both predictive-coding and the free-energy principle have these ambitions.

A related distinctive feature of the enactive-affordance based account compared to the rationalist and cybernetic accounts is its emphasis on subjectivity. As Thompson writes in *Mind in Life* (Thompson 2007, p. 81): "[N]aturalism cannot explain matter, life and mind, as long as explanation means purging nature of subjectivity and then trying to reconstitute subjectivity out of nature thus purged." Making skilled intentionality basic to our account implies highlighting the perspective and the concerns of the individual. On the Helmholtzian account all purposiveness is reducible to tending towards a truthful representation of the structure environment (or left external to the theory). On the cybernetic account, purposiveness is accounted for only to the extent that it pertains to the control of homeostatic essential variables. On the ecological-enactive account there is no such unifying account of purpose. Although, on this account, agency is understood in terms of tending towards grip on the situation, what actually counts as the optimum that the agent tends towards in acting, is generated by the system itself and is a function of the agent's history of interactions with the environment, embodied in the agent's generative model.

## 4.1 The Primacy of the 'I Can'

In this section, I wish to highlight an aspect of the sense of self that is, arguably, more basic than the sense of agency from the last section. The phenomenon that I am after is quite simple: when I pick up a cup, I do not experience my fingers, but I experience the cup. Still, in the experience of the cup my body is not totally transparent to me. My body is not given to me as an object, but rather it is the subject of my experience. Similarly, when I perceive the solicitation of the coffee, I experience the coffee through my bodily capabilities (e.g. the ability to drink from a mug). This sense of bodily self works

at the level of motor intentionality or skill, i.e. as intentional activities involving our bodily, situational understanding of space and spatial features. As Gallese and Sinigaglia note:

> [T]he bodily self has to be primarily and originally construed in terms of motor potentiality for actions, inasmuch the nature and the range of such potentiality define the nature and the range of pre-reflective bodily self-awareness (Gallese and Sinigaglia 2010, p. 753).

Their claim is that I pre-reflectively experience my body while grasping the cup, not as an arm, not as a hand, but as a bodily power for action. The horizon of action possibilities that the agent encounters (a field of relevant affordances), structured according to the demands of the situation and the agent's abilities, coincides with a coherent self as a bodily power for action. Importantly, for Merleau-Pontyians the relation between the horizon of possibilities and the coherent self is already intentional through and through. The primary sense of engaging with the world is in a bodily and skillful way, or as Merleau-Ponty, inspired by Husserl , famously states: "Consciousness is in the first place not a matter of 'I think that' but of 'I can'" (Merleau-Ponty 1945/1962, p. 137).

So, how does this conception of the self as a bodily power for action relate to active inference and FEP? We have seen in the section on the main tenets of the free-energy principle that the starting point for the free-energy principle is biological self-organization, formalized in terms of the minimization of surprisal. As such, the reliance on action is not accidental, but constitutive for the being of the agent:

> I model myself as embodied in my environment and harvest sensory evidence for that model. If I am what I model, then confirmatory evidence will be available. If I am not, then I will experience things that are incompatible with my (hypothetical) existence. And, after a short period, will cease to exist in my present form (Friston 2011, p. 117)

If we take this view seriously then the animal *needs* to expect itself to be a coherent agent acting in the world. Constitutive of self-organization, and basic to the free-energy principle, is an agent with the capacity to selectively interact with its environment to fulfill metabolic needs (Schrödinger 1944). In order to be a free-energy minimizing agent, then, an agent *needs* (in a constitutive sense) to expect itself as having the capacity to selectively act on its environment to fulfill metabolic needs. This expectation is not available to consciousness as a belief or hypothesis, but is rather embedded in the structure of the agent's generative model. The consequence of this is, I believe, that I encounter myself in the first place not in introspection, but in the way the world shows up to me as relevant: in the solicitations I encounter.

If we assume that the generative model constitutes the agent's perspective on its environment, then the free-energy principle dictates that this perspective is structured in a particular way. The agent needs to be able to act on the world and it needs to be able to act in ways that improve the agent's relationship to its environment. If there is phenomenal component to active inference (this might depend on the kind of animal, but see Bruineberg et al. 2016, on FEP and the mind-life continuity thesis), it will include something like "I can move to improve". The world-side of this phenomenological structure might consist of solicitations that stand out as relevant, pointing towards an improved grip on the environment. The animal-side might very well be captured by Gallese and Sinigaglia's notion of a coherent self as a bodily power for action.

I take it that this self-structure precedes any account of the self in terms of action-monitoring or comparing of intentions, for such accounts already presuppose the very ability to act. The free-energy principle requires an account of the self that captures its reliance on actions that are aimed to improve the condition of the organism in its environment.

What I hope to have shown is that the self in active inference is not accessible as an explicit belief or encountered as a thing, but shows up in the way the agent is drawn to improve its grip on the situa-

tion. This is required by FEP since only if the agent is a model of its econiche will the agent be able to maintain itself as the kind of being it is.

## 5 Conclusion

In this paper, I have investigated three perspectives on "predictivism": the Helmholtzian, the Ashbyian and the enactive affordance-based account on active inference. What is exciting about the paradigm of "predictivism" is its attempt to unify a plurality of cognitive concepts such as value and reward to a common currency: priors, prediction and precision. However, this by no means establishes the truth of the brain as analogous to a scientific hypothesis-tester. In particular, there is a tension between accounts that stress self-organization and metabolic needs and those that stress hypothesis-testing. I have argued that if the brain is thought of as a scientist, it needs to be a crooked scientist (contra the Helmholtzian interpretation). The Ashbyian account is better situated to account for bodily needs, because it starts from homeostasis, allows for both interoception and exteroception, and their integration. Yet, the Ashbyian account has its limits as well since not all of our actions can be grounded by metabolic needs. Some of our actions even squarely oppose our metabolic needs. I take it that theorists of active inference can draw important lessons from Merleau-Ponty's philosophy. Most notably from the kind of skilled action that Merleau-Ponty calls the tendency towards an optimal grip on a situation and the extent to which an animal brings forth its own world.

I hope to have shown that even if one does not care about the merger of phenomenology and cognitive science, the Merleau-Pontyian perspective on active inference still allows one to derive a number of research questions that are not easily derived from the other accounts presented. It is specifically able to frontload the question of relevance, and how an agent is able to select its own action possibilities given its previous history of interactions with the environment. Even purely behaviorally these are important questions that need to be highlighted in research on active inference.

It might seem odd to integrate neuroscience and phenomenology in the way attempted in this paper[5]. For one, the phenomenological tradition is often taken to be at odds with naturalism. For the moment, it suffices to understand this approach as taking inspiration from Merleau-Ponty, without claiming to actually be completely in line with his philosophy. A more radical thesis, to be defended in a future paper, is that if Merleau-Ponty were to be alive today, he would be a philosopher of complex systems theory.

Another point that requires some expansion is the connection between neuroscience and phenomenology. Much of contemporary phenomenology-friendly neuroscience takes a phenomenon of interest (such as the 'sense of self') and then tries to find the neural correlates of that phenomenon. The current paper stands in a rather different tradition: it attempts to develop a coherent and encompassing conceptual framework for skilled action including its neuroscientific, phenomenological and normative components. The comparison between the Helmholtzian, the cybernetic and the enactive affordance-based accounts layed out in this paper is not only about which account best fits the data, or providing knock-down arguments against one or the other, but about which one provides the most plausible, coherent and encompassing interpretation of active inference. The importance of such a framework is not just philosophical, but can have important practical ramifications. Consider one last time the case of standing under a too hot shower. To the modeler the option is always open to introduce an ad-hoc hyperprior that introduces an expectation that drives the agent away from the shower. The aim of a conceptual framework, like the Skilled Intentionality Framework, is to provide the right intuitions and theoretically justify choices made in modeling. We can understand moving away from the shower only if we think that active inference is about tending towards the most flourishing state of the animal-environment system, rather than the most likely causal structure of the environment per se.

5 Thanks to an anonymous reviewer for pressing me on this point.

The framework presented in this paper allows one to draw parallels between phenomenological structures and structures as they follow from theoretical biology. One of the great insights of Merleau-Ponty is that, as skilled humans, we have a prereflective bodily engagement with the world. Based on our concern of having grip on the environment, we are selectively sensitive to only particular solicitations that, when responded to, lead towards grip on the situation. As the skilled agent perceives solicitations in the environment, it experiences itself as a bodily power for action: self and lived world evolve together. A similar structure follows from the free-energy principle: only by enacting its own viability conditions by expecting particular sensory information and acting to bring about the sensory information it expects, does the agent guarantee its own continued existence, and flourishing in its environment. The agent *needs* to model itself as an active agent with the capacity to selectively interact with its environment. This bodily self as power for action, this 'I can', has priority over any other account of the sense of agency, such as action-monitoring. For, unlike the others, this one does not presuppose intentions. It is in itself able to ground a specific kind of intentionality, which I have elsewhere labelled "Skilled Intentionality".

## References


Apps, M. A. J. & Tsakiris, M. (2014). The free-energy self: A predictive coding account of self-recognition. *Neuroscience & Biobehavioral Reviews, 41*, 85–97.

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

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

Auletta, G. (2013). Information and metabolism in bacterial chemotaxis. *Entropy, 15* (1), 311–326.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science, 22* (3), 295–318.

Bitzer, S., Bruineberg, J. & Kiebel, S. J. (2015). A Bayesian attractor model for perceptual decision making. *PLoS Comput Biol, 11* (8), e1004442.

Bruineberg, J. & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience, 8*, 599.

Bruineberg, J., Kiverstein, J. & Rietveld, E. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*.

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. *Behavioral and Brain Sciences, 36* (03), 181–204.

——— (2015). Radical predictive processing. *The Southern Journal of Philosophy, 53* (S1), 3–27.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

Dayan, P. & Hinton, G. E. (1996). Varieties of Helmholtz machine. *Neural Networks, 9* (8), 1385–1403.

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4* (4), 429–452.

Dotov, D. & Chemero, A. (2014). Breaking the perception-action cycle: Experimental phenomenology of nonsense and its implications for theories of perception and movement science. In M. Cappuccio & T. Froese (Eds.) *Enactive cognition at the edge of sense-making* (pp. 37–60). Basingstoke, UK: Palgrave Macmillan.

Downey, A. (2017). Radical sensorimotor enactivism & predictive processing. Providing a conceptual framework for the scientific study of conscious perception. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Dreyfus, H. L. (2002). Intelligence without representation–Merleau-Ponty's critique of mental representation the relevance of phenomenology to scientific explanation. *Phenomenology and the Cognitive Sciences, 1* (4), 367–383.

Dreyfus, H. & Kelly, S. D. (2007). Heterophenomenology: Heavy-handed sleight-of-hand. *Phenomenology and the Cognitive Sciences, 6* (1-2), 45–55.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology.* MIT Press.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Comput Biol, 4* (11), e1000211.

——— (2011). Embodied inference: Or I think therefore I am, if I am what I think. In W. Tschacher & C. Bergomi

(Eds.) *The implications of embodiment (cognition and communication)* (pp. 89–125). Exeter, UK: Imprint Academic.

Friston, K. J. & Stephan, K. E. (2007). Free-energy and the brain. *Synthese, 159* (3), 417–458.

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

Friston, K. J., Daunizeau, J. & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PloS One, 4* (7), e6421.

Friston, K., Adams, R., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3*, 151.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T. & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience, 6* (4), 187–214.

Gallagher, S. (2000). Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences, 4* (1), 14–21.

Gallese, V. & Sinigaglia, C. (2010). The bodily self as power for action. *Neuropsychologia, 48* (3), 746–755.

Gallistel, C. R. (1980). *The organization of action: A new synthesis.* Hillsdale, NJ: Erlbaum.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

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

Hohwy, J. (2007). The sense of self in the phenomenology of agency and perception. *Psyche, 13* (1), 1–20.

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

——(2016). The self-evidencing brain. *Noûs, 50* (2), 259–285. https://dx.doi.org/10.1111/nous.12062.

Hohwy, J., Paton, B. & Palmer, C. (2016). Distrusting the present. *Phenomenology and the Cognitive Sciences, 15* (3), 315–335. https://dx.doi.org/10.1007/s11097-015-9439-6.

Hurley, S. L. (1998). *Consciousness in action.* Cambridge, MA: Harvard University Press.

Kripke, S. A. (1982). *Wittgenstein on rules and private language: An elementary exposition.* Cambridge, MA: Harvard University Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Latash, M. L. (1996). The Bernstein problem: How does the central nervous system make its choices. In M. L. Latash & M. T. Turvey (Eds.) *Dexterity and its development* (pp. 277–303). New Jersey: Lawrence Erlbaum Associates.

Limanowski, J. & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neuroscience, 7*, 547. https://dx.doi.org/10.3389/fnhum.2013.00547. http://journal.frontiersin.org/article/10.3389/fnhum.2013.00547.

McGregor, S., Baltieri, M. & Buckley, C. L. (2015). A minimal active inference agent. *arXiv preprint arXiv:1503.04187*.

Merleau-Ponty, M. (1942/1966). *The structure of behavior.* Boston: Beacon Press.

——(1945/1962). *Phenomenology of perception.* London, UK: Routledge.

Metzinger, T. (2017). The problem of mental action. Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing.* Frankfurt am Main: MIND Group.

O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24* (05), 939–973.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Francisco: Morgan Kaufmann.

Ramstead, M. J. D., Veissiére, S. P. L. & Kirmayer, L. J. (2016). Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention. *Frontiers in Psychology, 7*.

Rietveld, E. & Kiverstein, J. (2014). A rich landscape of affordances. *Ecological Psychology, 26* (4), 325–352.

Rietveld, E., Denys, D. & Van Westen, M. (forthcoming). Ecological-enactive cognition as engaging with a field of relevant affordances: The skilled intentionality framework (SIF). In A. Newen, L. de Bruin & S. Gallagher (Eds.) *Oxford handbook of 4E cognition.* Oxford University Press.

Schrödinger, E. (1944). *What is life? With Mind and Matter and Autobiographical Sketches.* Cambridge, MA: Cambridge University Press.

Seth, A. K. (2015a). Inference to the best prediction. In T. K. Metzinger & J. M. Windt (Eds.) *Open mind.* Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570986. http://open-mind.net/papers/inference-to-the-best-prediction.

—— (2015b). The cybernetic Bayesian brain. In T. K. Metzinger & J. M. Windt (Eds.) *Open mind.* Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570108.

Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology, 2*.

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind.* Cambridge, MA: Harvard University Press.

Tschacher, W. & Haken, H. (2007). Intentionality in non-equilibrium systems? The functional aspects of self-organized pattern formation. *New Ideas in Psychology, 25* (1), 1–15.

Turvey, M. T. & Carello, C. (2012). On intelligence from first principles: Guidelines for inquiry into the hypothesis of physical intelligence (PI). *Ecological Psychology, 24* (1), 3–32.

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

Varela, F., Thompson, E. & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience.* Cambridge, MA, London, UK: The MIT Press.

Von Helmholtz, H. (1860/1962). *Handbuch der physiologischen Optik.* New York, NY: Dover.

Wiese, W. & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Wittgenstein, L. (1953). *Philosophical investigations.* Oxford, UK: Blackwell.

# Sleep and Dreaming in the Predictive Processing Framework

Alessio Bucci & Matteo Grasso

Sleep and dreaming are important daily phenomena that are receiving growing attention from both the scientific and the philosophical communities. The increasingly popular predictive brain framework within cognitive science aims to give a full account of all aspects of cognition. The aim of this paper is to critically assess the theoretical advantages of Predictive Processing (PP, as proposed by Clark 2013, Clark 2016; and Hohwy 2013) in defining sleep and dreaming.
After a brief introduction, we overview the state of the art at the intersection between dream research and PP (with particular reference to Hobson and Friston 2012; Hobson et al. 2014). In the following sections we focus on two theoretically promising aspects of the research program.
First, we consider the explanations of phenomenal consciousness during sleep (i.e. dreaming) and how it arises from the neural work of the brain. PP provides a good picture of the peculiarity of dreaming but it can't fully address the problem of how consciousness comes to be in the first place. We propose that Integrated Information Theory (IIT) (Oizumi et al. 2014; Tononi et al. 2016) is a good candidate for this role and we will show its advantages and points of contact with PP.
After introducing IIT, we deal with the evolutionary function of sleeping and dreaming. We illustrate that PP fits with contemporary researches on the important adaptive function of sleep and we discuss why IIT can account for sleep mentation (i.e. dreaming) in evolutionary terms (Albantakis et al. 2014).
In the final section, we discuss two future avenues for dream research that can fruitfully adopt the perspective offered by PP: (a) the role of bodily predictions in the constitution of the sleeping brain activity and the dreaming experience, and (b) the precise role of the difference stages of sleep (REM (Rapid eye movement), NREM (Non-rapid eye movement)) in the constitution and refinement of the predictive machinery.

**Keywords**

Bayesian brain | Consciousness | Dreaming | Embodiment | Evolution of sleep and dreaming | Hard problem | Integrated information theory | Predictive processing | Sleep | Synaptic pruning

**Acknowledgements**

The authors wish to thank Jennifer Windt and two anonymous reviewers for the precious criticisms and suggestions and Joe Dewhurst, Marco Viola, Benjamin Singer and Emma Hemmings for useful comments on an early draft of this paper. Special thanks also go to Thomas Metzinger, Wanja Wiese and the participants of the MIND23 workshop in Frankfurt for the invaluable feedback and critical discussion of our work.

## 1 Introduction

Dreaming is a fundamental aspect of our mental activity. We spend almost one-third of our life sleeping, and a good portion of that time dreaming. We often report dreaming experiences upon awakening, and they can have a huge impact on our daily life. Dreaming has also been an object of philosophical investigations, representing a conundrum for our theories on the nature of reality and for the accuracy and reliability of perception.

In spite of this, not so much has been written about dreaming in either the philosophical or the scientific literature until the 20th century. Since the discovery of REM sleep in the early 1950s (Aserinsky and Kleitman 1953), there has been in fact an increasing interest for the topic in psychology and philosophy. The research on REM sleep has sparked a proliferation of theories about dreaming, although it is now fairly established that dreams can happen at any stage of sleep (Nielsen 2000).

On the scientific side, authors are divided in their theoretical proposals on the topic, in particular regarding the *explananda* of their research. A first group of theories concerns the neural underpinning of dreaming: such as the famous AIM model (Hobson et al. 2000), the neuropsychoanalysis of dreaming (Solms 2000), and more cognitive-functional approaches (Domhoff 2001). A second, although not entirely distinct, set of theories concerns the functional role of sleeping and dreaming:

Bucci A. & Grasso M. (2017). Sleep and Dreaming in the Predictive Processing Framework.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 6. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573079

from the early proposal of a "reverse learning" theory (Crick and Mitchison 1983) to the more updated "synaptic pruning" hypothesis (Tononi and Cirelli 2014), and the complementary idea of sleep as memory consolidation (Stickgold et al. 2001; Perogamvros and Schwartz 2012). Many researchers are also concerned with finding the specific evolutionary role of dreaming, rather than of sleep in general. Interesting proposals have been formulated over the last decades, such as the "threat simulation theory" (Revonsuo 2000; Valli and Revonsuo 2009) and its more recent version "social simulation theory" (Revonsuo et al. 2015). Generally speaking though, the idea of an evolutionary advantage of dreaming per se has been received with scepticism (Flanagan 1995; Flanagan 2000). Finally, another approach is focused on the developmental and cognitive aspects of dreaming: the analysis of dream reports highlighted a progressive enrichment in structure and length of children's dreams (Foulkes 1999) as well as continuities between dream content and wakeful activities (Foulkes 1985; Domhoff 2011a).

This plethora of proposals highlights the difficulty in formulating a cohesive theory of dreaming. This is reflected by the early philosophical scepticism in regards to the topic (famously expressed by authors such as Malcolm 1959 and Dennett 1976). A recent comprehensive analysis of the field of dream research (Windt and Noreika 2011) pointed out the so-called "integration problem: the problem of how to integrate dreaming into broader theories of consciousness" (Windt and Noreika 2011, p. 1091) and, more generally speaking, of cognition.

Taking the integration problem as a springboard, we aim to give an account of dreaming through the lens of Predictive Processing (as proposed by Clark 2013, Clark 2016; Hohwy 2013) in order to provide a more comprehensive alternative to the present theoretical fragmentation. The paper will provide an overview of how sleep and dreaming are explained within the framework, focusing on pertaining issues and their possible solutions.

In order to do that, in the first section we will briefly recapitulate the main tenets of Predictive Processing and illustrate the state of the art of this approach in dream research. We will give a definition of dreaming in Predictive Processing terms and highlight the main theoretical advantages offered by the framework.

In the second section, we will illustrate these theoretical advantages while tackling a weak spot in Predictive Processing: the explanation of phenomenal consciousness, with specific reference to the hard problem of consciousness (as formulated by Chalmers 1996). We will introduce Integrated Information Theory (hereafter IIT, Oizumi et al. 2014; Tononi et al. 2016) as an example of a theoretical proposal that deals with the correlations between the phenomenal aspects of dreaming and the neural work of the sleeping brain. We will show that Predictive Processing and Integrated Information Theory share important analogies and theoretical points of contact, and therefore could be natural allies in providing a more detailed picture of how and why we dream. Afterwards, we will examine in more detail the evolutionary function of sleep according to Predictive Processing and Integrated Information Theory and show why the latter identifies an evolutionary role for dreaming too.

In the final section, we will focus on two topics in dream research that can be promisingly investigated within the Predictive Processing framework in future studies. First, the role of the body in regards to dream formation, which can be accounted for by the predictive architecture proposed by Predictive Processing. Second, the specific role of the different sleep stages (REM and NREM) in the refinement of the predictive machinery: in order to resolve some ambiguity presented by the scientific literature, we will propose a two-step mechanism of refinement operating during sleep.

## 2 Dreaming in the Predictive Brain

### 2.1 What is Predictive Processing?

Predictive Processing (hereafter PP; see Clark 2013, Clark 2016; Hohwy 2013) is an emerging framework in cognitive science, rooted in a vast and diverse scientific and philosophical literature (for a

summary, see Friston 2010; see also Clark 2013, pp. 181-186, for the historical antecedents). The main tenet of the framework is that brains are predictive machines with a hierarchical structure, continuously in the business of predicting their own internal states in relation to the external sensory input. This result is achieved through a combination of top-down flows of predictions and bottom-up flows of error signals. Predictions here are based on hypotheses construed on a generative model[1] of the world, which tracks by means of Bayesian inferences the worldly causes behind the sensory input (for a detailed account of Bayesian statistics in this context see Hohwy 2013, ch. 1; see also Clark 2016, pp. 301-303). This inferential process creates expectations (or priors) which guide prediction at each level of the cognitive hierarchy.

The hierarchical structure is the following: priors are organised from bottom levels (which track fast time-scale, perceptual details) to top levels (which track slow time-scale, abstract regularities) (Hohwy 2013, pp. 27-28). Predictions are streamed top-down (and laterally) and matched with the bottom-up sensory information. That first matching generates an amount of prediction error that indicates how much the current prediction differs from the input. The prediction error is then streamed upward (forward) in the architecture, repeating the matching process at each level, through the mutual informational exchange between error units and representation units (the latter being the carrier for the top-down predictions).

The goal of the whole system is to minimize the amount of prediction error (and the overall level of surprisal to the system), i.e. to generate successful predictions of its own states, ultimately corresponding to successful inferences about the world (Clark 2013, p. 186; see also Hohwy 2013, pp. 51-53).

According to PP, therefore, brains are sophisticated neural networks that rely on statistical inferences to produce the best prediction of the incoming sensory input and of their own internal states. The uncertainty of the signal from the environment (its reliability) is handled through a mechanism of assignment of expected precision to incoming input and gain regulation of the error units (Hohwy 2013, pp. 64-66; Clark 2016, pp. 53-59). In other words, when the precision of the signal coming from the sensory input is judged as low, the gain on error units is also low and the brain relies more on its previously acquired priors. In other cases in which the signal is considered more precise, the gain is high and the brain relies more on the inputs. This process, in a nutshell, recapitulates the role of attention in the hierarchical architectures described by PP: "Attention […] names the means or process by which an organism increases the gain (the weighting, hence the forward-flowing impact) on those prediction error units estimated to provide the most reliable sensory information relative to some current task, threat, or opportunity" (Clark 2016, pp. 59-60).

As stated above, the goal of the brain is to minimize the amount of prediction error generated in the system. According to PP there are two ways of achieving this goal. The first is to explain away the prediction error by deploying better predictions that fit the upcoming signal, i.e. perception. A second and complementary strategy is to modify the stream of sensory data so that it matches the predictions better. This is, in PP terms, action: actively seeking to match the predictions by interacting with the environment and sampling it through bodily movements, in order to produce or evoke the sensory consequences expected by the brain. But how is the mechanism of action implemented in the first place? In PP, proprioceptive predictions play a central role in determining actual bodily movement. Motor control, as Clark (Clark 2016, p. 121) puts it, is "subjunctive": given a prediction of a non-actual proprioceptive state, the body will move accordingly in order to minimise prediction error.

According to the framework[2] then, perception and action are two recurring and complementary strategies adopted by the brain to minimise prediction error. Their combination and cyclical succession — labelled "active inference" (Friston et al. 2011) — seem to be the very basis of our interaction

1 A generative model "[…] aims to capture the statistical structure of some set of observed inputs by inferring a causal matrix able to give rise to that very structure" (Clark 2016, p. 21).

2 Clark (Clark 2013, Clark 2016), in this regard, proposes the label "action-oriented predictive processing" to differentiate from other approaches, like the one proposed by Hohwy , which do not place as much emphasis on the mechanism of active inference.

with the world (Hohwy 2013, pp. 90-92; Clark 2016, pp. 120-124). According to the action-oriented formulation of PP the interaction with the environment is crucial in determining the specific quality and accuracy of our percepts. This becomes particularly relevant for the PP explanation of dreaming, as we shall see below.

## 2.2 The State of the Art: Neurobiology of Sleep According to Predictive Processing

Recent works by Hobson and Friston (Hobson and Friston 2012; Hobson and Friston 2014) summarise the evidence in support of a PP explanation of dreaming and the labour of the sleeping brain.

In a nutshell, the idea is that "[…] the brain is essentially doing the same thing in sleep and waking; with one key difference — there is no sensory input during sleep. However, the recurrent hierarchical message passing is still in process; with continually changing expectations and hierarchical predictions that constitute dream content." (Hobson and Friston 2014, p. 8). In other words, the very same cognitive architecture that drives the perception-action loop when awake is still active during sleep, but devoid of the important role of environmental perceptual input and motor feedback.

Hobson and Friston base their view on the combination of the famous AIM model (Hobson et al. 2000) and the free energy principle (Friston 2010)[3]. The AIM model makes use of a multidimensional state-space for keeping track of the brain's changes in activation (A), input-output gating (I) and neurochemical modulation (M). The shifts in parameters are mirrored by shifts in subjective experience. In particular, two positions in the state-space are relevant for their discussion: wake (characterised by high activation, externally driven processing and prevalently aminergic modulation) and REM sleep (characterised by high activation, internally driven processing and cholinergic modulation). The peculiar condition of the brain during REM sleep determines the formation of dreams and their sometimes bizarre and perceptually unstable narrative. During REM sleep, in fact, the change in neurochemical modulation affects the input gating so to draw the attention of the brain from the sensory periphery to internally generated activations.

In PP terms this means that the stream of bottom-up sensory information is attenuated through the assignment of low precision, and the brain has to rely mainly on internally generated predictions (based on progressively more abstract, middle-to-high level priors) to carry out the task of minimizing prediction error (Clark 2016, pp. 98-102). To complete the picture, the specific activations of the brain during REM sleep are different from waking (Hobson and Friston 2014, pp. 9-10): activations of the primary visual and non-visual sensory cortices, the thalamocortical sensory system and basal forebrain explain the perceptual-like character of dreams. Particularly relevant in this context is the lack of activation of areas of the prefrontal cortex (deputed to executive functions), which would explain the diminished meta-cognitive awareness during dreams.

The motor cortex is still active and presumably deploying motor and proprioceptive predictions to engage in active inference; however, motor commands are inhibited at the pontine level, resulting in REM atonia and an effective paralysis of the body[4]. As a consequence, proprioceptive predictions can never be fully satisfied by proprioceptive feedback (as would happen during wake, when the strategy of active inference would elicit actual movements in the environment), forcing the brain to jump from one prediction to another, determining the inconsistent nature of dream narrative. Of particular relevance in this context is the role of ponto-geniculate-occipital (PGO) waves, which originate in the brainstem and proceed all the way up to the visual cortex. Their presence is correlated with anticipation and elicitation of ocular movements in both wake and REM sleep (Hobson and Friston 2012, pp. 85-90; Hobson and Friston 2014, p. 8). During wake, PGO waves peak in response to a change in peripheral vision (increase of surprisal), bringing about new predictions which need to be matched with the sensory stream in order to minimise prediction error. This in turn drives attention towards

3 The free energy principle is a much more general theory that encompasses the PP proposal (see Clark 2016, pp. 305-306, for more details).

4 In fact, subjects affected by REM sleep behaviour disorder lack the blockade of motor output and "act" their dreams while asleep.

the (visual) sensory periphery and a motor prediction is issued from the motor cortex, which results in a saccade towards the origin of the input. This is a typical example of the perception-action loop in wake, a strategy that usually is effective in minimising prediction error. However, during REM sleep, the precision assigned to the bottom-up sensory stream of information is very low. PGO waves remain present though, as do ocular movements[5]. The system therefore has to make sense of randomly generated activations of several areas as well as the actual deployment of ocular movements to fit visual predictions without the aid of the fine-grained environmental feedback. In our understanding of PP, these conditions force the brain to rely only on its available middle-to-high level priors (which have a more abstract nature) for the formation of the perceptual scene (the *dreamscape*), since the low-level priors are flagged (via precision weighting) as unreliable. This means that the dreamscape will be populated by objects that lack the fine-grained perceptual details and depth provided from the external environment during wake. Furthermore, these objects will be more likely to present bizarre and mixed features and the stability and continuity of the perceptual scene will be partially compromised. In the next section we will analyse the phenomenal aspects of dreaming in more details; for now, we argue that PP can account for the peculiar character of the phenomenal aspects of dreaming by linking it to the differences in the neural work during sleep. However, as we shall see later, this is not sufficient to explain why we dream in the first place.

It must be noted that the picture presented by Hobson & Friston is controversial insofar as it takes into account REM stages only to explain dreaming. However, it is now widely accepted in the dream research community that dream mentation can happen during NREM stages as well as during transitional stages such as hypnagogia (onset of sleep) and hypnopompia (onset of wakefulness) (see Windt 2015, pp. 50-56 and 530-550, for a detailed critique of transitional states), although they differ from REM dreams because they tend to be more similar to static images and they lack a narrative development. Moreover, recent articles (Domhoff 2011b; Fox et al. 2013) that connect dreaming to other cognitive phenomena instantiated by the default-mode network (DMN) put pressure on the neurological description provided above. A detailed analysis of the neurological details goes beyond the purpose of this paper: for example, the peculiar differences in content and vividness between NREM and REM sleep would deserve a separate discussion that accounts for the different occurrences of PGO waves. It is worth noticing, though, that the general architecture described by PP would be compatible with a theory of dreaming formation involving more widespread brain activations, as long as it doesn't contradict the main tenet of a predictive perception-action loop[6] ongoing in the brain.

## 2.3 Dreaming in the Bayesian Brain: Theoretical Advantages of Predictive Processing

From the general description of PP and the analysis of the scientific literature explaining dreaming through the framework, we can now propose a clearer definition of what dreaming is in this context.

> **Dreaming** $=_{Df}$ A process of hypothesis testing through perception-action loops under the constrained, altered neurophysiological conditions of sleep.

The mechanism of prediction error minimization is always in place - but the conditions under which the mechanism operates are different. Therefore, it is arguable that the brain tries to instantiate the loop with the external environment, but fails to do so effectively. Without the important feedback of the external environment, the brain "runs wild" from one prediction to another, in accordance with the probability distribution among priors expressed by the generative model, while trying to make

5 Extra-ocular muscles are one of the few groups of muscles barely affected by REM atonia.

6 The presence of a perception-action loop in dreams has not been widely discussed in the literature so far. A provisional suggestion is that the dreamscape plays part of the role of the external environment, insofar as the dreamer "acts" in the dream world.

sense of the (mostly) internally generated stream of information. As Windt (Windt 2015, p. 603) nicely puts it, dreaming is "a process of mental improvisation".

The PP explanation of dreaming has a clear theoretical advantage, in response to the integration problem mentioned above, insofar as it is by definition inclusive: it encompasses all forms of cognition under the same architecture and ongoing mechanism. PP therefore blurs the line between cognitive phenomena that were traditionally conceptualised as distinct: imagination, mind-wandering, dreaming, hallucinations, standard waking perception are all generated by the same predictive engine under different circumstances. In fact, as Clark puts it, "perceivers like us, if this is correct, are inevitably potential dreamers and imaginers too" (Clark 2012, p. 764). PP poses a necessary link between a set of potential cognitive phenomena that will arise from the specific Bayesian hierarchical architecture described above, distinguishing itself as a particularly parsimonious framework, while at the same time retaining the ability to explain the specific character of each of those phenomena.

In the next section, we will examine two more theoretical advantages of PP. The first is that it provides a clear insight on the correlations between neural states and subjective experience in dreaming. This ties in directly with the explanation of consciousness. Why do we dream in the first place? We will introduce Integrated Information Theory (Oizumi et al. 2014; Tononi et al. 2016) to address this question, showing the benefits of a comparison with PP. Secondly, PP is also a good framework in which to understand the evolutionary role of sleep and dreaming. To show this, we will provide an explanation of the advantageous mechanism of generative model optimisation operated during sleep that is compatible with the latest empirical evidence offered by sleep research. With the aid of IIT, we will also tackle the issue whether dreaming has an evolutionary role *per se*.

## 3 The Phenomenal Character of Dreaming

### 3.1 What Does it Feel like to Be Dreaming?

As seen above, during REM sleep the brain works under different conditions compared to wake. These conditions determine the resulting phenomenal character. The selective deactivation of large portions of the prefrontal cortex and the subsequent diminished meta-cognitive awareness contribute to the immersive nature of the dreaming experience. This picture led researchers to compare dreaming to a form of intensified mind-wandering (Fox et al. 2013, pp. 10-11). Indeed, in both cases the attention is drawn from the external input to internally generated stimuli, creating a form of (partial) seclusion from the environment. In PP terms, we may construe this as the assignment of low precision to the stream of information coming from the sensory periphery. It makes sense, phenomenally speaking, to compare mind-wandering to dreaming: while our mind strays from the present tasks, we feel immersed in our own thoughts to the point that we lose contact with our surroundings and only direct, sudden or life-threatening stimuli bring us "back to reality". During REM sleep, the physiological changes conspire to seclude us even more, raising the threshold for external stimuli to pass into the system.

This is far, however, from stating that the sleeping brain is totally disconnected from its environment. There are many cases in the literature that report integration of external stimuli into dreams. A study on the effect of somatosensory stimulation on dreaming (Sauvageau et al. 1998) provides an interesting example. For the study, the participants were monitored in a sleep lab and the stimulation was administered through the inflation of a blood pressure cuff fitted above the knees. Here is the dream report of one of the subjects upon awakening: "I was in our school gym bleachers. I decided to go join some gymnasts on the floor. It was really crowded with people; I've never seen so many. I was making my way through the crowd all out of breath and there was this big woman with a scarf. *The scarf got hooked on my leg and I couldn't get it off*. I could feel it there; it didn't hurt, but it bothered me that I couldn't take it off." (Sauvageau et al. 1998, p. 11, italics in original). From the PP perspec-

tive, this would be a clear case of deployment of new predictions (the scarf wrapped around the leg) to match a stimulus that has passed the high threshold imposed by sleep (the pressure applied by the cuff). However, the functioning of the predictive hierarchy is still disrupted, from which follows the (wrong) attribution of the stimulus to the scarf. Please note that, although the origin of the stimulus is somehow explained away by the sleeping brain, it can be argued that a certain degree of error is still present, hence the emotional reaction of the dreamer (she is bothered that she could not take off the scarf).

The occasional dream oddities resulting from the integration of external stimuli are a case of a more general phenomenal feature: bizarreness (Scarone et al. 2008; Noreika et al. 2010). Although this feature is probably not as frequent as Hobson's AIM model (which compares dreaming to a form of psychosis) would claim (see Domhoff 2007, for a critique), PP can also account for it. The brain relies on middle-to-high level priors which are more abstract in nature and it can't match the resulting hypotheses with the informationally-rich stream of external sensory input. This generates a disruption in the "binding process" of dream content (Revonsuo and Tarkko 2002), resulting in oddities such as mixed features of objects and dream characters (like dreaming of a duck-winged man), and contextual displacement (i.e. people or objects appearing where they are not supposed to in a standard waking environment). It is worth noticing, however, that the study reports the presence of bizarreness in only about 50% of the reports analysed, and of these cases "only 37% concerned the internal features of the representation itself (structure and outlook, familiarity, semantic knowledge, temporal continuity)" (Revonsuo and Tarkko 2002, p. 14), the remaining cases being related to contextual bizarreness. In PP terms this could be interpreted, all in all, as a sign of the strong reliability of the generative model upon which predictions are based, albeit in impaired conditions.

Another salient phenomenological feature of REM dreams is their narrative development (Hobson et al. 2000; Nir and Tononi 2010). The narrative structure is overall (bizarreness aside) continuous and similar to waking experience, a fact reflected by the ability to report dreams in a narrative fashion. In PP terms this could be directly linked to the way organisms construct their own ongoing subjective experience - that is by deploying active inference. The sleeping brain does not simply process scattered, random stimuli, but a constant flow of endogenous activations that it makes sense of through the aforementioned perception-action loop. It is not clear, however, how much of the cognitive architecture has to be in place and functioning to guarantee the presence of a cohesive narrative. NREM dreams are also widely reported, and although sometimes different in their content (more static, conceptual, less vivid), they can also be narrative in nature. Hypnagogic hallucination seems to be a better case of dream-like imagery without a narrative (Nir and Tononi 2010, p. 94). Studying these cases might provide a good way to pinpoint the exact neuronal circuits that implement the active inference strategy in contrast to cases in which the brain simply "stands still" on the internally generated input.

PP's explanation of dreaming has the potential to accommodate well the vast empirical literature on dream phenomenology and its neural substrate. However, there is still no indication in regards to why dreaming should be a form of conscious experience at all.

### 3.2 Why Doesn't PP fully Explain Consciousness?

Since the formulation of the "hard-problem" of consciousness (Nagel 1974; Chalmers 1996), the question of "why it feels like this, or like anything at all, to be something" has puzzled philosophers and scientists alike. In particular, Chalmers has proposed and discussed various arguments against physicalism, either casting doubts on the (nomological, logical, or metaphysical) supervenience of phenomenal properties on physical properties, or arguing against their identity with them, given the impossibility of deducing all truths about phenomenal facts from the complete knowledge of truths about physical facts (Chalmers 1996). The problem concerning phenomenal consciousness becomes even more evident in the case of dreaming: while it might make some sense, at least evolutionarily

speaking, to be able to have subjective experiences during wake in order to better cope with the environment, it is not immediately evident why we should experience the internally generated world of dreams instead of simply shutting down (phenomenally speaking) for a few hours per night.

Tentative solutions to the hard problem have been advanced in the PP literature. Hobson & Friston (Hobson and Friston 2014) and Hobson et al. (Hobson et al. 2014) equate consciousness to a form of active inference, while appealing to a "Cartesian theatre" metaphor to account for the connection between the labour of the brain and the subjective phenomenology. It is not clear, however, in what sense their proposal would protect the PP framework from the classical zombie objection (a system might have all the functional and behavioural properties of a conscious system, but no internal subjective experience)[7].

Hohwy proposes that "conscious perception is determined by the hypotheses about the world that best predicts input and thereby gets the highest posterior probability." However, he continues, "[…] this is not intended as a proposal that can explain why perceptual states are phenomenally conscious rather than not." (Hohwy 2013, pp. 201-202). In other words, Hohwy aims to explain precisely which specific representational content (among the many predictions elaborated by the brain) generates a subjective phenomenal experience, how the latter is generated and under which conditions. Interestingly, he later proposes to connect this to the Global Neuronal Workspace theory developed by Baars and Dehaene (Hohwy 2013, pp. 211-214). This proposal seems to suggest that only a small part of the information processed by the hierarchical architecture flows into consciousness, or in other words that consciousness is composed of/emerges from a series of subpersonal, subconscious processes.

On the same line, Clark suggests that Predictive Processing might be on the right track to begin to solve the hard problem (Clark 2016, p. 239). By explaining all the components of conscious experience, such as the sense of self, the sense of presence, agency, emotions, as well as the perceptual milieu and other cognitive features (imagination, dreaming and the like) under the same predictive architecture, PP is in a sense dissolving the hard problem piece by piece. However, Clark notes: "True believers in the hard problem will say that all we can make progress with using these new-fangled resources is the familiar project of explaining patterns of response and judgment, and not the very existence of experience itself." (Clark 2016, p. 324). This PP solution would work only for detractors of the hard problem (Dennett 2013). We want to remain agnostic on this point, and propose an alternative approach that tries to face the hard problem directly: Integration Information Theory (IIT).

### 3.3 Integrated Information Theory: What it Is

IIT (Tononi 2008; Tononi 2012; Oizumi et al. 2014; Tononi and Koch 2015; Tononi 2015; Tononi et al. 2016) attempts to account for consciousness by linking the phenomenological evidence we get from our own experience to the evidence provided in the last decades by cognitive neuroscience. In fact, IIT maintains that in order to solve the hard problem of consciousness a change of perspective is needed: "[A]s long as one starts from the brain and asks how it could possibly give rise to experience [...] the problem may be not only hard, but almost impossible to solve. But things may be less hard if one takes the opposite approach: start from consciousness itself, by identifying its essential properties, and then ask what kinds of physical mechanisms could possibly account for them. This is the approach taken by integrated information theory (IIT)" (Tononi and Koch 2015, p. 5).

The theory starts by identifying the properties of conscious experience (which are described as "axioms") and establishes connections with properties of the physical system that supports them ("postulates"). Each axiom about phenomenal experience has a corresponding postulate about the physical substrate (Oizumi et al. 2014; Tononi et al. 2016). The first axiom is *intrinsic existence*: experience exists, it is actual, undeniable, self-evident, and intrinsic, namely independent of external observers or

7 For a critical discussion of these ideas, see Dołęga & Dewhurst (Dołega and Dewhurst 2015) and response from Hobson & Friston (Hobson and Friston 2016).

objects. This axiom constitutes the starting point of IIT and corresponds to the Cartesian assumption that conscious experience is a given, a self-evident and indubitable truth. The corresponding postulate claims that the system supporting such experience must exist "intrinsically", namely must have cause-effect power upon itself[8].

The second axiom, *composition*, claims that the structure of experience is composed of multiple (higher-order) "phenomenological" distinctions, namely different aspects of each individual experience (such as the perception of various objects seen in the visual field, their shape, parts, colour, extension, etc.). The corresponding postulate claims that the system must be composed of sets of elements with cause-effect power within the system, forming a structure of mechanisms of different order that corresponds to the structure of phenomenal experience.

The third axiom, *information*, states that conscious experience is specific and that every conscious state is informative, inasmuch as it identifies specific sets of phenomenological distinctions and differentiates from (rules out) other possible experiences. The corresponding postulate states that the system must specify a cause-effect structure, roughly the repertoire of activation states of the mechanisms that compose the system, which characterizes the cause-effect profile of such states and differentiates it from other possible ones.

The *integration* axiom claims that conscious experience is unified and irreducible to its component phenomenological distinctions (i.e. experience comes as a whole, the experience of a blue book is not reducible to the experience of a colourless book plus the experience of the colour blue, nor they can be experienced separately). The postulate states that the cause-effect structure specified by the system must be unified and intrinsically irreducible to the one specified by non-interdependent sub-systems obtained by unidirectional partitions (namely a partition that is unable to affect and be affected by the activity of other parts of the system). The degree of intrinsic irreducibility is measured as integrated information (Φ), which quantifies the changes in the cause-effect structure of a system when the system is partitioned.

Finally, the *exclusion* axiom claims that consciousness has a definite spatio-temporal grain, and that there is only one conscious subject at a time (which cannot have parts or be part of a bigger subject). The corresponding postulate claims that the cause-effect structure of a system must be definite: it is specified over the set of elements that is maximally irreducible from its intrinsic perspective, hence having the highest level of integrated information ($\Phi^{MAX}$) among all its sub-systems or the systems that comprise it.

Starting from these axioms and postulates, IIT describes the phenomenology of consciousness as constituted by the sum of informational relationships among activation states of the system. The cause-effect repertoire, or *qualia space*, is defined as a high-dimensional space with one axis for each possible past and future state of the system in which a structure of *concepts* can be represented (Tononi 2012). Qualia are sets of informational relationships in high Φ-level generating systems, or maximally irreducible cause-effect repertoires, called "concepts", generated by a complex of elements. Φ here represents the measure for intrinsic irreducibility called *integrated information*. Φ also quantifies how the cause-effect structure changes when the system is partitioned. In fact, the greatest role is played by what Tononi calls maximally irreducible conceptual structures (MICS). A conceptual structure is a constellation of points in concept space, where each axis is a possible past/future state of the set of elements, and each point is a concept specifying differences that make a difference within the set.

Besides axioms and postulates, IIT posits a central identity: every experience is identical with a conceptual structure that is maximally irreducible intrinsically, namely with a MICS (Tononi et al. 2016). In particular, the "quality" of the experience - its content of phenomenal distinctions - is spec-

8 Causal power is a condition for existence, and cause-effect power upon itself is a condition for the existence of a system independent of external observers. Tononi writes: "in order to exist, [a system] must have cause-effect power, as there is no point in assuming that something exists if nothing can make a difference to it, or if it cannot make a difference to anything. Moreover, to exist from its own intrinsic perspective, independent of external observers, a system of elements in a state must have cause-effect power upon itself, independent of extrinsic factors." (Tononi 2015, p. 4164).

ified by the form of the conceptual structure (by the concepts and their relationship in cause-effect space), whereas the "quantity" of the experience - its level - is given by its irreducibility ($\Phi^{MAX}$), i.e. the quantity of integrated information of the MICS. IIT therefore posits an identity between integrated information in a system and conscious phenomenal experience. This equals to say that, if a system satisfies the requirements for having a non-null-quantity of Φ, it will be conscious by definition.

### 3.4 Integrated Information Theory: How it Helps

IIT makes a number of predictions concerning the neural processes fundamental for consciousness and their impairment in pathological and altered (i.e. non-standard waking) conditions (Casali et al. 2013). In particular, perturbational studies using TMS (Transcranial magnetic stimulation) as a method of breaking down cortical connectivity and reactivity (and therefore, the level of $\Phi^{MAX}$ in the brain) have shown that in REM sleep (Massimini et al. 2009; Massimini et al. 2010) and some instances of NREM sleep (Nieminen et al. 2016) the level of integrated information in the brain is high enough for exhibiting conscious experience (dreams, reported upon awakening). On the contrary, extensive cortical connectivity breakdowns (as the ones happening during slow-wave sleep and anaesthesia) impair the system to the point where no global conscious experience is possible. What this means is that waking-like consciousness does not appear in such cases, but few sub-complexes could still lead to limited phenomenal experience[9]. This fits nicely with PP insofar as only certain areas of the brain might be sufficient in order to produce conscious experience, but at the same time the variations in the stream of information available to minimise prediction error determines variations in phenomenal experience, due to reliance on different priors. It also accounts for NREM dreams in regards to the different phenomenal content described above: more sparse activations during NREM stages would result in segregated information, but the local maxima might still be sufficient for static, non-narrative perceptual-like mentation.

Moreover, IIT states that the brain is organized like a device for "interpreting" spatial and temporal correlations representative of the causal structure of the environment, in the light of its memories (stored in connections). This causal structure is incorporated in the connectivity of the system via natural selection and learning mechanisms. IIT calls *matching* the measure of how well the integrated conceptual structure generated by an adapted complex fits or "matches" the cause-effect structure of its environment (Oizumi et al. 2014; Tononi 2012). In other words, IIT tries to account for the adaptation of the cognitive structures and behaviour of biological organisms (and artificial systems) to the environment. This is reminiscent of the PP description of the role of the generative model (what IIT calls the "integrated conceptual structure"): the process of increasing the matching value can be seen as equivalent to the refinement of the generative model[10]. This is achieved by a progressive reorganisation of priors (concepts or structure of concepts, in IIT terminology) via learning and, as we shall see below, sleep.

Given the assumption of the identity between consciousness and integrated information, IIT predicts that conscious experience must correspond to the high levels of Φ measured during REM sleep (Massimini et al. 2010) and NREM sleep (Nieminen et al. 2016), we suggest that PP might benefit

9 This view implies that when the level of Φ of the whole system becomes lower than the level of Φ of some of its parts, the main conscious subject ceases to exist, but other subjects (corresponding to the new sub-complexes with maxima of Φ) come into existence. The issue concerning this consequence of the theory deserves deeper analysis, which goes beyond the scope of this paper, but its discussion should nonetheless be of top priority if IIT were to be a complete theory of consciousness.

10 As suggested by Metzinger (personal communication), there is a long-standing debate on the possibility that adaptive mechanisms will lead to self-deception rather than an accurate modelling of the world. It is therefore debatable to what extent the mechanisms described by both PP and IIT will lead to a good mapping of the causal structure of the world. A full discussion of the topic goes beyond the purpose of this paper; it will suffice to say that the convergence of descriptions of mechanisms between PP and IIT remains valid independently from the actual product of these mechanisms (whether it would be accurate or adaptive but self-deceptive). For further reading on the topic, see Von Hippel and Trivers 2011; for a discussion of self-deception in the light of PP, see Pliushch 2017.

from assuming a similar non-eliminativist position in regards to phenomenal aspects of consciousness and the hard problem.

It could be argued that IIT has several problems on its own. There is no space here for a thorough critique of the theory, but a relevant objection in this context is that it does not really solve the hard problem. In fact, the solution proposed by the theory comes from the assumption of the fundamental identity between consciousness and integrated information (more precisely, maximally irreducible conceptual structures). This solution constitutes a third way of responding to the hard problem, different from both classic dualism and reductive/eliminativist physicalism. As seen above, different versions of PP have different degrees of resistance to the hard problem and consequently different degrees of compatibility with IIT. The analysis offered in this paper justifies the provisional assumption that a connection between the two theories is feasible and potentially fruitful. Despite a more detailed analysis is needed, the similarities in the conceptual vocabulary used by IIT and PP allow for a direct comparison between the two theories.

In the next section, we will turn to another theoretical advantage of PP, that is the explanation of the evolutionary role of sleep. With the support of IIT, we will later argue for an evolutionary role of dreaming as well.

## 4 The Evolutionary Role of Sleep and Dreaming

### 4.1 The "Synaptic Homoeostasis Hypothesis"

The evolutionary role of sleep and dreaming has been an object of discussion among the scientific community for a long time. It is not evidently clear why an organism should, for at least some time a day, stop doing evolutionarily advantageous activities (like foraging for food) and become potentially vulnerable to predators. However, there is an increasing amount of evidence that points in the direction of a cognitive benefit deriving from sleep.

A particularly relevant proposal (independent from but compatible with IIT) has been advanced by Tononi & Cirelli (Tononi and Cirelli 2014) in a recent review of the state of the art in neurobiology: the "synaptic homoeostasis hypotheses" (SHY). SHY claims that sleep is a way to improve overall synaptic organisation and restore energetic equilibrium (homoeostasis) in the brain. During wake, our brains constantly form new synaptic pathways and neuronal connections, strengthening them in response to the stimuli from the environment. However, the brain's resources are limited and neural plasticity takes its toll: over time the energetic expense for synaptic maintenance, combined to synaptic saturation and decreased signal-to-noise ratio will become disadvantageous. This explains the presence of sleep: according to SHY, during slow wave sleep (i.e. NREM stages 3 and 4) the combination of synaptic depotentiation (triggered by cholinergic neuromodulation) and spontaneous activation throughout the brain contributes to an overall downgrading and optimisation of the synapses and, as a consequence, the restoration of synaptic homoeostasis. This process is called "activity-dependent down-selection" (Tononi and Cirelli 2014, p. 15). During wake, the brain searches for potential statistical regularities in the environment (repetitive and suspicious coincidence in the sensory input) and potentiates synapses accordingly. The stronger the regularities, the stronger the synaptic bonds. However, noise too could drive synaptic formation, thus leading to maladaptive connections and over-fitting of the model in the long run, since the waking brain will continue to form new synaptic connections in order to grasp all the possible regularities in the environment. During slow wave sleep, the brain operates a general connective depotentiation through low frequency diffused activations in neuromodulatory conditions that promote synaptic depression (Tononi and Cirelli 2014, p. 19). The strongest connections (i.e. the one representing the strongest regularities learned while awake) will be less affected, resulting therefore in a relative reinforcement and subsequent increase in signal-to-noise ratio. To give a simplified example of this, think of two different sets of connections, the first (representing strong regularities)

with value *n* and the second (representing weaker ones) with value *n-2*. If the downgrading operates on a -3 factor, over time the weaker connections will be deleted, leaving only the stronger (more adaptive) ones in place and more room for new connections (learning) upon awakening.

This process contribute to a better consolidation of previous useful information while integrating it in long-time learning schemes, and systematically protecting it from noise interference (by forgetting noise-related connections). Finally, the "gist extraction" of important regularities from the environment is ultimately improved (Tononi and Cirelli 2014, pp. 21-23).

### 4.2 Over-Fitting Avoidance and Optimisation of the Generative Model

The mechanism described by SHY bears striking similarities to the PP account of the evolutionary role of sleep. Hobson & Friston (Hobson and Friston 2012), in response to the apparently non-adaptive loss of thermoregulation during REM stages, propose that this is nonetheless a necessary step in order to reduce the general complexity of the generative model and avoid over-fitting. This idea has roots in the "wake-sleep algorithm" proposed by Hinton (Hinton et al. 1995). The mechanism is simple: during wake the system learns new things (optimising posterior beliefs), but over time its generative model of the world becomes overly-complex, incapable of distinguishing between meaningful signal and noise. A second phase is therefore needed: "During sleep, the brain's model is insulated from further sensory testing but can still be improved by simplification and streamlining. […] Sleep may thus allow the brain to engage in synaptic pruning so as to improve (make more powerful and generalizable) […] the generative model" (Clark 2016, p. 101). This "synaptic pruning" (i.e. removal of redundant, weak connections) is precisely what SHY describes as happening during NREM sleep.

We will discuss the different roles of REM and NREM stages of sleep in the optimisation of the generative model in the next section. For now, let's turn back to dreaming.

### 4.3 Are Dreams Adaptive?

We showed that PP provides an elegant and rich explanation for the adaptive role of sleeping. But what about sleep mentation (like REM dreaming or hypnagogic imagery)? Does it have an evolutionary role or is it just a "spandrel of sleep" (Flanagan 1995)? In the light of the discussion above, there are two main reasons to believe that dreaming has an adaptive value.

First, provided that assuming a convergence between PP and IIT is fruitful, the identity postulated between integrated information and consciousness suffices to establish that sleep mentation is a form of conscious experience, insofar as the adaptive neural mechanism ongoing during sleep implies high levels of Φ. More generally, it seems that in this context the question about an evolutionary role of dreaming per se is ill-conceived: given that the level of integrated information in a system depends on the structure of the system itself, a structure that has a clear evolutionary advantage in sleep, it makes poor sense to distinguish between the two. The question of what is the evolutionary role of dreaming (if there is any) is then linked to the more general question of what is the evolutionary role of consciousness.

Second, and in support to the previous point, "IIT predicts that adaptation to an environment should lead to an increase in matching and thereby to an increase in consciousness." (Tononi 2015). This means that an increase in matching, or in PP terms the optimisation of the generative model, is strictly connected to a wider and richer phenomenal repertoire (a higher value of Φ). Studies on adaptive logic-gate networks, or animats (Albantakis et al. 2014), have shown that over the course of their adaptation, integrated information increases with respect to task fitness and matches the complexity of the environment. Although this form of task-dependent fitness was tested exclusively online, it is plausible that in much more complicated neural networks like our brains the sleep phase contributes to the same ultimate result: increase in matching. Given that dreaming is the inevitable phenomenal aspect of this procedure (for during both REM and NREM dreams the level of Φ is high enough, see

§3.4), it makes no sense to ask for an evolutionary role disconnected from the neural processes that give raise to it.

In sections 3 and 4 we have shown in depth what we think are two important theoretical advantages of adopting PP in the study of dreaming, discussing the possible issues arising. In the next section we will outline two open problems in dream research and we will suggest that they can be fruitfully re-examined through the lens of Predictive Processing,

## 5 Two Avenues of Research for the Future

### 5.1 The Dreaming Body: Is Dreaming an Exclusively Internal/Off-Line Simulation?

Dreaming has often been referred to in the literature as a form of cranial-bound, off-line, disembodied experience. Windt (Windt 2015, pp. 350-354) neatly sums up this position as the "functional-disembodiment hypothesis", according to which the sleeping-body inputs and motor outputs are disconnected from the phenomenal experience presented in dreams. This hypothesis could be interpreted as assimilating dreaming either to a radical form of cranial envatment or to a particularly vivid case of imagination. Indeed, the very depiction of dreaming by Hobson & Friston presented in the first section of this paper seems to fall in line with it. This position, although widespread, has limited scope since it does not account for phenomena like the integration of external stimuli and, more generally, the role of the complex dynamics in the body while asleep.

However, we want to suggest that this position does not represent PP in its entirety. In fact, a growing body of literature within the framework seems to indicate that bodily predictions are almost always in place. As suggested by Seth (Seth 2013), interoceptive predictions (i.e. predictions about the states of our own internal organs, muscular and visceral sensations, hunger, pain, breathing, etc.) constitute the basis for our sense of conscious presence in the world and the building blocks for the formation of a rich, embodied conscious experience. Arguably, proprioceptive predictions (i.e. predictions about the position of our body in space) might play a role too, in particular for what concerns vestibular positioning and motion (Dharani 2005). Even emotions, explained as arising from a combination of interoceptive predictions (as described above) and cognitive ones (how we are "supposed to feel" about our own body reactions) (Clark 2016, pp. 231-235), might play a role inasmuch as they constitute a conspicuous part of dreaming phenomenology.

Bodily predictions, according to PP, are therefore an important part of the cognitive hierarchy and the study of their neural instantiation should provide a better insight of why dream phenomenology has a "diminished" embodied flavour. Interestingly, a recent study (Windt et al. 2014) used lucid dreams as a condition to explore the subjective difference between self-tickling (usually ineffective) and being tickled by another (usually intense). During wake condition, this pattern of subjective difference was respected. The results in the lucid dreaming condition indicate however that a form of sensory attenuation is in place: in spite of the dreamer "commanding" other dream characters to tickle her, the subjective feeling is almost indistinguishable from self-tickling in intensity. Interpreted through the lens of PP (Clark 2016, pp. 112-114), the impossibility of self-tickling is due to the inclusion of motor commands issued for the tickling itself into predictions of the sensory output, thus leading to a general sensory attenuation. The results of the lucid dreaming experiment indicate that bodily predictions are still in place during sleep. However, "a strong conviction driving these effects in lucid dreams might be that to the extent that one is able to control an agent, this agent cannot be fully distinct from oneself" (Windt et al. 2014, p. 7). The authors here refer explicitly to the presence of hyperpriors that guide body-ownership attribution — and those hyperpriors, in the case of dreaming, would be among the ones leading predictions, in the absence of reliable sensory input.

Following this, we do not want to suggest that dreams are fully embodied experience, but rather support the proposal that "dreams are weakly functionally embodied states" (Windt 2015, p. 383): bodily predictions, especially in the form of high-level priors, are still present during dreaming, therefore it would be inappropriate to think about it as a form of disembodied mentation. More generally speaking, PP helps rethinking dreaming as yet another particular state in which the cognitive architecture operates: one that shuts off the environment but only up to a certain threshold, and that doesn't cancel out the role of the body in the formation of the dreamscape. Future empirical research are needed in order to establish exactly how "weak" this functional embodiment is.

### 5.2 Synaptic Pruning or Synaptic Strengthening? The Exact Role of Sleep Stages in the Optimisation of the Generative Model

Recall from section 4.2 that according to both SHY and PP, sleep serves a role of optimisation of the generative model via global activity-dependent down-selection. However, the supporters of SHY claim that this synaptic pruning is mainly operated during slow wave sleep, while Hobson & Friston (Hobson and Friston 2012) focus their attention on REM sleep. Given the current state of the evidence provided, it is still unclear when exactly the optimisation is conducted. Interestingly, a recent controversy (Heller 2014 and response by Cirelli and Tononi 2015) puts pressure on the general picture drawn by SHY: some studies suggest that sleep does involve synaptic potentiation and strengthening as mechanisms of memory consolidation. There is indeed a broad literature about the role of both NREM and REM sleep in memory evolution, recombination and integration in mental schemes (Stickgold and Walker 2013). The reward system also seems to play an important part in this respect, explaining the highly emotional character of REM dreams (Perogamvros and Schwartz 2012; Perogamvros and Schwartz 2013).

More empirical research is needed to set the debate. Our tentative proposal is to consider the different stages as performing two different but complementary functions: NREM sleep would operate synaptic pruning, while REM sleep would strengthen the connections via randomly generated, wake-like levels of activity. This would make sense if we think that a continuous depotentiation overnight would deplete the brain of possibly useful but freshly formed (therefore, weak) synapses. Alternating NREM synaptic pruning to REM synaptic reinforcement would avoid that loss. Additionally, reinforcement during REM sleep would be conducted without the noise disturbance of the bottom-up sensory stream but only among middle-to-high level priors, leading to a better internal coherence of the generative model. The optimisation would be thus conducted via a "two-steps mechanism" that alternates overnight. Interestingly, this idea has been already proposed in the scientific literature: Giuditta and colleagues (Giuditta et al. 1995) talked about a "sequential hypothesis" (SH) of the function of sleep. In their view, mostly related to the positive effects of sleep on memory consolidation, slow-wave (NREM) sleep would serve as a preliminary mechanism of general depotentiation and simultaneous flagging of memories, while REM sleep would help to store and potentiate those important memories. We argue that SH can be easily meshed with PP. A two-steps mechanism, as suggested above, would also be compatible with SHY, as the authors themselves admit (Tononi and Cirelli 2014, p. 27). However, it must be noted that in a more recent paper Giuditta (Giuditta 2014) remarks important differences between SH and SHY on several points: the energetic needs and the nature of the activity of the brain, the methodological approaches to support the respective claims, and the role of REM stages in memory consolidation. Establishing the exact function of each stage of sleep is important, from PP's perspective, to understand their variation in length over the ontogeny and their possible role in the early development of the generative model (Segawa 1999; Hobson 2009). More research is needed to figure out the exact neurological mechanisms and processes: we suggest that the adoption of PP would be a profitable way to frame the future research in order to overcome the differences between competing hypotheses.

## 6 Conclusion

This paper aimed to provide an overview of the understanding of sleep and dreaming within the Predictive Processing framework. We described three theoretical advantages in adopting PP. First, it is an integrative, inclusive framework, insofar as it explains several cognitive phenomena under the same cognitive architecture. Second, it has a good grip on the phenomenal aspects of dreaming. Third, it provides a clear and elegant explanation of the evolutionary role of sleep. In relation to the last two aspects, we observed that PP still lacks a good answer to the hard-problem of consciousness and suggested a possible merging with Integrated Information Theory. We showed the terminological and conceptual affinities between the two theories and the solution that IIT offers in response to the hard-problem and to the question about whether dreaming has a specific evolutionary role, the strategy of starting from the phenomenal aspects of consciousness and their explanation in terms of integrated information. If our proposal proves to be theoretically robust, it might serve as a springboard for a more general theory of cognition that includes an explanation of consciousness. In the final section, we illustrated two topics for future research to focus on at the intersection of PP and dream studies. We think that they could provide mutual and helpful clarification to both fields.

## References


Albantakis, L., Hintze, A., Koch, C., Adami, C. & Tononi, G. (2014). Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Comput Biol, 10* (12), e1003966.

Aserinsky, E. & Kleitman, N. (1953). Regularly occurring periods of eye motility, and concomitant phenomena, during sleep. *Science, 118* (3062), 273–274.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G. & Massimi, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine, 5* (198), 198ra105. http://stm.sciencemag.org/content/5/198/198ra105.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory.* New York: Oxford University Press.

Cirelli, C. & Tononi, G. (2015). Sleep and synaptic homeostasis. *Sleep, 38* (1), 161.

Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind, 121* (483), 753-771. https://dx.doi.org/10.1093/mind/fzs106.

——— (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36* (03), 181–204.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

Crick, F. & Mitchison, G. (1983). The function of dream sleep. *Nature, 304* (5922), 111–114.

Dennett, D. C. (1976). Are dreams experiences? *The Philosophical Review, 85* (2), 151–171.

———(2013). Expecting ourselves to expect: The Bayesian brain as a projector. *Behavioral and Brain Sciences, 36* (03), 209–210.

Dharani, N. E. (2005). The role of vestibular system and the cerebellum in adapting to gravitoinertial, spatial orientation and postural challenges of REM sleep. *Medical Hypotheses, 65* (1), 83–89.

Domhoff, G. W. (2001). A new neurocognitive theory of dreams. *Dreaming, 11* (1), 13–33.

———(2007). Realistic simulation and bizarreness in dream content: Past findings and suggestions for future research. In D. Barrett & P. McNamara (Eds.) *The new science of dreaming: Content, recall and personality correlates* (pp. 1–28). Westport, CT: Preager Press.

———(2011a). Dreams are embodied simulations that dramatize conception and concerns: The continuity hypothesis in empirical, theoretical, and historical context. *International Journal of Dream Research, 4* (2), 50–62.

——— (2011b). The neural substrate for dreaming: Is it a subsystem of the default network? *Consciousness and Cognition, 20* (4), 1163–1174.

Dołega, K. & Dewhurst, J. (2015). Curtain call at the Cartesian theatre. *Journal of Consciousness Studies, 22* (9-10), 109–128.

Flanagan, O. (1995). Deconstructing dreams: The spandrels of sleep. *The Journal of Philosophy, 92* (1), 5–27.

——— (2000). Dreaming is not an adaptation. *Behavioral and Brain Sciences, 23* (06), 936–939.

Foulkes, D. (1985). *Dreaming: A cognitive-psychological analysis.* Mahwah, NJ: Lawrence Erlbaum Associates.

——— (1999). *Children's dreaming and the development of consciousness.* Cambridge, MA: Harvard University Press.

Fox, K. C. R., Nijeboer, S., Solomonova, E., Domhoff, G. W. & Christoff, K. (2013). Dreaming as mind wandering: Evidence from functional neuroimaging and first-person content reports. *Frontiers in Human Neuroscience, 7*, 412.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11* (2), 127–138.

Friston, K., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics, 104* (1-2), 137–160.

Giuditta, A. (2014). Sleep memory processing: The sequential hypothesis. *Frontiers in Systems Neuroscience, 8*, 219.

Giuditta, A., Ambrosini, M. V., Montagnese, P., Mandile, P., Cotugno, M., Zucconi, G. G. & Vescia, S. (1995). The sequential hypothesis of the function of sleep. *Behavioural Brain Research, 69* (1), 157–166.

Heller, C. (2014). The ups and downs of synapses during sleep and learning. *Sleep, 37* (7), 1157.

Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science, 268* (5214), 1158.

Hobson, J. A. (2009). REM sleep and dreaming: Towards a theory of protoconsciousness. *Nature Reviews Neuroscience, 10* (11), 803–813.

Hobson, J. A. & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology, 98* (1), 82–98.

——— (2014). Consciousness, dreams, and inference: The Cartesian theatre revisited. *Journal of Consciousness Studies, 21* (1-2), 6–32.

——— (2016). A response to our theatre critics. *Journal of Consciousness Studies, 23* (3-4), 245–254.

Hobson, J. A., Pace-Schott, E. F. & Stickgold, R. (2000). Dreaming and the brain: Toward a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences, 23* (06), 793–842.

Hobson, J. A., Hong, C. C.-H. & Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Frontiers in Psychology, 5*, 1133.

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

Malcolm, N. (1959). *Dreaming.* London: Routledge and Kegan Paul.

Massimini, M., Boly, M., Casali, A., Rosanova, M. & Tononi, G. (2009). A perturbational approach for evaluating the brain's capacity for consciousness. *Progress in Brain Research, 177*, 201–214.

Massimini, M., Ferrarelli, F., Murphy, M. J., Huber, R., Riedner, B. A., Casarotto, S. & Tononi, G. (2010). Cortical reactivity and effective connectivity during REM sleep in humans. *Cognitive Neuroscience, 1* (3), 176–183.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83* (4), 435–450.

Nielsen, T. A. (2000). A review of mentation in REM and NREM sleep: "covert" REM sleep as a possible reconciliation of two opposing models. *Behavioral and Brain Sciences, 23* (06), 851–866.

Nieminen, J. O., Gosseries, O., Massimini, M., Saad, E., Sheldon, A. D., Boly, M., Siclari, F., Postle, B. R. & Tononi, G. (2016). Consciousness and cortical responsiveness: A within-state study during non-rapid eye movement sleep. *Scientific Reports, 6*, 30932.

Nir, Y. & Tononi, G. (2010). Dreaming and the brain: From phenomenology to neurophysiology. *Trends in Cognitive Sciences, 14* (2), 88–100.

Noreika, V., Valli, K., Markkula, J., Seppälä, K. & Revonsuo, A. (2010). Dream bizarreness and waking thought in schizophrenia. *Psychiatry Research, 178* (3), 562–564.

Oizumi, M., Albantakis, L. & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology, 10* (5), e1003588.

Perogamvros, L. & Schwartz, S. (2012). The roles of the reward system in sleep and dreaming. *Neuroscience & Biobehavioral Reviews, 36* (8), 1934–1951.

———(2013). Sleep and emotional functions. In P. Meerlo, R. M. Benca & T. Abel (Eds.) *Sleep, neuronal plasticity and brain function* (pp. 411–431). Springer.

Pliushch, I. (2017). The overtone model of self-deception. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Revonsuo, A. (2000). The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. *Behavioral and Brain Sciences, 23* (06), 877–901.

Revonsuo, A. & Tarkko, K. (2002). Binding in dreams–The bizarreness of dream images and the unity of consciousness. *Journal of Consciousness Studies, 9* (7), 3–24.

Revonsuo, A., Tuominen, J. & Valli, K. (2015). The avatars in the machine. In T. K. Metzinger & J. M. Windt (Eds.)

*Open MIND*. Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570375.

Sauvageau, A., Nielsen, T. A. & Montplaisir, J. (1998). Effects of somatosensory stimulation on dream content in gymnasts and control participants: Evidence of vestibulomotor adaptation in REM sleep. *Dreaming, 8* (2), 125.

Scarone, S., Manzone, M. L., Gambini, O., Kantzas, I., Limosani, I., D'Agostino, A. & Hobson, J. A. (2008). The dream as a model for psychosis: An experimental approach using bizarreness as a cognitive marker. *Schizophrenia Bulletin, 34* (3), 515–522.

Segawa, M. (1999). Ontogenesis of REM sleep. In B. N. Mallick & S. Inoué (Eds.) *Rapid eye movement sleep* (pp. 39–50). New York: Marcell Dekker, Inc.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17* (11), 565–573.

Solms, M. (2000). Dreaming and REM sleep are controlled by different brain mechanisms. *Behavioral and Brain Sciences, 23* (6), 843–850.

Stickgold, R. & Walker, M. P. (2013). Sleep-dependent memory triage: Evolving generalization through selective processing. *Nature Neuroscience, 16* (2), 139–145.

Stickgold, R., Hobson, J. A., Fosse, R. & Fosse, M. (2001). Sleep, learning, and dreams: Off-line memory reprocessing. *Science, 294* (5544), 1052–1057.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin, 215* (3), 216–242.

——(2012). Integrated information theory of consciousness: An updated account. *Arch Ital Biol, 150* (2-3), 56–90.

——(2015). Integrated information theory. *Scholarpedia, 10* (1), 4164.

Tononi, G. & Cirelli, C. (2014). Sleep and the price of plasticity: From synaptic and cellular homeostasis to memory consolidation and integration. *Neuron, 81* (1), 12–34.

Tononi, G. & Koch, C. (2015). Consciousness: Here, there and everywhere? *Phil. Trans. R. Soc. B, 370* (1668), 20140167.

Tononi, G., Boly, M., Massimini, M. & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*.

Valli, K. & Revonsuo, A. (2009). The threat simulation theory in light of recent empirical evidence: A review. *The American Journal of Psychology*, 17–38.

Von Hippel, W. & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34* (01), 1–16.

Windt, J. M. (2015). *Dreaming: A conceptual framework for philosophy of mind and empirical research.* Cambridge, MA: MIT Press.

Windt, J. M. & Noreika, V. (2011). How to integrate dreaming into a general theory of consciousness—A critical review of existing positions and suggestions for future research. *Consciousness and Cognition, 20* (4), 1091–1107.

Windt, J. M., Harkness, D. & Lenggenhager, B. (2014). Tickle me, I think I might be dreaming! Sensory attenuation, self-other distinction, and predictive processing in lucid dreams. *Frontiers in Human Neuroscience, 8*, 1–11.

# Embodied Decisions and the Predictive Brain

Christopher Burr

A cognitivist account of decision-making views choice behaviour as a serial process of deliberation and commitment, which is separate from perception and action. By contrast, recent work in embodied decision-making has argued that this account is incompatible with emerging neurophysiological data. We argue that this account has significant overlap with an embodied account of predictive processing, and that both can offer mutual development for the other. However, more importantly, by demonstrating this close connection we uncover an alternative perspective on the nature of decision-making, and the mechanisms that underlie our choice behaviour. This alternative perspective allows us to respond to a challenge for predictive processing, which claims that the satisfaction of distal goal-states is underspecified. Answering this challenge requires the adoption of an embodied perspective.

**Keywords**

Action-oriented representation | Active inference | Decision-making | Distributed consensus | Embodied decisions

**Acknowledgements:**
A special thanks to Max Jones and Richard Pettigrew for their comments on earlier drafts, and to two anonymous reviewers and the editors for their invaluable feedback. I believe the paper is in a far better form as a result of their help. Thanks also to Andy Clark and Mark Miller for insightful discussions and helpful recommendations during the earliest stages of this work.

## 1 Introduction

Consider the following situation. You are busy writing a paper, which has a deadline that is fast approaching. It is important that your time is used productively to ensure that this deadline is met. You have already been writing for a couple of hours without a break and notice that your productivity is decreasing. What should you do? As it's almost lunchtime, you consider the fact that your tiredness is a product of your hunger. However, there is a chance that if you take too long a break you will be unable to regain your train of thought. Perhaps you should continue working for a bit longer, and have lunch once the word count has been reached, or maybe it's better to compromise and take a short break now to make a coffee. What about doing something else entirely? Far from it being a simple matter of choosing between a few well-delineated options, it appears that you must also determine what options are available to you. A complete account of how we make decisions, among other things, should be able to explain these challenges. The main concern of the paper is the possibility of situating such an account of the mechanisms that underlie decision-making within the framework of predictive processing (PP).

It has been claimed that the PP framework has the capacity to unify a wide range of different phenomena, such as perception, cognition and action (see Clark 2016; Hohwy 2013, for introductions). According to Hohwy, "prediction error minimization is the only principle for the activity of the brain" (Hohwy 2016, p.2). This principle can be couched in terms of statistical inference, and provides a functional ground for unification—one which Hohwy sees as resisting an embodied reading (Hohwy forthcoming). Alternatively, Clark has also claimed that PP may offer a unifying perspective on many of the brain's capacities, including perception, action, learning, inference, and cognitive control, but opts for a less neurocentric perspective (Clark 2016). Favouring the latter approach, this paper con-

Burr, C. (2017). Embodied Decisions and the Predictive Brain.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 7. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573086

cerns itself with exploring how the PP framework can contribute to our understanding of the process of *decision-making* from an embodied perspective, thus supporting Clark's view.

We also focus on a specific problem that has been raised by (Basso 2013). This is the problem of the *underspecification* of action sequences, which are initially represented by some distal goal-state. In PP, a future goal-state is essentially a higher-level prediction used as a means of enabling action through the reduction of proprioceptive prediction-error (i.e. Active Inference) (see Clark 2016; Hohwy 2013, for introductions). However, as Basso states:

> […] the future goal state created in the beginning is accurate only in some particular circumstances (i.e., when both the task and algorithm are well-defined). In most cases, people are used to facing underspecified tasks in which a future goal state cannot be employed to derive the intermediate states (Basso 2013, p. 1)

This challenge is also important for an account of decision-making, which is traditionally assumed to operate as a deliberation over representations of expected goal-states associated with the performance of some action. However, if the sequence of actions is underspecified this raises a challenge of what constrains the selection of one set of possible actions over another. For example, imagine you decide to take a break from writing and go and make lunch. There are myriad ways this initial (perhaps vaguely specified) goal-state could be satisfied, and some may be less effective than others (e.g. deciding to make a long lunch that keeps you from returning to writing for a prolonged period of time).

We will argue that recent work on embodied decision-making (Cisek and Pastor-Bernier 2014; Lepora and Pezzulo 2015), offers both an interesting perspective on the neural mechanisms that underlie decision-making, and may also be able to provide important insights for developing PP. However, we will also argue that doing so requires adopting an embodied perspective if we are to respond to the above challenge. We will argue that a neurocentric perspective is too narrow to acknowledge the many important ways that our choice behaviour is constrained by our bodies and the world.

## 2 Embodied Decisions

Consider another decision. You must choose between two routes to work where Route A takes you through a city that has a high risk of heavy traffic but is short in distance. The other route is less likely to be affected by the increased congestion, but is longer than the former. Suppose you know from previous experience that, given the time you are leaving, it is more likely that the traffic will be light, and your preference is always for the shortest time spent travelling. What should you do?

**Table 1:** A simple representation of a decision under risk.

| | Heavy Traffic (30%) | Light Traffic (70%) |
|---|---|---|
| Route A | 24 minutes | 14 minutes |
| Route B | 18 minutes | 17 minutes |

Table 1 represents a *decision under risk*, as the agent has full knowledge of the available options and probabilities attached to the relevant states. In situations like this, deciding what to do is relatively straightforward, and a number of decision rules exist as proposals for norms of rationality. For example, the *principle of maximizing expected utility* would suggest taking Route A, as the following demonstrates that it has the shortest expected duration (and therefore the greatest expected utility, assuming that utility is a negative linear transform of duration):

Expected Duration of Route A = 0.3 x 24 + 0.7 x 14 = 17
Expected Duration of Route B = 0.3 x 18 + 0.7 x 17 = 17.3

Savage famously referred to these situations as 'small worlds', where it is possible to "look before you leap", by which he meant an agent has knowledge of the states of the world and all of the options available to her (Savage 1954/1972, p. 16). Even in the case where the probabilities attached to the states are unknown (decisions under uncertainty), many decision-theoretic norms (e.g. dominance and subjective expected utility maximization) exist to help guide this process. However, unlike small worlds, the real world is not so neatly circumscribed. In contrast, everyday decisions can be viewed as 'large worlds', where agents may be unaware of all the relevant information, including an uncertainty of what options are available. This level of uncertainty is a challenge for decision theory, as the possibility of framing a genuine decision problem requires that an agent already have options to deliberate over. Even hallmarks of rationality such as Bayesianism have been criticised as inapplicable in these types of large worlds (Binmore 2008).

It may be argued that this is not really a problem for decision theory *per se*. The issue of determining options is a problem for the perceptual system to solve. Perception is faced with the task of specifying and constructing a representation of features of the environment, which can then be used as the basis for making decisions (along with abstract representations of related decision variables such as potential risk). According to this traditional view of decision theory, decision-making is seen as a prototypical cognitive task, which can be decomposed into a process of *deliberation* (i.e. calculating the values of the relevant decision variables) and *commitment* (i.e. selecting an action). Furthermore, motor behaviour is simply the manner in which a decision is reported, and can be used to reveal an agent's preferences (Sen 1971).

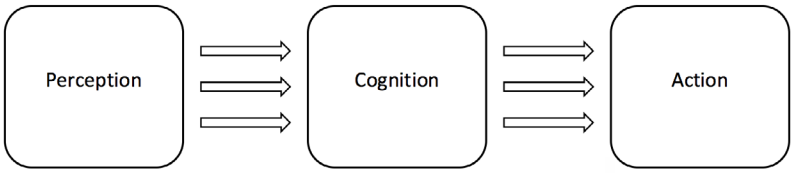


**Figure 1** - The Classical Sandwich Model

This account of decision-making is based on a number of *cognitivist* assumptions, which are nicely captured by Hurley's critique of what she termed the "classical sandwich model" of the mind (Hurley 1998, p. 401). In this model (Figure 1), the outer slices of perception and action are peripheral to the inner filling of cognition, and thus separate from one another. They are also separate from cognition, which interfaces between perception and action. First, perception builds a reconstructive representation of features of the external world. These discrete, abstract representations are then transformed by cognitive processes into a motor plan for action, according to the system's beliefs and desires, and subsequently carried out by the system's effectors. Within this model, decision-making would reside within the middle box, and deliberation and commitment could take place in some 'central executive region' such as the prefrontal cortex, which integrates relevant information from other systems such as working memory (Baddeley 1992).

Hurley saw a number of problems or limitations with this account. For example, this serial process would be insufficiently dynamic to cope with the time pressures of a constantly changing environment. In the time taken to construct a representation and plan an action by integrating the necessary information, the environment may have changed, which would render the current model (and any actions based on it) inaccurate. This worry about the urgency of performing an action in ecologically-valid scenarios is particularly pressing when applied to the case of decision-making. In traditional decision theory, models of decision-making do not incorporate the time constraints of agents, and therefore fail to account for a number of additional pressures that agents face.

The cognitivist view of decision-making then, highlighted by the classical sandwich model, leads to a tendency to think of sensorimotor control in terms of the transformation of input representations into output representations through a series of well-demarcated, encapsulated processing stages. It also often leads to the assumption that key decision variables are encoded as an abstract value in some central executive region separate from sensorimotor processes (Levy and Glimcher 2012; Padoa-Schioppa 2011). Cisek has argued that this picture is hard to reconcile with a growing body of neurophysiological data (Cisek 2007; Cisek and Kalaska 2010). He claims that key functions of decision-making, which the cognitivist would expect to be neatly delineated, in fact appear widely distributed throughout the brain, including key sensorimotor regions. To accommodate this otherwise anomalous data he proposes a notion of *embodied decisions*, which have a number of properties that are quite different to the kinds of decisions modelled by traditional decision theory (Cisek and Pastor-Bernier 2014). The following sections will explore each of them.

## 2.1 Decision-Making as a Distributed Consensus

Cisek proposes the *affordance competition hypothesis* (ACH) as a model that aims to explain both the cognitive and neural processes implicated in decision-making, and one which attempts to make sense of emerging neurophysiological data that conflicts with traditional decision theory (Cisek 2007; Cisek and Kalaska 2010). According to the ACH, decisions emerge from a distributed, probabilistic competition between multiple representations of possible actions, and importantly overlaps with sensorimotor circuits (see Figure 1 of Cisek 2007, p. 1587). To expound this view, a number of components require clarification—evidence supporting the claims are reviewed in (Cisek and Kalaska 2010; Cisek 2012).

Cisek's focus on the *distributed* manner of decision-making stands in obvious contrast to the earlier cognitivist framework, and also to other models that propose that decision-making occurs downstream of the integration of multiple sources of information, which yields a common representation of abstract value (Padoa-Schioppa 2011). Instead, according to the ACH, the sensorimotor system is continuously processing sensory information in order to *specify* the parameters of potential actions, which compete for control of behaviour as they progress through a cortical hierarchy, while at the same time, other regions of the brain provide biasing inputs in order to *select* the best action (Cisek and Kalaska 2010). These processes of *specification* and *selection* occur simultaneously and continuously, and are not localisable to a specific region. Rather, the *competition* occurs by way of mutual inhibition of neural representations, which specify the parameters of potential actions, until one suppresses the others. At such a time, a global, distributed consensus emerges.

Integral to this process is the role of continuously biasing influences (e.g. rule-based inputs from prefrontal regions, reward predictions from basal ganglia, and a range of further biasing variables from sub-cortical regions). Each of these biasing inputs contribute their votes to the selection process. As the authors state:

> […] the decision is not determined by any single central executive, but simply depends upon which regions are the first to commit to a given action strongly enough to pull the rest of the system into a 'distributed consensus'. (Cisek and Pastor-Bernier 2014, p. 4)

Again, this idea stands in stark contrast to the cognitivist picture, where the perceptual system merely processes information in order to construct a perceptual representation, which provides the evidence about the environment needed to make decisions. Rather, here we have the *beginnings* of an account that explains how the relevant options of a decision problem are being selected in parallel with the specification of sensorimotor information:

> […] although traditional psychological theories assume that selection (decision making) occurs before specification (movement planning), we consider the possibility that, at least during natural

> interactive behavior, these processes operate simultaneously and in an integrated manner. (Cisek and Kalaska 2010, p. 277)

One of the specific claims made by Cisek and Pastor-Bernier is that, as part of the competitive process, the brain is simultaneously specifying and selecting among representations of multiple *action opportunities* or affordances, which compete within the sensorimotor system itself (Cisek and Pastor-Bernier 2014).[1] These representations serve as indications of the possible actions available in the agent's environment, rather than as objective, organism-independent properties of the world.

For example, Cisek and Kalaska discuss recordings taken from the dorsal premotor cortex (PMd) in monkeys during a reaching task (Cisek and Kalaska 2010). In the experiment, monkeys were presented with two potential reaching actions by way of spatial cues, where one would later be indicated as the correct choice (using a non-spatial cue). During a memory period, where the spatial cues were removed and the future correct choice was uncertain, recorded activity in the PMd continued to specify both directions simultaneously, suggesting an anticipatory nature for the neural activity. When the information specifying the correct choice was eventually presented, activity relating to the respective action was strengthened, and the unwanted action was suppressed.

Importantly, this process occurs within the *same system* that is used to prepare and execute the movement associated with the action representations. Furthermore, Cisek and Kalaska state that the task design allowed for the monkeys to exploit a different (cognitivist) strategy, where the target locations are stored in a more general-purpose working memory buffer, distinct from motor representations, and converted to a motor plan after a decision has been made. However, though conceptually possible, the findings did not seem to support this latter view. Instead, the study seems to point to a need for representations that encode predictive (or anticipatory) action opportunities, rather than abstract representations that specify the state of the world independently from an agent's particular goals and capacities.

The ACH also makes key predictions that can be tested in future experiments. For example, it predicts that actions that are further apart from one another will show stronger mutual inhibition than those that are closer together. This is because action representations are specified in terms of spatial parameters (Cisek and Kalaska 2010), which means that a decision between similar actions (with overlapping representations) can be encoded using a weighted average. This weighted average could evolve over time, initially tolerating some uncertainty between two future actions, whereas drastically different options could not. A prediction made by the ACH is, therefore, that if one records from neural cells related to a given option, while modulating the desirability of a different option, the gain of that modulation will be strongest when the other option is most dissimilar to the one coded by the recorded cell (Cisek and Pastor-Bernier 2014, p. 5).

Importantly, this view stands in contrast to decision-theoretic accounts that model humans as making decisions between different options by integrating the relevant factors into a single variable, such as subjective utility (Levy and Glimcher 2012). For example, some have argued that the orbitofrontal cortex (OFC) and ventromedial prefrontal cortex (vmPFC) could integrate the relevant information and encode such an abstract value (Padoa-Schioppa 2011). However, this view is hard to reconcile with neurophysiological data (Cisek and Kalaska 2010). Given these findings, as well as others (cf. Klaes et al. 2011; Pastor-Bernier and Cisek 2011), it seems that at least part of a prototypical cognitive process (decision-making) is inextricably intertwined with sensorimotor control, suggesting a blurring of the boundaries between perception, action and cognition. However, as Cisek himself notes,

1 The use of the term 'affordances' will be familiar to those with an understanding of Gibson's work in ecological psychology (Gibson 1979). However, Cisek states his account is directly influenced by a wider commitment to the tradition of ecological psychology, and relies on different notions (i.e. action opportunity, affordance, pragmatic representation and sensorimotor plan) throughout his work to emphasise similar points (Cisek and Kalaska 2010). Given the controversy on how best to understand the metaphysical commitments of Gibson's notion (cf. Chemero 2003; Chemero 2011), we will focus on the less demanding notion of action opportunities instead of affordances.

"we are capable of making decisions that have nothing to do with actions, and in such situations the decision must be abstract." (Cisek 2012, p. 927) Whether we have to accept this limitation of the ACH, or whether PP can develop on its findings and scale up to include purportedly abstract decisions, is a primary focus of this article, and will be dealt with specifically in section 4.

## 2.2 Simultaneous Decisions

Consider the earlier example of attempting to write a paper, where deciding what to do was plagued by an indeterminacy in knowing whether all possible actions had been considered. Cisek and Pastor-Bernier claim these types of everyday decisions are archetypical kinds of *simultaneous* decisions. By this they mean that there are a multitude of possible actions an agent could take at any one time, and therefore, we are continuously deciding what to do.

To attempt to explain this, Cisek and Kalaska review research on the pervasive effect of attentional modulation, which supports the idea that activity in the visual system is strongly influenced by attentional modulation, even in familiar and stable environments (Cisek and Kalaska 2010). This is usually recorded as an enhancement of activity correlated with the attended regions of space, and a suppression of activity from the unattended regions. For example, studies by Stefan Treue show the ubiquitous effects of attentional modulation in primate visual cortex (Treue 2001; Treue 2003). This attentional modulation results in the enhancement of activity towards behaviorally relevant stimuli, along with a corresponding suppression of those cells tuned to non-attended spatial features. Attending only to those features of the world that are deemed salient may not appear to be a rational strategy. However, echoing the sentiments of work in ecological rationality, Treue acknowledges that it is nevertheless "an effective use of limited processing resources." (Treue 2003, p. 428)

Despite the attractiveness of appealing to saliency and attention on ecological grounds, flexible and adaptive choice behaviour requires reciprocal communication between affective and sensorimotor regions. This is because determining what is salient requires an awareness of the changing demands of both the *external* and *internal* environment, in order to respond to the homeostatic demands of the agent. At present this is one area that is left underdeveloped by Cisek, Kalaska and Pastor-Bernier. In section 3.2, we will see how this aspect of embodied decisions can be developed further, by exploring the unique roles of attention within predictive processing, and its emphasis on interoceptive inference.

## 2.3 Dynamic Choice Behaviour

Finally, Cisek and Pastor-Bernier argue that the notion of embodied decisions requires a dynamic account of the decision-making process. They claim that the continual processing of noisy or uncertain sensory information after commitment, suggests that agents continue to deliberate during the overt performance of a task. This means the agent constantly monitors the overt performance of their actions through sensory feedback (e.g. proprioception). The existence of this evidence, they argue, requires the revision of some commonly used formal models in decision theory that are unable to account for this post-selection monitoring and alteration.

Lepora and Pezzulo also acknowledge this requirement, and claim that *action performance* should be considered a proper part of a dynamic model of decision-making, rather than being understood as merely the output of the decision process (Lepora and Pezzulo 2015). As a proof of principle to support this claim, they develop a computational model, which they call the embodied choice (EC) model. This is compared against two serial evidence-accumulation models, based on the well-known drift-diffusion model[2], and their respective performance is evaluated (see Lepora and Pezzulo 2015 for details).

2 The drift-diffusion model aims to capture how a subject integrates (noisy) accumulating evidence, for multiple distinct options, in a forced choice task. The model assumes that evidence is integrated at various time steps, until some threshold is reached and a commitment is made to one of the options.

The first of these models is represented by a simple *serial process*, in which deliberation fully precedes a choice that commits the agent to the preparation and performance of the chosen action—much in the same way that the 'classical sandwich model' highlights (Hurley 1998). A *parallel model* develops this by connecting the decision process to action *preparation*. This speeds up the agent's performance by anticipating what action will be most likely given the incoming sensory evidence. As evidence in support of one option increases, the agent can begin to make preparations for the respective action, before fully committing to it. Though the latter model gains a speed increase, it does so at the expense of accuracy. It could easily turn out that evidence that initially supports one option is overshadowed by later competing evidence, leading to inaccurate or clumsy actions.

To deal with this speed-versus-accuracy trade-off, Lepora and Pezzulo develop the EC model, which, in addition to the parallel feed-forward connection, has a feedback connection that allows action dynamics (e.g. current trajectory and kinematics) to influence the decision-making process. Whereas the previous models consider decisions to be independent of ongoing action (only allowing for influence from previous experiences), embodied choices consider ongoing action as an integral part of the decision-making process, with proprioceptive signals feeding into the deliberative process to provide information about the evolving biomechanical costs of associated actions.

This is important for real-world decisions. To illustrate, Lepora and Pezzulo (Lepora and Pezzulo 2015, pp. 4-5, emphasis added) give the following example of a lion that has begun tracking a gazelle, deciding to switch and track another:

> [..] if the lion waits until its decision is complete, it risks missing an opportunity because one or both gazelles may run away. The lion faces a decision problem that is not stable but dynamic. In dynamic, real-world environments, costs and benefits cannot be completely specified in advance but are defined by various situated factors such as the relative distance between the lion and the gazelles, which change over time as a function of the geometry of the environment (e.g. a gazelle jumping over an obstacle can follow a new escape path) and the decision maker's actions (e.g. if the lion approaches one gazelle the other can escape).

They continue:

> [..] action dynamics in all their aspects (i.e. both their covert planning and their overt execution) have a backwards influence on the decision process by changing the prospects (the value and costs of the action alternatives). For example, when the lion starts tracking one of the gazelles, undoing that action can be too costly and thus the overall benefit of continuing to track the same gazelle increases. This produces a *commitment effect* to the initial choice that reflects both the situated nature of the choice and the cognitive effort required for changing mind at later stages of the decision.

A couple of comments are necessary. First, being receptive to ongoing action means the EC model can consider the evolving biomechanical costs that are salient to the current decision. Although the serial and parallel models can incorporate action costs as well, they must do so *a priori*, as there is no way for the ongoing action to feedback into the deliberative process. Critics may argue that part of the developmental process for any organism is learning about the body, and associated biomechanical costs, which are not going to change that drastically, given the limited number of states that the body can be in. Therefore, prior knowledge of biomechanical costs can be incorporated through learning. This is surely correct, but is also incomplete. As the gazelle example should highlight, biomechanial costs are also partly dependant on the evolving state of the environment, and where other agents are involved, will be difficult to precisely evaluate in advance. Second, commitment effects make it harder to change your mind once an action is performed, because the later sensory information must outweigh the initial commitment that arises from having started an action. Situated agents that are

receptive to subjective commitment effects may gain an important adaptive advantage, especially if the agent is able to learn about them for future interactions (see section 5).

As well as dovetailing nicely with the embodied decisions account, Lepora and Pezzulo found their EC model to perform better in terms of speed and accuracy than the alternative models. Initially, the models were evaluated in two simulation studies representing a two-alternative forced choice task (see Lepora and Pezzulo 2015, for details), which on its own stands as an interesting proof-of-principle. However, they also compared their models with empirical evidence from human studies, and found that the EC model was a good fit with human behaviour.

Taken together, the aforementioned properties of embodied decisions stand in contrast to the cognitivist assumptions of traditional decision theory. To reiterate, the cognitivist perspective on decision-making is strictly separated from evidence accumulation in perceptual systems, and the control of action in motor systems. However, embodied decisions view deliberation as a continuous competitive process within sensorimotor circuits, modulated by relevant biases from cortical and sub-cortical regions as well as from ongoing action, and from which a distributed consensus emerges. It is hard to maintain the traditional functional separation of perception, cognition and action if we are to appreciate this process fully.

A number of issues remain. First, although there is mention of 'continuously biasing influences' in the embodied decisions research, there is little explicit mention of the role of affective signals in the aforementioned work. This is of vital importance, as an agent should have some way of determining which action opportunities it cares about most. Second, some important questions remain about whether the concept of embodied decsions can scale up and accommodate more explicit, goal-directed decision-making. Before addressing these issues directly, we will explore how predictive processing shares many of the same motivations as embodied decisions. By doing so, we hope to uncover where the two frameworks can offer mutual development.

## 3 Decision-Making in the Predictive Brain

Like embodied decision-making, an embodied account of predictive processing eschews the idea that perception is a passive accumulation of evidence with the purpose of reconstructing some detailed inner model of the world (Clark in press).[3] Instead, perception is in the service of guiding actions that keep the organism within homeostatic bounds and maintain a stable grip upon its environment (Clark 2015; Friston et al. 2010). In this manner, Clark sees the PP story as a contemporary expression of the *active vision* framework (Ballard 1991; Churchland et al. 1994), and argues that many of the latter's motivations are also present in an *action-oriented* account of predictive processing. He emphasises the following role for prediction error minimisation (PEM):

> [...] it is the guidance of world-engaging action, not the production of 'accurate' internal representations, that is the real purpose of the prediction error minimizing routine itself. (Clark 2016, p. 168)

This leads to a different relation between the key notions of *perceptual inference* and *active inference* from other proponents of the PP framework (e.g. Hohwy 2013). These terms refer to the two ways that

3 The exact manner in which PP should accommodate a notion of embodiment is still an open question, and is unlikely to be resolved without greater clarity regarding the notion of embodied cognition. One may worry that failing to resolve this issue would impede a satisfactory unification of the two ideas being considered here.
In response to this worry, we can offer two remarks. First, we echo the sentiments expressed by Shapiro (Shapiro 2011), who argues that it is acceptable to acknowledge the varied (sometimes neubulous) methodological practices that operate under the 'embodied cognition' banner, and that for the time being we should refer to embodied cognition as a research programme, rather than as a theory, to avoid the appearance of dogmatic unity. Second, the two accounts considered here both acknowledge the existence and explanatory importance of action-oriented representations, which is perhaps one of the most hotly debated areas in embodied cognition. Therefore, given this common ground, for the time being we belive it best to pursue the possible unification, while being careful to acknowledge any theoretical disagreements.

prediction-error can be minimised: either the system can update the parameters of its inner models, in order to generate new predictions about what is causing the incoming sensory data (perceptual inference), or it can keep its generative model fixed, and resample the world such that the incoming sensory data accords with the predictions (active inference).

Although both play an important role in PP, for Clark, the primary role of perceptual inference is to "prescribe action", and as such, he states, that our percepts, "are not action-neutral 'hypotheses' about the world so much as ongoing attempts to parse the world in ways apt for the engagement of that world." (Clark 2016, p. 124) This is a throroughly action-oriented account, and importantly, it is this shift in emphasis that exposes a unity between Clark's account of predictive processing and the insights of the ACH, which claims that neural processes represent action opportunities, rather than organism-independent, objective properties of the world (Clark 2016, p. 181). In addition, Clark views 'active inference' as a more-encompassing label for the combined mechanisms whereby the perceptual and motor systems cooperate in a dynamic and reciprocal manner to reduce prediction-error by exploiting the two strategies highlighted above.[4] Active inference is accomplished using a combination of perceptual and motor systems rather than being confined to the latter that are traditionally associated with action. This view is also supported by recent neuroanatomical evidence that suggests a close relationship in the functional anatomy of the sensorimotor systems (Adams et al. 2013; Shipp et al. 2013). As Friston et al. argue:

> The primary motor cortex is no more or less a motor cortical area than striate (visual) cortex. The only difference between the motor cortex and visual cortex is that one predicts retinotopic input while the other predicts proprioceptive input from the motor plant. (Friston et al. 2011, p. 138)

This supports one of the core claims of PP, which states that action is accounted for by a downwards cascade of predictive signals through motor cortex, which elicit motor activity in much the same way as predictions descend through perceptual hierarchies. In short, motor control is just more prediction, albeit about proprioceptive signals (see chapter 4 of Clark 2016). Rather than updating the generative model in response to error signals, these control-states predict subjunctive sensory trajectories that *would* ensue *were* the agent performing some desired action. It is this development of the active inference framework that allows PP to provide an account of choice behavior.

In PP, choices are made between competing higher-level predictions about expected sensory states. The formal basis for this perspective is based on the *free-energy principle* (Friston 2010). In (Friston et al. 2014) this is extended to account for decision-making in terms of hierarchical active inference. Friston describes choices as 'beliefs about alternative policies', where a policy is defined as a *control sequence* (i.e. sensory expectations associated with a sequence of descending proprioceptive predictions) that determines which action is selected next. Policies are selected under the prior belief that they minimise the prediction error between attainable and desired outcomes across multiple, hierarchichally-nested levels. Importantly, these policies are selected on the basis of a belief in both their expected outcome, and also their expected precision (see section 3.2). Recently, Pezzulo and Cisek (Pezzulo et al. 2016) have argued that the ACH can be thought of in similarly hierarchical terms, and argue that this is further reason for adopting a control-theoretic, or action-oriented approach to cognition.

As applied to decision-making, this line of thought bears a close resemblance to work by Daniel Wolpert , who has demonstrated the close ties between motor control and decision-making (Wolpert and Landy 2012). One key difference between the views expressed by Wolpert and PP, however, is the latter's rejection of the need for the separate, explicit representation of *cost functions*.

4 We have also argued elsewhere that it is misleading to simply equate *perceptual inference* with perception and *active inference* with action (Burr and Jones 2016). Instead, we take perception to be an active exploration of the environment, involving a continuous (and simulataneous) unfolding of *both* perceptual inference *and* active inference. Similarly, action involves both altering the environment by changing one's bodily state, and monitoring the ongoing changes. In this manner, perception and action, understood in folk psychological terms, involve a combination of *both* perceptual *and* active inference at the level of underlying cognitive processing (see Vance 2017, for further discussion on these points).

In PP, cost functions are absorbed into the generative models harboured by the brain, and continuously updated by prediction error signals, becoming intertwined with the expectations of some policy. As these expectations will have been shaped by learning, there is already a *prior belief* about a policy's probability—a probability based on previous experience and captured by the extent to which it minimises prediction error through action (Friston et al. 2014). Some may worry that this view eliminates too much, or is too deflationary, and that the need for encoding some measure of the value associated with an outcome is necessary to explain why certain behaviors are preferred over others. Several responses can be offered to placate such worries.

Firstly, Clark explains that many working roboticists have already turned away from the explicit encoding of separate value/cost functions, arguing that they are too inflexible and biologically unrealistic due to their computational demands (Clark 2015). Instead they favour approaches that exploit the complex dynamics of embodied agents (such as the computational approach of (Lepora and Pezzulo 2015)), as they are computationally less demanding. These approaches acknowledge that the physiological constraints of an agent provide implicit means of understanding the value associated with dynamic action performance, without the need for positing additional abstract neural representations (see section 5.1).

Secondly, there are a number of debates in decision theory about whether the brain does in fact calculate value, with some arguing in favour of some abstract form of a neural 'common currency' (Levy and Glimcher 2012). However, as Vlaev et al. (Vlaev et al. 2011) argue, these views are beset with difficulties both from behavioural studies that explore contradictory, empirically-observed context effects (e.g. preference reversals and prospect relativity), as well as competing neurophysiological studies (see section 5.3). Vlaev et al. review a range of theories and models and provide the following positions to help capture these commitments:

- Value-first position: the brain computes the value of different options and simply picks the one with the highest value.
- Context-dependent value: the brain computes values, but the choice is heavily context-dependent on the set of available options.
- Comparison with value computation: the brain computes how much it values options, but only in relation to other values.
- Comparison-only: choice depends on comparisons without any computation of value.

Micro-debates exist within each of these positions. For example, is value represented on some ordinal, interval or ratio scale, and what objects are represented? Regardless of how these debates turn out, it should be clear that the *value-first* position is incompatible with the embodied decisions perspective. This is because value-first positions maintain that the value of an option is stable, and explicitly represented. We have already seen that the embodied decisions account is opposed to such a view, due to conflicting neurophysiological evidence. In addition, we have seen how PP eschews the explicit representation of value/cost functions altogether. However, it is unclear which of the alternative positions would best describe an embodied account of PP.

As a possible solution, den Ouden et al. (den Ouden et al. 2012), review the neurophysiological evidence relevant to an understanding of prediction errors, and argue that there is support for multiple kinds of prediction errors (PEs) in the brain: perceptual PEs, cognitive PEs, and motivational PEs. The first two types are referred to as *unsigned* PEs. These do not reflect the valence of any sensory input, but simply the surprise of its occurrence. The final kind of PEs, however, are known as signed PEs, for they reflect whether an outcome was better or worse than expected. They state:

> Signed PEs play a central role in many computational models of reinforcement learning. These models describe how an agent learns the value of actions and stimuli in a complex environment,

> and signed PEs that contain information about the direction in which the prediction was wrong, serve as a teaching signal that allows for updating of the value of the current action or stimulus. (den Ouden et al. 2012, p. 4)

Having access to multiple kinds of PEs, including those with affective significance, may provide the brain with the means to implicitly *compare* and *evaluate* which policy is most desirable based on prior learning. If value is determined indirectly through the comparison of multiple PEs, this would allow the agent to assess which of the myriad possible action opportunities is most salient given its current needs. The comparison could take the form of a distributed competition, in line with the proposal offered by Cisek, with no need for an abstract encoding of value that is generated downstream of sensorimotor processing (Cisek 2012).

In addition, adopting the suggestion of multiple PEs seems to frame PP as either an example of the `context-dependent value' view or the `comparison with value computation' view—depending on which additional mechanisms are posited to co-ordinate or integrate the options based on the type of PE considered. For example, although PP eschews talk of explicit cost functions, there is nevertheless a non-trivial sense in which the brain is comparing the expected values of the predictions that stand in place of the cost functions. Given the uncertainty regarding the precise implementational details of an exact architecture for PP (see Clark 2016, pp. 298-299, for a list of possible schemas), this could be a possibility. It is also one area where a synthesis between the work on embodied decisions and PP could be mutually beneficial—the former is presently developing novel computational methods that may help specify architectural details, whereas the latter provides a wider framework that unifies perception, cognition, action, and as we will see shortly emotion. However, there is another, possibly more radical approach that makes use of *precision-weighting* mechanisms, which may frame PP as an example of the `comparison-only' view. This approach may be able to answer the underspecification challenge posed at the start of the paper, but doing so requires bringing the body more closely within the remit of PP.

## 4 Scaling Up

One challenge is particular pressing given what we have so far discussed. The work on embodied decisions has focused primarily on exploring the neural mechanisms that underlie decision-making in simple, *visually-guided* motor tasks, such as grasping an object or pressing a button—so called habitual (or situated) decisions. Though this may be sufficient for explaining a wide variety of simple behaviors across a number of different species, humans (and some non-human animals) possess far more complex decision-making capacities. It is possible that the embodied decisions approach will be unable to account for the rich, and seemingly heterogeneous practices, that traditional decision theory tends to concern itself with.

For example, when you decide to buy a house, or choose where to go on holiday, it is not immediately obvious how a notion of embodied decisions could be of any use. Buying a house or going on holiday are both activities that require long-term planning. This implies the prolonged maintenance of a desired goal-state (e.g. buying a house), in order to coordinate and constrain relevant behaviours (e.g. acquiring a mortgage and communicating with solicitors). It is not immediately clear how the predictive brain handles the representation of distal goal-states by making solely embodied decisions of the kind hitherto discussed.

To respond to this challenge, we need to turn to a distinction often made within the decision-making literature between *deliberative* and *habitual* forms of decision-making. Competing *model-based* and *model-free* accounts are respectively put forward to try and capture the associated phenomena (Daw et al. 2011; Doll et al. 2012; Lee et al. 2014). In deliberative cases, where choice depends on the evaluation of various options, model-based methods use richly structured internal models (often rep-

resenting conditional, probabilistic relations between states of the relevant domain) to simulate future action possibilities and their associated values—such accounts are increasingly studied in neuroeconomics (Glimcher and Fehr 2014). These general methods are flexible enough to apply to a wide range of circumstances. In contrast, habitual decisions rely on previously-learned "cached" or "heuristic" strategies to choose between actions and are frequently used in reinforcement learning. Although less flexible than model-based accounts, the benefit of model-free methods as Clark acknowledges, is that they nonetheless "implement "policies" that associate actions directly with rewards, and that typically exploit simple cues and regularities while nonetheless delivering fluent, often rapid, response" (Clark 2013b, p. 5).

Initially, the model-free strategy seems incompatible with PP due to the latter's strict adherence to the existence of generative models. However, Clark argues that precision-weighting may explain how the brain flexibly switches between these two strategies on the basis of expected precision and accuracy (Clark 2013b; Clark 2016). With regards to the model-free, or heuristic strategy he claims:

> […] the use (when ecologically apt) of simple cues and quick-and-dirty heuristics is not just compatible with prediction-based probabilistic processing: it may also be actively controlled by it. (Clark 2013b, p. 8)

Here, Clark appeals to work in reinforcement learning (Daw et al. 2005; Gläscher et al. 2010) that shows how model-free strategies can implictly learn (and embody) probabilities associated with certain action sequences or policies through trial and error, without the need to retain an explicit value or construct a detailed representation. These "cached" policies can then be redeployed at a later stage by the agent if they are estimated to be more reliable than the alternatives.

Recently, a number of studies have argued that the brain decides between these two modes by employing some form of arbitration mechanism (so called "neural controllers") that predicts the respective reliability of the respective policies and chooses between them (Daw et al. 2005; Lee et al. 2014; Pezzulo et al. 2013). Clark believes these controllers could be accounted for in PP by appealing to *precision-weighting*, allowing them to be brought within the scope of the generative models (Clark 2013b; Clark 2016).

In PP, precision-weighting is considered to be a process by which the brain increases the gain on the prediction errors that are estimated to provide the most reliable sensory information, conditional on the higher-level prediction (Feldman and Friston 2010; Hohwy 2012). These precision-estimations work in close unison with higher-level predictions to provide context for the incoming prediction-error. For example, a high-level prediction carrying contextual information about the environment (i.e. at a high spatiotemporal scale) can also provide contextual information about which sensory inputs are most reliable (e.g. noisy environments mean auditory information is unreliable). It is claimed that the mechanisms behind this precision-weighting involve altering the post-synaptic gain on prediction-error units, and may also provide a way to reconcile the competing effects of signal suppression and signal enhancement (Clark 2016). If so, this would provide a more unified account for accommodating findings such as those mentioned earlier in the experiments of Cisek and Kalaska, where multiple choices are selected between on the basis of competing suppression and enhancement (Cisek and Kalaska 2010).

As well as providing a way of balancing the influence between top-down and bottom-up signals, it has also been argued that the mechanisms behind precision-weighting could provide a means of altering the brain's *effective connectivity* (i.e. the influence that one region exerts over another) (see Clark 2013b). For example, (den Ouden et al. 2010) found evidence that striatal prediction errors play a modulatory role on the large-scale coupling between distinct visuomotor regions. Additional research has also explored the context-dependent, transient changes in patterns of cooperation and competition between control systems, which result from higher-level cognitive control (Cocchi et al.

2013). Some have argued that these transiently assembled networks, are formed in response to the task demands faced by the situated agent, and that key neuromodulatory mechanisms (e.g. volume transmission) may be responsible (Anderson 2014).

This additional role for precision-weighting allows PP to accommodate the ubiquitous effects of attentional modulation that Cisek and Pastor-Bernier discuss (see chapter 6 of Clark 2016). For example, (Friston et al. 2012) investigated the neuromodulatory role of the dopaminergic system, with a specific focus on decision-making and reinforcement learning. They argue that dopamine controls the precision of incoming sensory inputs, which engender action, by balancing the respective weight of top-down and bottom-up signals during active inference. This balancing means, crucially, that the predictions that drive action, also determine the context in which the movements are made and provide a way of balancing top-down expectations with bottom-up prediction errors to flexibly manage higher-level goals (based on urgency and saliency).

Given these more global influences of precision-weighting, some have argued that the distinction between model-free and model-based methods is too coarse-grained to be usefully applied at the neural level (Gershman and Daw 2012; Clark 2016). Instead, the distinction becomes a difference in degree rather than in kind. In line with an earlier suggestion regarding the brain's maintenance of a careful balance predictions and prediction-error, Clark (Clark 2016) suggests that the same strategy should be applied to the balancing between model-free and model-based methods. As such, the former would be associated with a greater degree of bottom-up processing (i.e. driven by sensory information and sensorimotor coupling), and the latter would be associated with more knowledge-driven, top-down processing:

> The context-dependent balancing between these two sources of information, achieved by adjusting the precision-weighting of prediction error, then allows for whatever admixtures of strategy taks and circumstances dictate. (Clark 2016, p. 253)

This is an interesting claim, and one that PP appears well-equipped to handle, but it is unlikely on its own to satisfy the challenge of underspecification introduced at the start of the paper. How does the brain select, from the wide range of action opportunities, the sequence that most effectively leads to the satisfaction of some distal (possibly abstract) goal representation?

A speculative proposal offered by (Pezzulo et al. 2016) argues that different types of policies can be distinguished according to whether they are associated with *extrinsic value* (i.e. the expected physical reward for completing the action) or *epistemic value* (i.e the additional information gain or resolution of uncertainty). This distinction may be useful in explaining how offline forms of motor planning—resembling earlier theories of motor simulation (Grush 2004)—evolved progressively as elaborations on earlier sensorimotor control loops (Pezzulo 2012). These loops could then be utilised as a sort of *epistemic action*, simulating the expected precision of certain policies, and reducing the epistemic uncertainty associated with overt behaviour. In cases like this, there may be a payoff for considering actions with high epistemic value in order to ascertain whether there are other options that have not yet been considered (Friston et al. 2015). Such cases represent a sort of best-guess for the agent, based on prior knowledge of how similar situations have played out in the past. This suggestion, if valid, is likely to play an important explanatory role for more distal explanations of how the problem of underspecificatuon is resolved in biological organisms. As PP and the ACH both claim that probabilistic neural representations of action-opportunities are in continuous competition—biased by sub-cortical and cortical influences—the above suggestion seems like a worthwhile area for future investigation.

However, an embodied account can appeal to further constraints that may help to provide a more complete answer to the underspecification challenge (Basso 2013). It is to these constraints that we now turn.

## 5 Constraining Effective Choice Behaviour

Consider the decision of whether to go for a run. The distal goal-state of 'go for a run' is satisfied once you begin your workout. However, there is a series of more fine-grained causal events that exists between the time when you purportedly "decide" to go for a run and the satisfaction condition of having gone for a run. We wish to argue that the decision to go running, should include the full series of fine-grained causal microstructures—beginning with the mental representation of the goal-state considered, and ending with the overt performance of the necessary behaviour.[5] As such, the decision of whether to 'go for a run' is temporally extended over time, and as we will see, is partially constituted by events that extend beyond the brain and body.[6] This reconception of decision-making from a diachronic perspective allows us to more effectively appreciate the dynamic nature of embodied decisions, and the continuous effects of active inference. However, it requires a number of claims to be explored and defended.

The first is that goals do not exist independently of their being enacted through an organism's interactions with the world; that is, they have no independent objectivity (Gallese and Metzinger 2003). Only *goal representations* have a physical existence, realised by particular patterns of neural activity. Secondly, although we speak of goal *representations*, as we use the term, they differ from traditional notions of representation in a number of ways: a) they have no truth-conditions, only conditions for satisfaction that are directed towards the deployment of certain actions that minimise prediction error through active inference (Gallese and Metzinger 2003), and b) they are strictly grounded in facts about the agent's embodiment—although possibly multimodal at some high-level of abstraction, they are not amodal in the sense used by the cognitivist (Burr and Jones 2016).

The motivation behind (a) follows from the truth of the first claim, defended by (Gallese and Metzinger 2003 p. 371), that "no such things as goals exist in the objective order of things", therefore, "a goal representation cannot be true or false."[7] In PP, goal representations (in the form of higher-level predictions) are required by active inference, and thus have satisfaction (or fulfilment) conditions based on the imperative to minimise sensory prediction error. This leads to consideration of (b), and to the question of whether the existence of goal representations require more than can be provided by an embodied account of PP.

For example, the distal goal-state to go for a run appears to be abstract, despite being decomposable into more fine-grained sub-events (e.g. put on trainers; warm-up muscles; fill water bottle; lock door on leaving house; spend 20 minutes attempting to get your GPS (Global Positioning System) watch to detect your location). It is this abstract nature of the initial goal-state that leads to the underspecification challenge. This is because each of these multi-functional events can be considered independent of the specific goal-state—I may fill my water-bottle because I am thirsty and require a drink; I will lock my door whenever I leave my house (irrespective of whether I am going for a run). This fact regarding the multi-functionality of sub-events doesn't appear to change even when the series of sub-events is so frequently performed that I rarely deviate from the order of performance. Alternatively, another decision (e.g. whether to buy a house) may be performed so infrequently, and contain a wide diversity of sub-events, that I will have very little idea of the order of events in advance.

With the aforementioned in mind, we will show how an embodied account of PP is able to appeal to a wide-range of explanatory factors, in order to demonstrate how the neural mechanisms that underlie decision-making constrain our choice behaviour (considered here from a diachronic perspective) in important, and perhaps adaptive ways.

5 For simplicity we put no requirements on how far you travel, or how fast you run in order for the statement, "I went for a run" to obtain. This vagueness is likely to be a characteristic of many choice behaviours, and we believe that a certain flexibility is necessary to account for differences in an individual's own satisfaction conditions (e.g. less than 1km may not suffice for a professional athlete.

6 Obviously some decisions will be extended over shorter or longer period of times dependent on the framing of the decision (e.g. deciding between two sandwiches at a shop versus deciding on a new career path).

7 This point seems to fit with the subjunctive nature of higher-level predictions as employed in active inference accounts of motor control.

## 5.1 Physiological Constraints

In PP, predictions arise from generative models in the brain. These models are encoded as probability density functions, which are structured according to an increasing level of spatiotemporal scale. The predictions at the lowest levels correspond to the activity of sensory receptors encoding perturbations at small and fast spatiotemporal scales, whereas the higher-level models provide more general contextual information concerning larger and slower structures in the environment. The theoretical and empirical support for this picture has already been documented extensively in work by Karl Friston and colleagues (Friston et al. 2010), who argue that the formal similarities of their hierarchical models to the hierarchical structure of the motor system lends them biological plausibility (Kanai et al. 2015). Here, they argue, there will be a highly restricted set of possible parameters, which specify the range of possible actions, given the limited ways in which parts of the body could be configured. These parameters further restrict the set of actions, and may allow for automated or simple reflexive patterns resembling the sorts of habitual decisions we saw in the previous section. Far from being a hindrance to an agent, these restricted features can have adaptive value, allowing the agent to more easily detect and learn about the relevant features that emerge in the course of interacting with the world. This will in turn help with efficient action selection, as the specification of the relevant parameters can be reliably constrained by relevant factors of their embodiment.

For instance, as the eyes saccade from left to right, the visual scene will shift from right to left in a predictable manner, relative to the speed and direction of saccadic motion. An active perceiver can exploit regular relations between sensory input and motion of this kind in order to detect objective structural and causal features of the environment. These predictable relationships between bodily movement and sensory input are known as sensorimotor contingencies (SMCs) (O'Regan and Noe 2001), and are commonly discussed within the embodied cognition literature. We have elsewhere argued that this aspect of embodied cognition is implied by the PP framework (Burr and Jones 2016), and Seth has also proposed utilising SMCs to extend the PP framework to account for phenomena such as perceptual presence, and its absence in synaesthesia (Seth 2014).

Recent work by Cos et al. provides an interesting development to this idea (Cos et al. 2014). They argue that human subjects make a rapid prediction of biomechanical costs when selecting between actions. For instance, when deciding between actions that yield the same reward, humans show a preference to the action that requires the least effort, and are remarkably accurate at evaluating the effort of potential reaching actions as determined by the biomechanical properties of the arm. Cos et al. argue that their study (a reach decision task) supports the view that a prediction of the effort associated with respective movements is computed very quickly. Furthermore, measurements taken of cortico-spinal activity initially reflects a *competition between candidate actions*, which later change to reflect the processes of preparing to implement the winning action choice. Although there may be a possible disagreement concerning the exact manner in which cost functions are encoded and modelled, studies like this provide further reasons for taking the work of Lepora and Pezzulo (Lepora and Pezzulo 2015) seriously, due to the close connection with the aforementioned *commitment effects*, and the dynamic unfolding of decision-making.

Learning about the average biomechanical costs associated with performing certain actions could be a useful first-step in the formation of simple heuristics that stand in lieu of rational deliberation, and may also explain the presence of purportedly maladaptive decisions (e.g. sunk-cost fallacy). For example, some tasks undoubtedly require too much effort to properly deliberate over (e.g. choosing between a pair of socks), but a failure to properly identify these types of situations based on environmental markers, may lead to the misapplication of a strategy that is maladaptive in the current environment. For researchers, in cases where this strategy leads to undesired commitment effects, there may also be an opportunity to learn about the cognitive architecture of the agent in question. This is

because commitment effects reflect both biomechanical costs, as well as cognitive costs associated with changes of mind.

Being receptive to these changes in context is therefore of the upmost importance, as the value of many actions will vary contextually, dependent on factors such as fatigue, injury and environmental resistance (e.g. hill-climbing). However, rather than attempting to internalise all of the environmental variables, an alternative strategy is to simply allow the constraints of the body and environment to stand-in as a constituent part of the decision-making process. This is where dynamic, responsive feedback from the body, as input back into an ongoing decision is so important, and where work in situated cognition can provide constructive assistance. As Lepora and Pezzulo note:

> In situated cognition theories, the current movement trajectory can be considered an external memory of the ongoing decision that both biases and facilitates the underlying choice computations by offloading them onto the environment. (Lepora and Pezzulo 2015, p. 16)

This work also connects with a further topic explored in the embodied decisions literature concerning decision-making in situations of increasing urgency.

## 5.2 Urgency

Accommodating urgency exposes another important connection between PP and embodied decisions. Given the level of urgency of an agent's higher-level goal states, the gain of incoming sensory information should be adjusted accordingly. Higher-level goals should therefore encode more abstract expectations regarding the optimal amount of time taken to deliberate in any given decision. Cisek and Pastor-Bernier point to the importance of an urgency signal in their work:

> [..] in dynamically changing situations the brain is motivated to process sensory information quickly and to combine it with an urgency signal that gradually increases over time. We call this the 'urgency-gating model'. (Cisek and Pastor-Bernier 2014, p. 7)

When the urgency of a decision is low, only an option with strong evidence will win the probabilistic competition, and the agent may seek out alternative options (i.e. exploration). However, as the urgency to act increases, the competition between the options can increase, such that a small shift may be sufficient to alter the distribution. Cisek and Pastor-Bernier highlight a number of neuroimaging studies that support the existence of such an urgency signal, and argue that evidence accumulation may therefore not be the only cause of the build-up of neural activity seen during decision-making experiments.

By emphasising the importance of precision-weighting as a neuromodulatory mechanism for altering the brain's effective connectivity, PP may be able to further develop this line of thought in a more unified framework, which demonstrates the closely intertwined nature of perception, action, cognition and emotion, learning, and decision-making. This is because PEM is receptive to causes in the environment across a number of spatiotemporal scales. For example, perhaps some perturbing influence happens regularly at the order of milliseconds, but is also nested within a further perturbing influence that occurs on the timescale of minutes. The hierarchical structure of the brain is well-suited to accommodate these changes, but it is also well-suited to regulate additional factors such as the biomechanical costs involved with certain actions, which themselves may differ across spatiotemporal scales. Anyone who has done long-distance running and suffered as a result of inadequate pacing will attest to the importance of being receptive to the body's changing demands across extended timescales. This connects to a further important constraint, which has so far been overlooked—the importance of affective information originating in the body.

## 5.3 Interoceptive Inference

In PP, the predictions generated by the inner models of the brain do not merely attempt to anticipate the flow of sensory input from the outside world, but also the flow of interoceptive inputs (i.e. pertaining to endogenously produced stimuli, e.g. bodily organs). These inputs further constrain the set of viable actions in important ways. For example, deciding to quench one's thirst or sate one's hunger is often more important than allowing oneself to be distracted by alternative action opportunites. Being receptive to the current state of your body is fundamental to making adaptive decisions, as it allows us to determine which options have the greatest value *relative to our present needs*.

Tracking this type of sensory information requires incorporating *interoceptive information* into the PP framework. Seth has argued that Active Inference can be extended to accommodate interoception, and that key areas such as the anterior insular cortex (AIC) are well-suited to play a central role as both a comparator that registers top-down predictions against error signals, and as a source of anticipatory visceromotor control (i.e. the regulation of internal bodily states) (Seth 2013).

This is important for integrating decision-making within the PP framework, as it allows for a consistent understanding of the role that affective information (and possibly emotions) play in guiding our actions. Such a view, often attributed to the likes of William James, has seen a resurgence of interest, with many theories being proposed for how we are able to integrate emotions into the models of the mechanisms that underlie decision-making (e.g. Lerner et al. 2015; Phelps et al. 2014; Vuilleumier 2005). Some of these studies (Phelps et al. 2014) echo the sentiments of the earlier embodied decisions work, but go further in demonstrating how specific biasing inputs, such as affective information, play a fundamental modulatory role in the competitive process of action selection. An increasingly widespread claim, is that affective signals provide a basis for determining the salience of potential actions (Barrett and Bar 2009; den Ouden et al. 2012; Lindquist et al. 2012).

More specifically, the notion of *core affect*, which Lindquist et al. (Lindquist et al. 2012) define as "the mental representation of bodily sensations that are sometimes (but not always) experienced as feelings of hedonic pleasure and displeasure with some degree of arousal", provides a way for an agent to know if some action is salient (i.e. good or bad for it). Importantly, this evaluation need not be considered as a separate step in a computational process. Barrett and Bar argue that activity in OFC is reflective of ongoing integration of sensory information from exteroceptive cues, with interoceptive information from the body (Barrett and Bar 2009). They claim that this supports the view that perceptual states are "intrinsically infused with affective value", such that the affective significance (or salience) of an object (or action opportunity) is intertwined with its perception. This is one area where a synthesis between the work on embodied decisions and PP could be most beneficial, as the latter provides a more developed account of the importance of interoceptive information, with which to flesh out the notion of biasing inputs in distributed decision-making.

A further suggestion may come from research that postulates a more dynamic, action-oriented account of emotional episodes. For example, Lewis and Todd (Lewis and Todd 2005) view emotional episodes as the self-organising synchronization of neural structures, which help consolidate and coordinate neural activity throughout the nervous system. The emotional episode is posited to explain how an agent selectively attends to certain perceptual states, and is not perturbed by alternative goal obstructions (e.g. alternative action opportunities). An emotional episode thus acts as an important coordinating process for distributed neural activity, importantly including the sort of biases involved in the action selection required in embodied decision-making.[8]

8 It is important to note that an emotional episode is differentiated from "core affect". An emotional episode is typically associated with an intentional object that is considered to elicit the agent's attention, and a corresponding cognitive appraisal of the stimulus. Core affect refers to a neurophysiological state that is consciously *accessible* as a simple primitive non-reflective feeling, and is an important component in emotional episodes (see Russell and Barrett 1999, for an analysis). However, even when an emotional episode is not present, there is always some felt core affect in the background, potentially ready to evolve into an intentional emotional episode.

The dynamic approach to emotions is important. It treats emotions as evolving states, rather than simply end-points, and can thus play a coordinating role that biases decision-making. As such, emotions are well-suited to perform long-term, action-guiding roles, as according to Lewis and Todd, they are directly concerned with "improving our relations with the world through some action or change of action." Therefore, an emotional episode can direct attention away from obstructions that prevent the agent from obtaining some goal, and towards an associated intentional object. Lewis and Todd take this to be a fundamental factor that (partially) defines an emotional episode, and means that an emotion can assist an agent in overcoming and responding to goal obstruction, perhaps explaining why emotional episodes persist over time. This is important for understanding how distal goal-states, which require coordinated actions can be fulfilled.

These constraints help the predictive brain to recruit the relevant neural systems, best suited to respond to the current challenges it faces, based on current expectations defined by its ongoing activity. It also dovetails with the increasing attention being paid to the relation between emotions and decision-making (Phelps et al. 2014; Lerner et al. 2015), and emotions and cognition more generally (Pessoa 2013). Action opportunities are thus selectively attended to, based partly on the needs of the organism as determined by affective information, and may be modulated by ongoing dynamics that can be associated with action-guiding emotional episodes. Far from being mere limitations, emotional episodes can be seen as playing a coordinating role, the absence of which would likely result in unmanageable disorder. Although this picture is not sufficient to account for how all distal goal-states are obtained through co-ordinated action selection, it appears to be an important contributing factor—as suggested by the work on embodied decisions. The final suggestion, which could also bear on the social nature of emotional learning, is to look beyond the brain and the body, to acknowledge the constraints of the external environment.

## 5.4 Enculturation

An embodied account of PP embraces our cognitive limitations, and looks to explore how culture and the external world have been shaped to enable us to smoothly interact with the world. The need to move beyond a neurocentric perspective to an embodied (or situated) perspective is nicely expressed in a quote by anthropologist Clifford James Geertz :

> Man's nervous system does not merely enable him to acquire culture, it positively demands that he do so if it is going to function at all. Rather than culture acting only to supplement, develop, and extend organically based capacities logically and genetically prior to it, it would seem to be an ingredient to those capacities themselves. A cultureless human being would probably turn out to be not an intrinsically talented, though unfulfilled ape, but a wholly mindless and consequently unworkable monstrosity. (quoted in Lende and Downey 2012, pp. 67-68)

Of interest to this project is recent work by (Lende and Downey 2012) on *The Encultured Brain*: an exploration of recent interdisciplinary work in the fields of neuroscience and anthropology. The relevance of this interdisciplinary work (known as *neuroanthropology*) to an embodied account of PP is captured in the following:

> A central principle of neuroanthropology is that it is a mistake to designate a single cause or to apportion credit for specialized skills (individual or species-wide) to one factor for what is actually a complex set of processes. (Lende and Downey 2012, p. 24)

Like embodied PP, neuroanthropology realises that exploring the brain alone (a form of methodological solipsism, cf. Fodor 1980) is insufficient to explain the myriad skilful interactions that define adaptive life, and instead requires turning to the notion of *enculturation*. Enculturation can be defined as the idea that certain cognitive processes emerge from the interaction of an organism situated in a

particular environmental, or socio-cultural niche. Neuroanthropology forcibly claims that many neurological capacities, such as language or skills, simply do not appear without the immersion of an organism within a particular culture (i.e. enculturation). In fact, Lende and Downey even state that "embodiment constitutes one of the broadest frontiers for future neuroanthropological exploration", and that neuroanthropology is interested in "brains in the wild", to appropriate a phrase from (Hutchins 1995). This requires understanding how our brain's support skillful activity, and also how this activity has in turn re-wired our brains. Initial evidence points to: differences of neural structure and function between East Asian and Western cultures that may account for differences in notions of self (Park and Huang 2010); cross-cultural differences in subject's ability to accurately judge relative and absolute size of objects (Chiao and Harada 2008), as well as evidence for differences in spatial representation of time (Boroditsky and Gaby 2010).

The notion of enculturation is often appealed to by those most accurately described as enactivists. For example, (De Jaegher and Di Paolo 2007) appeal to a notion of *participatory sense-making* to account for how social meaning can be generated and transformed through the interactions of a group of individuals collectively participating in collaborative activities. The notion of *participatory sense-making* is an extension of the enactivist notion of *sense-making*, which is the process that describes how an autopoietic system creates meaning through its lived experiences (Thompson 2007). For the enactivist, meaning does not exist independently of a system, but is defined by the selective interactions that are specific to (and defining of) certain phenotypes. These selective interactions could create a source of emergent meaning (Steels 2003), and also alter the functional and structural properties of our brains by retuning existing motor programs, which facilitate adaptive action selection and performance (Soliman and Glenberg 2014).

As our cognitive capacities have become increasingly advanced, we can begin to appreciate how our ability to shape our environment led to ways of simplifying it, in order to meet the requirement of minimising prediction error (i.e. making our environment more predictable). Hutchins (Hutchins 2014) offers a nice example of restructuring our material environment through interactive behaviours, which can be understood as a case of dimensionality reduction (an important component of PEM). For example, he offers the case of queueing as an instance of enabling a more straightforward perceptual experience. This is because the experience of a one-dimensional line, is more predictable than the experience of a two-dimensional crowd, and in turn the experience of a queue has a lower entropy (and thus a lower source of surprise) than the experience of a crowd. He states, "[t]his increase in predictability and structure is a property of the distributed system, not of any individual mind." (Hutchins 2014, p. 40)

Of interest to this discussion, Fabry (Fabry 2017) considers the relation between enculturation and PP arguing in favour of a complementary approach. Her focus is on the acquisition and devlopment of cognitive capacities, which she argues falls naturally out of the basic principles of PP (i.e. updating of model paramters via PEM). Her exploration of cognitive development considers the role of *learning driven plasticity*, which she defines as "the idea that the acquisition of a certain cognitive capacity is associated with the changes to the structural, functional, and effective connectivity of cortical areas." (Fabry 2017, p. 4) Echoing the sentiments of the current discussion, she states this plasticity is constrained by anatomical properties of the agent, functional biases of the underlying neural circuits (Anderson 2014), and the environmental (and cognitive) niche. Despite the wide scope, Fabry argues that learning driven plasticity, which governs the acquisition of cognitive capacities (e.g. decision-making), is realised by ongoing prediction-error minimisation mechanisms. This *predictive acquisition of cognitive capacities* requires an understanding of the ontogenetic development of the situated agent, and also consideration of the more distal phylogenetic development. Both of these pursuits must consider the interrelated development of both the body of the organism[9], and the environmenetal niche.

9 See Fabry's (Fabry 2017, p. 7) suggestion regarding learning dependent bodily adaptability, which considers the developmental trajectory of skilled motor action alongside neural plasticity.

As Fabry argues, this is best approached from an embodied perspetive. We agree with Fabry that, by connecting with relevant research in enculturation (and neuroanthropology), PP may find a complementary approach, which provides a way of answering the ultimate `why' questions behind how the brain co-evolved alongside our body and external environment. This could supplement the focus on the `how' questions that PP already seems well-suited to explain (cf. Clark 2013a).

Finally, we can reflect on how the social environment provides constraints on the sorts of action sequences we take when making decisions. Consider the decision between whether to go to Tokyo or Lima when booking a holiday online. Rather than seeing this as a synchronic choice, reported by clicking a button, we could instead view the choice behaviour as the first in a series of successive actions, which are subsequently constrained by virtue of the agent's previously acquired knowledge. In this instance, that a significant financial cost (and corresponding feeling of regret) would be incurred were she not to go ahead with the subsequently implied actions (e.g. pack bags, head to airport etc.). In addition, in cases where social costs would be incurred (e.g. backing out of a verbally agreed arrangement), it is possible to view the decision as constrained by social commitment effects, akin to the sorts earlier proposed by Lepora and Pezzulo.

The various constraints considered above demonstrate a possible way in which the flexible precision-weighting of both habitual and deliberative decision mechanisms in the brain can be importantly constrained to lead to adaptive choice behaviour. Each of these constraints offers an intriguing explanatory step towards settling the underspecification challenge raised at the start of this section. By doing so, it pushes the explanatory scope of the PP framework outside of the neural mechanisms, while retaining an important foothold on how the neural mechanisms (as described by PP) flexibly interact with external factors to reduce uncertainty and maximise the probability of desired goal-states in as efficient a manner as possible. The body is an indispensable explanatory factor in this account: it allows us to uncover ways in which overt behaviour progressively tunes sensorimotor knowledge over developmental timescales; provides constraints on urgent decisions that will help to provide a more continuous evolutionary picture; offers affectively-significant information into the decision process that guides the organism to salient action selection; and it offers us myriad ways to mutually-interact with our sociocultural environment.

## 6 Conclusion

Herbert Simon famously claimed that choice behaviour should be understood as constrained by a pair of scissors, where the blades represent the limitations of the environment and the cognitive capacities of the agent in question (Simon 1990). Although his ecological approach to (bounded) rationality was cognitivist in nature, the core truth of his statement remains valid.

PP has much to offer for the second of the blades, and will likely place important theoretical constraints on the first if we acknowledge the lessons of neuroanthropology and enculturation and seek a complementary approach. Furthermore, by developing on the work in embodied decisions, we seek to gain a greater understanding of how decision-making is a dynamically-unfolding, situated process, rather than simply a process of deliberation and commitment. It seems as though developing the PP framework, in line with the myriad ways suggested, calls for an approach that is capable of explaining how the body and the world constrain the selection of action opportunities in the brain, in order to flexibly and efficiently support the adaptive choice behaviour of a situated agent.

## References


Adams, R. A., Shipp, S. & Friston, K. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function, 218* (3), 611–643.

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain.* Cambridge, MA: MIT Press.

Baddeley, A. (1992). Working memory. *Science, 255* (5044), 556–559.

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence, 48* (1), 57–86.

Barrett, L. F. & Bar, M. (2009). See it with feeling: Affective predictions during object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364* (1521), 1325–1334.

Basso, D. (2013). Planning, prospective memory, and decision-making: Three challenges for hierarchical predictive processing models. *Frontiers in Psychology, 3*, 1–2.

Binmore, K. (2008). *Rational decisions.* Princeton: Princeton University Press.

Boroditsky, L. & Gaby, A. (2010). Remembrances of times east: Absolute spatial representations of time in an Australian aboriginal community. *Psychological Science, 21* (11), 1635–1639.

Burr, C. & Jones, M. (2016). The body as laboratory: Prediction-error minimization, embodiment, and representation. *Philosophical Psychology, 29* (4), 586–600.

Chemero, A. (2003). An outline of a theory of affordances. *Ecological Psychology, 15* (2), 181–195.

——— (2011). *Radical embodied cognitive science.* Cambridge, MA: MIT Press.

Chiao, J. & Harada, T. (2008). Cultural neuroscience of consciousness: From visual perception to self-awareness. *Journal of Consciousness Studies, 15* (10-11), 58–69.

Churchland, P. S., Ramachandran, V. S. & Sejnowski, T. J. (1994). A critique of pure vision. In C. Koch & J. L. Davis (Eds.) *Large-scale neuronal theories of the brain*. Cambridge: MIT Press.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362* (1485), 1585–1599.

———(2012). Making decisions through a distributed consensus. *Current Opinion in Neurobiology, 22* (6), 927–936.

Cisek, P. & Kalaska, J. (2010). Neural mechanisms for interacting with a world full of action choices. *Philosophical Transactions of the Royal Society B: Biological Sciences, 33*, 269–298.

Cisek, P. & Pastor-Bernier, A. (2014). On the challenges and mechanisms of embodied decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369* (20130479), 1–14.

Clark, A. (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36* (03), 181–204.

———(2013b). The many faces of precision. *Frontiers in Psychology, 4* (270), 1–9.

———(2015). Embodied prediction. In T. K. Metzinger & J. M. Windt (Eds.) *Open Mind*. Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570115.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

——— (in press). Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs*. https://dx.doi.org/10.1111/nous.12140.

Cocchi, L., Zalesky, A., Fornito, A. & Mattingley, J. B. (2013). Dynamic cooperation and competition between brain systems during cognitive control. *Trends in Cognitive Sciences, 17* (10), 493–501.

Cos, I., Duque, J. & Cisek, P. (2014). Rapid prediction of biomechanical costs during action decisions. *Journal of Neurophysiology, 112* (6), 1256–1266.

Daw, N. D., Niv, Y. & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8* (12), 1704–1711.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69* (6), 1204–1215.

De Jaegher, H. & Di Paolo, E. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences, 6* (4), 485–507.

den Ouden, H. E. M., Daunizeau, J., Roiser, J., Friston, K. & Stephan, K. E. (2010). Striatal prediction error modulates cortical coupling. *Journal of Neuroscience, 30* (9), 3210–3219.

den Ouden, H. E. M., Kok, P. & de Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology, 3* (548), 1–12.

Doll, B. B., Simon, D. A. & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology, 22* (6), 1–7.

Fabry, R. E. (2017). Predictive processing and cognitive development. In T. Metzinger & W. Wiese (Eds.) *Philoso-*

*phy and predictive processing*. Frankfurt am Main: MIND Group.

Feldman, H. & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4* (215), 1–23.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences, 3* (1), 63–109.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11* (2), 127–138.

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

Friston, K., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biol Cybern, 104* (1-2), 137–160.

Friston, K., Shiner, T., Fitzgerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. E. & Bestmann, S. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology, 8* (1), 1–20.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T. & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369* (1655), 1–12.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T. & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, 1–28. https://dx.doi.org/10.1080/17588928.2015.1020053.

Gallese, V. & Metzinger, T. (2003). Motor ontology: The representational reality of goals, actions and selves. *Philosophical Psychology, 16* (3), 365-388.

Gershman, S. J. & Daw, N. D. (2012). Perception, action and utility: The tangled skein. In M. Rabinovich, K. Friston & P. Varona (Eds.) *Principles of brain dynamics: Global state interactions* (pp. 293–312). Cambridge, MA: MIT Press.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Glimcher, P. W. & Fehr, E. (2014). *Neuroeconomics.* Academic Press.

Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66* (4), 585–595.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences, 27* (03), 377–396.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3* (96), 1–14.

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

———(2016). The self-evidencing brain. *Noûs, 50* (2), 259–285. https://dx.doi.org/10.1111/nous.12062.

———(forthcoming). The predictive processing hypothesis and 4E cognition. In A. Newen, L. Bruin & S. Gallagher (Eds.) *The Oxford handbook of 4E cognition*. New York: Oxford University Press.

Hurley, S. (1998). *Consciousness in action*. Cambridge, MA: Harvard University Press.

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.

———(2014). The cultural ecosystem of human cognition. *Philosophical Psychology, 27* (1), 34–49.

Kanai, R., Komura, Y., Shipp, S. & Friston, K. (2015). Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences, 370* (1668), 20140169–20140169.

Klaes, C., Westendorff, S., Chakrabarti, S. & Gail, A. (2011). Choosing goals, not rules: Deciding among rule-based action plans. *Neuron, 70*, 536–548.

Lee, S. W., Shimojo, S. & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron, 81* (3), 687–699.

Lende, D. H. & Downey, G. (2012). *The encultured brain.* Cambridge, MA: MIT Press.

Lepora, N. F. & Pezzulo, G. (2015). Embodied choice: How action influences perceptual decision making. *PLoS Computational Biology, 11* (4), 1–22. https://dx.doi.org/10.1371/journal.pcbi.1004110.

Lerner, J. S., Li, Y., Valdesolo, P. & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology, 66* (1), 799–823.

Levy, D. J. & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology, 22* (6), 1027–1038.

Lewis, M. D. & Todd, R. M. (2005). Getting emotional. *Journal of Consciousness Studies, 12* (8-10), 210–235.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E. & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences, 35* (03), 121–143.

O'Regan, J. K. & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24* (05), 939–973.

Padoa-Schioppa, C. (2011). Neurobiology of economic choice: A good-based model. *Annual Review of Neuroscience, 34*, 333–359.

Park, D. C. & Huang, C. M. (2010). Culture wires the brain: A cognitive neuroscience perspective. *Perspectives on Psychological Science, 5* (4), 391–400.

Pastor-Bernier, A. & Cisek, P. (2011). Neural correlates of biased competition in premotor cortex. *The Journal of Neuroscience, 31* (19), 7083–7088.

Pessoa, L. (2013). *The cognitive-emotional brain: From interactions to integration.* Cambridge, MA: MIT Press.

Pezzulo, G. (2012). An active inference view of cognitive control. *Frontiers in Psychology, 3* (478), 1–2.

Pezzulo, G., Rigoli, F. & Chersi, F. (2013). The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. *Frontiers in Psychology, 4* (92), 1–15.

Pezzulo, G., Cartoni, E., Rigoli, F., Pio-Lopez, L. & Friston, K. (2016). Active inference, epistemic value, and vicarious trial and error. *Learning & Memory, 23* (7), 322–338.

Phelps, E. A., Lempert, K. M. & Sokol-Hessner, P. (2014). Emotion and decision making: Multiple modulatory neural circuits. *Annual Review of Neuroscience, 37* (1), 263–287.

Russell, J. A. & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, 76* (5), 805–819.

Savage, L. (1954/1972). *The foundations of statistics.* Dover: John Wiley and Sons.

Sen, A. (1971). Choice functions and revealed preference. *Review of Economic Studies, 38*, 307–317.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17* (11), 565–573.

——— (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.

Shapiro, L. (2011). *Embodied cognition.* London: Routledge.

Shipp, S., Adams, R. A. & Friston, K. (2013). Reflections on agranular architecture: predictive coding in the motor cortex. *Trends in Neurosciences, 36* (12), 706–716.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41* (1), 1–20.

Soliman, T. & Glenberg, A. M. (2014). The embodiment of culture. In L. Shapiro (Ed.) *The routledge handbook of embodied cognition* (pp. 207–219). London: Routledge.

Steels, L. (2003). Intelligence with representation. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 361* (1811), 2381–2395.

Thompson, E. (2007). *Mind in life.* Cambridge, MA: Harvard University Press.

Treue, S. (2001). Neural correlates of attention in primate visual cortex. *Trends in Neurosciences, 24* (5), 295–300.

——— (2003). Visual attention: The where, what, how and why of saliency. *Current Opinion in Neurobiology, 13* (4), 428–432.

Vance, J. (2017). Predictive processing and the architecture of action. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Vlaev, I., Chater, N., Stewart, N. & Brown, G. D. A. (2011). Does the brain calculate value? *Trends in Cognitive Sciences, 15* (11), 546–554.

Vuilleumier, P. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Sciences, 9* (12), 585–594.

Wolpert, D. M. & Landy, M. S. (2012). Motor control is decision-making. *Current Opinion in Neurobiology, 22* (6), 1–8.

# Which Structures Are Out There?

## Learning Predictive Compositional Concepts Based on Social Sensorimotor Explorations

Martin V. Butz

How do we learn to think about our world in a flexible, compositional manner? What is the actual content of a particular thought? How do we become language ready? I argue that free energy-based inference processes, which determine the learning of predictive encodings, need to incorporate additional structural learning biases that reflect those structures of our world that are behaviorally relevant for us. In particular, I argue that the inference processes and thus the resulting predictive encodings should enable (i) the distinction of space from entities, with their perceptually and behaviorally relevant properties, (ii) the flexible, temporary activation of relative spatial relations between different entities, (iii) the dynamic adaptation of the involved, distinct encodings while executing, observing, or imagining particular interactions, and (iv) the development of a — probably motor-grounded — concept of forces, which predictively encodes the results of relative spatial and property manipulations dynamically over time. Furthermore, seeing that entity interactions typically have a beginning and an end, free energy-based inference should be additionally biased towards the segmentation of continuous sensorimotor interactions and sensory experiences into events and event boundaries. Thereby, events may be characterized by particular sets of active predictive encodings. Event boundaries, on the other hand, identify those situational aspects that are critical for the commencement or the termination of a particular event, such as the establishment of object contact and contact release. I argue that the development of predictive event encodings naturally lead to the development of conceptual encodings and the possibility of composing these encodings in a highly flexible, semantic manner. Behavior is generated by means of active inference. The addition of internal motivations in the form of homeostatic variables focusses our behavior — including attention and thought — on those environmental interactions that are motivationally-relevant, thus continuously striving for internal homeostasis in a goal-directed manner. As a consequence, behavior focusses cognitive development towards (believed) bodily and cognitively (including socially) relevant aspects. The capacity to integrate tools and other humans into our minds, as well as the motivation to flexibly interact with them, seem to open up the possibility of assigning roles — such as actors, instruments, and recipients — when observing, executing, or imagining particular environmental interactions. Moreover, in conjunction with predictive event encodings, this tool- and socially-oriented mental flexibilization fosters perspective taking, reasoning, and other forms of mentalizing. Finally, I discuss how these structures and mechanisms are exactly those that seem necessary to make our minds language ready.

**Keywords**

Anticipatory behavior | Compositional concepts | Cooperation | Embodiment | Event segmentation theory | Free energy principle | Homeostasis | Language | Predictive encodings | Sensorimotor learning | Social interactions

## 1 Structuring the Generative, Predictive Mind

The predictive mind (Hohwy 2013), which may be viewed as continuously "surfing" on its currently active predictions and the involved uncertainties about its environment (Clark 2013; Clark 2016), gives a very intuitive and integrative view on how our mind works. However, many details of this perspective remain to be determined. The free energy-based inference principle offers a mathematical framework to specify implementational details (Friston 2010), addressing the question how predictions may interact and how predictive structures may be learned in the first place. Furthermore, goal-pursuance

has been successfully integrated by formulations of active inference, which is anticipatory in that it takes probabilistic differences between expected and desired future states into account when inferring current behavior, thus yielding goal-directed and epistemic (that is, information seeking) behavior. In sum, the free energy principle allows the mathematical formulation of slower structural learning and faster activity adaptations (Friston et al. 2011) as well as anticipatory, active inference-based goal-directed behavior.

While all three inference aspects have been implemented successfully, the implementations so far have not come anywhere close to yielding a scalable learning system, that is, a system that is able to successfully and computationally efficiently learn in and interact with complex, real-world environments. Moreover, the learning of conceptual structures and behaviorally-relevant abstractions from continuous sensory-motor information has not yet been accomplished. Nonetheless, the available proofs of principle show that the free energy-based inference approach and the resulting conceptualization of the mind as a predictive encoding and processing system has very strong merits and seems neuro-computationally as well as cognitively plausible (Butz 2016; Clark 2016; Hohwy 2013).

One reason why scalability is still out of reach may lie in the fact that current formalizations and implementations rely on particular, hand-designed representations, within which formalizations of uncertainties, probabilistic information processing, predictive estimations, motor activities, and sensory feedback unfold. From machine learning and optimization theory perspectives, however, it is well-known that learning can only make efficient progress when particular structures can be expected in the addressed learning or optimization problems (cf. no free lunch theorem (NFL), Wolpert and Macready 1997). Thus, the theory implies that it is mandatory to uncover the structures that can be found in our environment and which a learning and optimization system should `look for.' Technically, this means that formalizations of free energy-based learning and inference should work on integrating those structural learning and information processing biases that are maximally suitable to learn from and interact with our world most effectively. Presumably, evolution has integrated those biases into our brain's free energy-based inference processes (including, for example, physiological growth and neural wiring mechanisms).

Fortunately, we do not need to start from scratch when exploring which structures are there and we do not need to be very speculative, either. Rather, psychological and cognitive science research offers various clues about fundamental structural components, which our brain tends to process in distinct manners. One very important aspect is the fact that predictive encodings must ultimately serve the purpose of flexibly and effectively planning and controlling interactions with the world. Thus, the mentioned structural learning biases need to be behavior-oriented. For example, behavioral predictions and goal-directed manipulations of entities can be encoded much more effectively by entity-relative spatial arrangements and local interactions between entities in contrast to global spatial localizations.

Thus, I have suggested that three fundamental types of predictive encodings[1] should be distinguished, which are *spatial*, *top-down*, and *temporal predictive encodings* (Butz 2016). The separation of these will lead to the development of (i) universal spatial mappings — and probably the possibility to think spatially in the first place — of (ii) higher-level, multisensory integrative perceptual encodings of entities and their particular properties, and of (iii) temporal predictive encodings, which enable the anticipation of future events on various time scales. Moreover, temporal predictive encodings enable goal-directed behavioral decision making and control as well as goal-oriented attention by means of active inference.

For abstracting the predictive encodings further, event segmentation theory (Zacks and Tversky 2001) offers an additional fundamental structural principle: the segmentation of the continuous sensory-motor experiences into *event encodings* and *event boundary encodings*. It appears that when learning focusses on the processing, detection, and induction of events, fundamental conceptualiza-

1 Please note that I use the term "predictive encodings" to explicitly distinguish such encodings from "representations." Predictive encodings are not representations per se. Rather, they are encodings of predictions about the activity state of other predictive encodings. Their partial representational properties are only a result of what is actually encoded.

tions of the environment can develop, which come in the form of spatial, property-oriented, and temporal force-based conceptualizations (Butz 2016).

Clearly, our body-grounded motivational system is the driving force that makes us interact with and explore our world in the first place. Hunger and other bodily signals, as well as social needs, determine our behavior from birth onwards, and to some extent even before that. Thus, encodings need to develop that are able to predict when and how certain motivations are typically satisfied. As a result, the predictive encodings sketched-out above can be expected to be further shaped by motivational influences. Moreover, behavior will be determined to a large extent by the bodily motivational system, such that the gathered sensory-motor statistics about the world will be strongly motivationally biased.

Finally, besides tendencies towards particular modularizations and segmentations, the human mind has developed highly versatile behavioral and social capacities. *Tool usage* is unprecedented and relies on the ability to flexibly integrate different tools into our own postural body schema. *Social interactions* require the integration of other humans into our cognitive apparatus — with the tendency to assign similar capabilities (physical and mental) to them. I thus emphasize the importance of our social abilities, and particularly cooperation and perspective taking in relation to predictive encodings. By interacting with others and acknowledging that others and even the society as an imaginary entity watch and evaluate us — and even determine further interactions with us dependent on these observations — our minds integrate us into a bigger social reality, within which any participant can take on particular roles during particular interactions (Tomasello 2014). I will discuss what this implies for our actual perceptions, interactions, and actual thoughts about our world, our 'selves', and our knowledge and cognitive capabilities.

In conclusion, I argue that while our minds are individually shaped in their details, the overall structure reflects the structures found in our environment, including physical, biological, as well as social and cultural structures. As a result, our abilities for role taking and flexible concept compositions come naturally to us and significantly contribute to the beauty and the peculiarities of our mind. No wonder that a universal grammar has been detected (Jackendoff 2002): it is the grammar of pre-linguistic human thought, which enables us to learn any available human language as a child.

## 2 How Do We Comprehend Compositional Concepts? An Illustrative Example

Let us look at an example of how our mind seems to combine entities and associated knowledge about these entities into a consistent concept composition. Interestingly, artificial intelligence has been struggling to do just that and has recently proposed the so-called *Winograd Challenge* (Levesque et al. 2012; Levesque 2014), which is named after the AI researcher Terry Winograd , who created a rather intelligent software in the 1970s (Winograd 1972), which was able to produce meaningful sentences and interactions in a blocks world. The derived challenge basically focusses on common sense reasoning, wrapped into the challenge of pronoun disambiguation. Take for example the following statement:

The ball fits into the suitcase, because it is small [large].

Clearly the pronoun 'it' refers to the ball or to the suitcase, depending on whether the adjective is 'small' or 'large'. Grammatically ball or suitcase could be chosen (and in natural English language it is actually more likely that the adjective (small or large) refers to the subject (ball). The Winograd Challenge explicitly gives a 50% chance of choosing the referenced noun correctly when no semantic information is considered. Our common sense knowledge helps us solve this pronoun disambiguation problem. In particular, as put forward by Barsalou and others (Barsalou 1999; Barsalou 2008; Butz 2008; Butz 2016; Gallese and Goldman 1998), we probably imagine the situation in some form of conceptualized, anticipatory simulation. Figure 1 illustrates some aspects of the simulation that may be activated in our minds. There is the entity 'ball' and the entity 'suitcase', which are probably encoded by means of distributed, conceptual, predictive encodings. Moreover, the verb phrase "fits into" implies

that the subject can be placed into the object, which furthermore implies that the object is a kind of container. We thus have the concept composition of a ball that can be placed into the suitcase — abstracted over spatial entities; the composition may specify that one entity is smaller than the other entity, such that it can be placed into the hollow area of the other entity.

The sentence continuation then makes a statement about why the described situation is true, as indicated by the word "because", and further specifies the spatial relationship between the two, confirming the situation of the first part. For "fits into" to be applicable, the first entity needs to be smaller than the second entity, thus, in order to maintain a consistent overall simulation, "it" must refer to the ball [suitcase] depending on the adjective.

Note that we are processing the information online while reading. This can be nicely illustrated when considering the altered sentence:

The suitcase fits into the ball.

When considering the situation described in this sentence, something feels wrong. From the description above it is not so hard to identify that the concept "fits into" is facing inconsistent objects: a ball is not a common container — especially not for objects! Moreover, it is unlikely that one has ever experienced a suitcase being placed inside a ball. However, the former aspect is probably the one that makes the sentence feel incorrect, because it calls upon our common sense knowledge and essentially makes us think "how can a suitcase be put into a ball!?"

Let us consider one more sentence modification:

The ball fits into the suitcase, because it is light [heavy].

What happens in our mind in this situation? Clearly, the adjectives "light" and "heavy" are referring to a weight concept, which is not tapped into in the first part of the sentence. Lightness and heaviness do usually not affect the concept of 'fitting into' something. However, before fully dismissing this sentence as semantically incorrect, we may develop explanations, such as weight restrictions, which may then allow us to use 'fitting into' in a more metaphorical manner. In this case, the metaphor transfers the 'container' and 'fitting into' concepts from the spatial volume realm into the weight scale realm, assuming there is a particular weight restriction (e.g. for traveling on an airplane). Interestingly, a common magnitude representation has been proposed, which may help to understand this metaphor (Walsh 2003).

Throughout the rest of the paper I will refer back to this example to highlight the importance of the involved predictive encoding structures and the activation mechanisms, which determine the currently most active predictive encodings.

## 3 Fundamental Types of Predictive Encodings

When considering predictive encodings, it seems worthwhile to contrast particular types of encoded predictions. Before introducing these types, however, I want to clarify how I use the terms *predictive encodings* and *currently active predictive encodings*.

By predictive encodings I loosely refer to a set of neurons with their neural connections, which encode particular (predictive) relationships. Usually, a particular predictive encoding will be implemented by a set of neurons in the brain. For simplicity, however, it suffices to think of an encoding as a neuron with its axon and its dendritic tree, which essentially encode how information is transferred from the pre-synaptically connected neurons via the dendritic tree, axon hillock, and axon to the post-synaptically connected neurons, without considering further details on how this works exactly (not to mention the important involvements and dynamics of the neural transmitters, oxygen supply, etc.). The encodings are predictive in nature because they structure themselves for the purpose of predicting other neural activities.

*Currently active predictive encodings* are those encodings that are currently actively firing in that they are determining the currently unfolding cognitive processing dynamics. The simplest correspondence in the brain may be a neuron that is producing an action potential. However, the active encodings addressed in this paper are probably realized in the brain by a suitable (probabilistic) combination of well-timed firing neurons. For simplicity reasons, I write about *sets of active encodings*, which may seem to imply that an encoding can only be either on or off. However, what I am actually addressing is those encodings that are currently firing sufficiently strongly to influence the unfolding cognitive dynamics — thus, for example, activating the imagining of — or 'thought' about — a particular concept.

From research in psychology and cognitive science three fundamental types of predictive encodings have often been considered separately — albeit they are certainly strongly interactive: *spatial predictive encodings*, *top-down predictive encodings*, and *temporal predictive encodings* (Butz 2016; Goodale and Milner 1992; Holmes and Spence 2004). In the following paragraphs, I detail their distinction and their most typical interactions.

### 3.1 Spatial Predictive Encodings

Spatial encodings have been distinguished from recognition-oriented encodings since the seminal papers of Mishkin et al. 1983, and later of Goodale and Milner 1992. However, we still do not know how our minds actually generate predictive spatial encodings from sensorimotor experiences. When I refer to spatial predictive encodings, I mean spatial mappings that map different frames of references onto each other, essentially predicting that the information perceived or encoded in one frame of reference is related to the one perceived in the other frame of reference. The most basic forms of such mappings are concerned with our postural body schema (Butz 2014; Holmes and Spence 2004). At birth, and probably even before birth, a baby shows signs that it has some knowledge about its own body (cf. Rochat 2010). Indeed, such spatial mappings are important not only to map different sensory sources onto each other, but also to enable spatially-oriented interactions with the environment. For example, very early and rudimentary spatial mappings appear to enable fetuses to insert their thumb into their mouth even in the womb.

Modeling such capabilities for enabling a continuous information exchange between different sensory information sources (including visual, proprioceptive, and tactile) has shown that spatial mappings reflect the structure of the external three-dimensional space — or six-dimensional when also considering the orientation of an object or a limb, relative to, for example, the body mid-axis (Ehrenfeld and Butz 2013; Ehrenfeld et al. 2013; Schrodt and Butz 2015). The spatial encodings enable the dynamic activation of the currently applicable spatial mappings. Although it has not yet been shown rigorously, it may be the case that spatial mappings that are learned by means of the free energy-based inference principle, especially when enforcing sparse, compact encodings, develop a universal spatial encoding system that enables the mapping of any frame of reference imaginable onto any other, related frame of reference, and which thus reflects the dimensionality of the outside environment.

From the psychological perspective there are various indicators that spatial encodings play a fundamental role in abstract spatial reasoning and spatial problem solving (Kneissler et al. 2014). Moreover, there are various indicators that suggest that objects are encoded in terms of relative spatial constellations — rather than fully visually. Furthermore, the perception of or the thought about unfolding spatial dynamics — such as rotations — appears to be encoded distinctly from the actual sensors and entities, which may actually cause the perception of the unfolding spatial dynamics. The result is an intermodal crosstalk between the involved modalities, including tactile, visual, and motor modalities as well as the mere thought about some dynamics — such as the mental rotation of oneself or an object (Butz et al. 2010a; Janczyk et al. 2012; Liesefeld and Zimmer 2013; Lohmann et al. 2016).

## 3.2 Top-down Predictive Encodings

Top-down predictive encoding is the most basic type of predictive encoding. It has been investigated in detail in the neuro-vision literature (Chikkerur et al. 2010; Giese and Poggio 2003; Rao and Ballard 1998). From the psychological perspective, however, it seems that deeper differentiations in the top-down encodings develop only after the fundamental spatial encodings are sufficiently well structured. Possibly one of the first characterizations of top-down predictive encodings comes from the Gestalt psychologists (cf. Koffka 2013), investigating to what extent we can deduce and imagine whole figures from particular sensory input. Point-light motion figures are a well-known example, in which we tend to perceive, for example, a walking human person although we only see a few points of light that are attached to particular body parts. Even the size, agility, gender, and emotions of the person can to some extent be deduced solely by observing the motion dynamics (cf. e.g. Pavlova 2012).

Note that top-down predictive encodings typically encode predictions of particular aspects of a stimulus while ignoring others. For example, specialized areas in the visual cortex have been identified that selectively process color, visual motion, or complex edges. Indications of similar somatosensory property encodings in a corresponding ventral stream have been identified as well. Top-down predictive encodings may thus be generally characterized as predictions about typical perceivable properties of some entity — and these property predictions may address any available modality — including bodily and motivational signals — as well as combinations of modalities. In combination with spatial mappings, the currently active top-down encodings may be flexibly projected onto the currently relevant sensory-grounded frames of reference — such as onto the correct position in the retinotopic frame of reference or onto the correct position in a tactile, somatosensory body map.

## 3.3 Temporal Predictive Encodings

The third types of predictive encodings, which strongly interact with spatial and top-down predictive encodings, are temporal predictive encodings. Essentially temporal predictive encodings predict activity changes in the two other types of fundamental predictive encodings. The first temporal predictive encodings that develop in our brain are probably those concerning our own body — which muscular activities have which body postural and other perceptual (mainly proprioceptive) effects? For example, it has been shown that the visual effect caused by saccades and the opening and closing of one's eyes is anticipated by an information processing loop through the thalamus (Sommer and Wurtz 2006). Underlying this processing principle is the *reafference principle* (von Holst and Mittelstaedt 1950): corollary discharges of motor activities are converted into sensory predictions of the consequent effects. This principle was further spelled-out by the theory of anticipatory behavioral control and related theories from cognitive psychology (Prinz 1990; Hoffmann 1993).

Note that temporal predictive encodings — irrespective of where they apply and what exactly they predict — process information over time and thus predict upcoming changes. The nature of the changes, however, may differ in very fundamental ways, ranging from sensory changes (such as a change in color, in stiffness, in vibration, in loudness etc.) to abstract property changes (such as weight, content type, amount, etc.) and spatial changes (such as displacements and changes in orientation). Temporal predictive encodings can be expected to be active together with any top-down and spatial predictive encodings, effectively encoding the currently imaginable potential property or spatial changes, respectively.

Take the example of the ball with some of its activated, characteristic spatial, temporal, and top-down encodings shown in Figure 1a. Top-down predictive encodings will, for example, generate visual predictions of roundness, and possibly more concrete images, such as those of a soccer ball. Pre-activated temporal predictions may anticipate the consequences of the ball interacting with other objects as well as typical motion dynamics (e.g. starting to roll, bounce, fly, react to a kick in a certain way, etc.). Finally, without additional information, active spatial predictive encodings may predict a typical standard ball size and a somewhat central location in front of us.

Imagine now the situation where we watch a ball lying outside — say on a sidewalk — and it suddenly begins to role in one direction seemingly without any cause. What would be the constructed explanation? Probably we come up with the explanation that it must be a windy day and the wind must have forced the ball to start moving (possibly supported by an additional suitable slope of the sidewalk). The activation of the rolling temporal predictive encoding requires the presence or the activation of a force. Initial experiences of such forces are generated by our own motor behaviors (experiencing bodily restrictions in the womb, for example). Over time, top-down predictive encodings can be learned that generalize more behaviors to forces, which may be generated by our motor behavior but also by other means. These encodings in turn predict the activation of temporal predictive encodings of the expected changes (including spatial and property changes) that are typically caused by the encoded forces. For example, a "pushing force" encoding can develop that predicts, on the one hand, motor behavior activities that can generate such a force (a top-down prediction) and, on the other hand, changes in motion of the entity that is being pushed by the force (a temporal prediction).

## 4 Event-Oriented Conceptual Abstractions

While the proposed fundamental types of encodings may be able to encode all kinds of predictions, and top-down predictions typically generate abstractions, several additional psychological theories suggest that in order to build conceptual schemata about the environment, the continuous sensorimotor information flow needs to be segmented systematically.

The theory of anticipatory behavioral control (ABC, Hoffmann 1993; Hoffmann 2003), the common coding approach (Prinz 1990), and the theory of event coding (TEC, Hommel et al. 2001) all imply that actions are encoded in close relation to the action effects they tend to produce. ABC further postulates that such commonly encoded action-effect complexes are endowed with the critical conditions that enable the action-effect complex to take place. For example, an object needs to be in reach in order to be graspable, or an object must not be too heavy such that it is still movable. A computation model of the ABC theory has indeed shown high learning capacity and the potential to model typical adaptive behavior, such as latent learning in a maze, which is observable, for example, in rats (Butz 2002; Butz and Hoffmann 2002). On the other hand, the common coding approach, and its further formalization into TEC, emphasizes that action-effect complexes are co-encoded in a common, abstract code, which coordinates, anticipates, and controls the action-effect complex.

All three theories essentially emphasize that our brains seem to develop encodings of distinct motor activity-effect complexes, which can be selectively activated dependent on the current context. As detailed above, temporal predictive encodings often result in predictive codes of motor activities and their effects. Moreover, motor activities may be substituted by force-effect encodings, which may indeed be the type of encoding envisioned by TEC. Thus, an event can be understood as the application of a particular force in a particular situation.

A more general definition of an event originates from studies on event segmentation. The event segmentation theory (EST), which was derived from these studies, characterizes an event as "a segment of time at a given location that is conceived by an observer to have a beginning and an end". (Zacks and Tversky 2001, p. 3). Later, EST was refined further such that event segments were related to unfolding predictions and it was suggested that "[…] when transient errors in predictions arise, an event boundary is perceived". (Zacks et al. 2007, p. 273).

When considering motor activities, a simple event may thus be understood as a simple, unfolding motor activity, such as a grasp. EST has not been closely related to theories that focus on motor actions, such as ABC or TEC. However, when considering all of these theories in the light of predictive encodings, it appears there is a close relationship: an event may be characterized by a particular set of predictive encodings starting with the onset of this set and ending with the offset of this set. The (also predictive) encodings of the situational properties at which point a particular event typically

commences and at which point it may end then characterize the context, the necessary aspects that can bring the event about, and the ones that can stop it.

Reconsidering the ball example, an event encoding may characterize the typical behavior of a rolling ball by temporal predictive encodings of spatial displacements and of accompanying visual motion signals, which indicate the rolling motion. Moreover, motion onset, that is, the prediction of the onset of the particular temporal predictive encoding of spatial displacement, may be predictable by a top-down force encoding, which can be activated given, for example, the impact of any other moving entity — including one's foot but also a strong gust of wind.

Although the exact formulations of how such event-oriented predictive encodings may develop from basic spatial, top-down, and temporal predictive encodings still need to be spelled-out in detail, a recent algorithm that builds such event encodings from signals of *temporary surprise* seems promising (Gumbsch et al. 2016). One critical aspect of the algorithm is the formulation of temporary surprise, which corresponds to a temporal state of large free energy but which is preceded and succeeded by low free energy states in the involved predictive encodings. Interestingly, this approach was additionally motivated by research on hierarchical reinforcement learning and the challenge to automatically form conceptual hierarchies from sensorimotor signals (Butz et al. 2004; Simsek and Barto 2004; Botvinick et al. 2009).

## 5 Continuously Unfolding Predictive Encoding Activities

When assuming the existence of a complex network of the described predictive encodings, the currently active encodings constitute the currently considered concepts and their composition. While this may sound intuitively plausible, it is still unknown how these concept compositions are selectively activated, maintained, and dynamically adapted over time. One could assume the basic mechanism would be free energy-based inference. But how does the processing unfold concretely?

From modeling flexible behavioral control and the maintenance of a postural body image over time (Butz et al. 2007; Ehrenfeld et al. 2013; Kneissler et al. 2014) and from the close relationship of these mechanisms to free energy-based inference (Kneissler et al. 2015), evidence has been accumulated that suggest that at least a three-stage information processing mechanism may continuously unfold over time. First, the currently active encodings may be considered as the prior predictions about the current environmental state, including the state of one's body. Next, sensor information integration may take place, yielding local posterior activity adaptations, taking the uncertainties of the prior predictions and sensory information content into account.

After sensor information integration, the predictive network strives towards global consistency, adapting its active predictive encodings further for the purpose of increasing consistency between the active encodings themselves, that is, to minimize internal free energy. As a result, the active encodings move toward a distributed attractor, which comes in the form of a free energy minimum. Note that it seems impossible to reach a global minimum in such a distributed system, in which only local interactions unfold. Note furthermore that this process naturally takes all knowledge about the environment, which is represented in the learned predictive encodings, into account. As a result, local minima, that is, local attractors, will be approximated, which take the predictive activities only from connected encodings into account. Suitably modularized partitions of information contents, such as the spatial and top-down encodings, may optimize this distributed free energy minimization process.

In relation to the concept of a Markov blanket , which is used by Friston and others to derive the theory that the brain develops a predictive model of the (indirectly perceived) outside environment (Friston 2010), consistency enforcement may be thought of as a process during which distributed, internal Markov blankets are at play. That is, while striving for local consistencies in the currently active network of predictive encodings, local adaptations are made, which take the predictive activities of the connected predecessors, successors, and the predecessors of the successors into account. As a

result, the adaptations yield approximate locally consistent (within the respective Markov blanket) but distributed state estimates, which approximate the theoretical, global free energy minimum.

Finally, following the reafference principle, temporal predictions will be processed, generating the next prior predictions. When this anticipation of next sensory information is expanded to all active temporal predictive encodings — including those that do not depend on one's own motor behavior but that depend on the estimated presence of other forces in the environment and the presence of current motion — then essentially the temporal dynamics of the environment are predicted. Given that the global posterior reflects the actual environmental model rather well and given further that the temporal predictive encodings are relatively accurate, hardly any surprise will be encountered when processing the next sensory information and the system will be so-to-say 'in-sync' with the world. Incorrect temporal, spatial, or top-down predictions, on the other hand, will yield larger surprises, that is, larger free energy when processing the incoming sensory information.

The implications of the sketched-out process are diverse. Let us reread the first part of the sentence about the ball and the suitcase above. What will actually happen according to the sketched-out process when considering a word-by-word processing level? The article "the" prepares the predictive encodings for a concrete concept, generating predictions about the next word. Spatial predictive priors will be uniform, or rather will estimate an uncertain central default location. Top-down predictive priors remain unspecified (for example, a uniform distribution in the respective available encodings). Next, the noun "ball" triggers the activation of a more or less concrete conceptual encoding of some sort of ball with its typical properties. Some of the different types of active predictive encodings are sketched-out in Figure 1a. Note also how the article will be merged with the ball, leading to the assumption that the sentence refers to one concrete ball. Thus, on the one hand, one concrete entity code needs to be activated, while, on the other hand, this entity may be associated with all imaginable top-down, predictive ball entity properties. Because these more concrete encodings are very diverse, the activated focus will remain on the abstract but particular ball concept. It remains to be shown how a concrete, but unspecified entity is imagined by our brain. Adhering to our distinct types of predictive encodings, spatial predictive encodings should restrict the imagined entity to one location, while all imaginable ball properties are associated with this activated location via the activated top-down predictive encodings characterizing the term "ball".

Next, "fits into" will imply that the ball can be put or is inside another entity. Thus, another relative spatial predictive encoding must be activated, which relates a yet undefined entity and the ball, such that the dimensions of the ball fit into the other entity's container part. In fact, the activation of the entity will most likely activate predictive encodings that characterize a *containment property*, as sketched-out in the temporal interaction consequence encodings in Figure 1b for the suitcase. Moreover, active temporal predictive encodings will generate the next prior, which essentially leads to the expectation that the container entity will be specified.

Finally, "the suitcase" indeed makes the expected container concrete, causing the activation of the suitcase concept as a particular example of a container. Thus, the container concept is integrated into the suitcase concept. Moreover, the other predictive encodings are verified and adjusted. For example, the predictive encodings that characterize the size of the ball will be adjusted such that the size of the ball is smaller than the size of the suitcase, essentially increasing the consistency of the imagined whole concept composition.

## 6 Motivations, Goals, and Intentions

I have focused on the information processing and neural activity adaption mechanisms in the above paragraphs. When considering action decision making, however, active inference mechanisms become necessary. Free energy-based formulations of active inference take imaginable future states into account. The tendency to achieve particular futures comes from the principle that internal predictions expect unsurprising outcomes. Combined with homeostatic internal states, the result is a system that strives to sustain internal homeostasis, because very imbalanced drives cause very 'surprising' and thus unfavorable signals.

This principle is closely related to the principle of living systems as autopoietic systems (Maturana and Varela 1980). While the general principle is closely related to reinforcement learning, the formulations avoid the introduction of a separate term that characterizes reward. Rather, reward is integrated into the free energy formulations, such that extrinsic reward is akin to a successful avoidance of overly surprising signals, which can be, for example, unfavorable signals about one's own bodily state. Various artificial system implementations have successfully generated agents that are self-motivated and that are both curious and goal-oriented at the same time. For example, latent learning in rats is fostered by an inherent curiosity drive, while the gathered knowledge can then be used to act in a goal-directed manner (Butz et al. 2010b). The maintenance of a maximally effective balance between the two components, however, remains to be controlled by an appropriate choice of parameter values (Butz et al. 2010b; Friston et al. 2015; Hsiao and Roy 2005).

As a result, from the perspective of the described predictive encoding network with its unfolding, dynamically adapting current activities, it seems necessary that the currently active encodings about the considered situation also need to enable the partial simulation of potential futures forward in time. The proposed hierarchical activity encoding based on event encodings seems to be ideally suited for this matter, because it enables planning and reasoning on abstract predictive encoding levels. For example, when we want to drink out of a glass in front of us, we first need to reach for and grasp the glass and then transport it suitably to our mouth, tilt it properly, and finally drink out of it in a coordinated manner. Thus, the final goal pre-determines the successive subgoals of being in control of the glass and so forth (Gumbsch et al. 2016). Interestingly, cognitive psychological modelling literature suggests that the encodings of spatial relations between entities (such as the hand, the glass, and the mouth) and their potential space-relative interactions seem to be particularly well-suited for cognitive reasoning (Kneissler et al. 2014). It may indeed be the case that spatial predictive encodings, which must have evolved primarily for the control of flexible interactions with objects, tools, and other entities in space, are recruited by our brains to enable other planning and reasoning processes as well. Note that such spatial encodings need to be highly swiftly and flexibly adaptable to the current circumstances, to changes in these circumstances, and for enabling the consideration of such changes caused be own motor activities. As a result, the imagining of — or thoughts about — entity interactions becomes possible in a similarly swift and flexible manner.

Combinations of the present and considered futures then lead us to act — to the best of our knowledge — in our own best interest. We act intentionally to minimize uncertainties about achieving our desired goals. Moreover, we do this on all levels of abstraction that are imaginable by our minds. Priorities certainly vary and depend on the extent to which we prefer particular states and dislike others. Nonetheless, this point of view gives us an inherent and highly individual intentionality, which essentially determines our character.

Let us go back to the ball again. When we want to project the ball into the goal, we may attempt to kick it when we are in range — because we want to produce a flying ball event which may end, if suitably executed, with the ball entering the goal's interior. Similarly, when we go on a trip and pack a suitcase, we activate predictions about the presence of the items we pack into the suitcase wherever we intend to go with the suitcase. If we expect to play with the ball when we have reached the destination, we may consider packing the ball and thus consider if the ball fits into the suitcase. Given it fits, we may actively put the ball into the suitcase by using, for example, our hands. Note again how first the displacement event is activated, which then causes the activation of particular means, such as particular motor actions, that are believed to cause the displacement. Note also that as learning strongly depends on the types of experiences that have been gathered, the developing predictive encodings will strongly depend on the generated active inferences, which, vice versa, depend on the available predictive encodings, that is, the gathered knowledge about the world and about our ability to interact with it and to manipulate it. As a result, we will choose those means to put the ball into the suitcase that we are most used to (for example, by means of our hands or our feet).

## 7 Perspective Taking, Hypothetical Thinking and Role Taking

As our bodies and our 'selves' become "a public affair" over the first years of our lives, so does our behavior and our social interactions with others (Rochat 2010). In fact, it appears that we tend to constitutionalize our social interactions in such a way that an imaginative entity, that is, the 'public' as a whole is the one that is perceived as (partially) watching us (Tomasello 2014). As we have particular expectations about the general knowledge of others — such as that everybody knows how to open a door, how to count, or the common words of his or her mother tongue — we have particular expectations about social rules of interactions and basic tendencies for interactions, even with nearly complete strangers within a society or a particular culture (Tomasello 2014). These social capacities, which are strengthened and shaped further by our versatile communication capabilities, inevitably enable us to view ourselves not only from the geometric (that is, purely spatial) but also form the epistemic perspective of somebody else. And this 'somebody else' does not even need to be a person but may be an imagined knowledge entity.

A somewhat similar capacity develops when mastering tool use. The tool temporarily becomes part of our body, thus *subjectifying* the tool (Butz 2008; Iriki 2006). From a social cooperative perspective, we can act as a tool — as when handing over the butter at the breakfast table. Similarly, we can perceive our body as a tool, as when we attempt to undertake a task with our hands that is usually accomplished with the help of a particular tool. In both of the latter cases, we essentially *objectify* our body — or a part of it — as a tool.

When manipulating and interacting with objects — possibly with the help of tools — as well as when interacting with (and sometimes also manipulating) others, we experience particular interaction roles and the currently involved goals. When we are the one to initiate and control an interaction, we are the intentional actor, while the manipulated object or person is the recipient. Likewise, we may be the recipient and another person may be the intentional actor. This is probably the reason why we tend to *subjectify* objects when they interact with us by chance in peculiar ways — such as the infamous apple that is said to have fallen onto Sir Isaac Newton's head 'giving' him the idea of universal gravity. Due to the peculiarity of the interaction, we tend to *subjectify* the object and thus assign intentions to it — essentially in the same way we attribute intentions to the behavior that we observe in other humans.

It seems that these social interactions and tool manipulation capabilities are thus the key to enable us to take on different perspectives and thus to flexibly switch roles (actor, recipient, means, space, time) mentally. As a further consequence, we become able to learn the human language we are confronted with. This is probably the case because human languages offer a means to satisfy our drive to communicate with others in order to coordinate social interactions successfully. The grammar of a language essentially constrains how we assign roles to the individual entities addressed in a sentence, including their relationship in space and time and the particular interaction addressed. The prior assumption of the pre-linguistic mind is that verbalized states of affairs in the environment and changes in the environment typically refer to particular entities, relative spatial relations between entities, and particular interactions of entities (cf. also Knott 2012). In all these cases, it is important to clarify exactly which entities are referred to and what role they play. The detection of ambiguities during communication leads to the addition of words and further learning of the grammar the developing child is confronted with.

Let us re-consider the original sentence and analyze the second part of it from a reasoning perspective with the ability to take on different perspectives and thus the ability to construct counterfactuals and the involved hypothetical alternative scenarios.

The ball fits into the suitcase, because it is small.

The word "because" suggests that there is a particular property of one of the entities that makes the first sentence true and if it was changed to another property — typically the opposite — then the sentence would be false. To simulate this, we need to be able to change our image of one or the other object. The property put forward is of type "size" and the term "small" implies being "not large". The

property can be directly associated with the constructed mental situation of one entity fitting into another entity. Revisiting the constructed network of active predictive encodings (Figure 1c) and changing the size of either object allows the construction of two counterfactual situations. Changing the ball to a larger ball may make it not fit, because the combined relative spatial and size encoding would yield a spatial overlap of the entities, implying that the ball does not fit. On the other hand, changing the suitcase to a larger suitcase would not change the truth of the first part, because no overlap would occur and the ball would still fit. Thus, it is much more plausible that "the ball" is referenced by the pronoun.

Let us now change the example to show that the same principle generally can apply in a social context with particular persons or a group of persons constituting particular entities. Consider the following nearly equivalent sentence in a social context:

The Smith family fits into the discussion group, because it is insightful.

Note how the sentence specifies two entities, which are actually a group of people: The "Smith family" and the "discussion group". Again, the first entity is said to fit into the second entity — where "fits into", as above, implies that the first entity can be inserted or added to the second entity, where the second entity functions as a container. In fact, a "discussion group" allows the addition of more people, so it can function as a container. Fitting into the social group in this case, however, is not a matter of 'size' but of 'social competence'. This is exactly what the second part of the sentence implies: it argues that the Smith family fits because of its insightfulness. The opposite social property — something like "dull" or "unwise", that is, "not insightful" — would probably not fit into a discussion group, where insight and fruitful discussions are typically sought. Thus, the sentence can generally be processed in a manner similar to the ball and suitcase sentence, constructing an active predictive encoding network (an attractor that indicates consistency) and disambiguating the 'it' by imagining the two counterfactual options, that is, the 'Smith family' or the 'discussion group' being not insightful. As a result, it is much more likely that the 'Smith family' is referenced by the pronoun, although the other interpretation cannot be fully excluded.

## 8 Conclusions

While predictive coding intuitively makes a lot of sense to many of us, I have argued that it is necessary to further describe the most likely kinds of predictive encodings that are developing in our brains and how they interact. I have argued that three basic types may constitute the building blocks of our thoughts: top-down, spatial, and temporal predictive encodings. Moreover, I have argued that, starting with basic predictive encodings of the encountered sensorimotor experiences, event-oriented abstractions can lead to the development of event and event boundary encodings, which can lead to conceptual encodings of the relevant properties to bring particular events about.

Moreover, I have argued that the simulation of a particular thought, a particular situation, or a particular event in a particular context is essentially encoded by a network of active predictive encodings. While, for example, reading a sentence, a constructive process unfolds, which attempts to activate those predictive encodings that form a maximally consistent, interactive network. The network reflects what is believed to be the sentence's content, including its implications. By changing parts of the activated predictive encodings, it is possible to alter the imagined situation in meaningful ways and to probe the resulting consistency. As a result, it is possible to argue semantically about certain events, situations, and statements. Our developing tool usage and social capabilities and the involved encodings seem to support the necessary perspective taking abilities — and thus the ability to imagine conceptual environmental interactions in the past, the future, and by others (Buckner and Carroll 2007). Due to the event-oriented predictive encodings, such imaginings are conceptual because the involved predictive encodings focus on those aspects of the environment that are believed to be relevant for particular situations in and interactions with the world. Moreover, mental manipulations of such imaginings allow the probing of

situational changes and their effects, including property, spatial, and temporal changes — thus enabling planning, reasoning, and the pursuance of hypothetical thoughts. Finally, the social perspective as well as our tool use abilities enable us to objectify ourselves and thus to develop explicit self-consciousness.

Clearly, details of the proposed theory need to be further defined in the near future, in order to verify the involved hypotheses and to further differentiate the identified kinds of predictive encodings. The best way to go forward may be to develop actual implementations of artificial cognitive systems in virtual realities, in addition to the necessary further interdisciplinary philosophical, neuro-psychological, and linguistic research.

**Figure 1:** Illustrations of the active predictive encoding networks for the thought about "the ball" and "the suitcase" as well as the composition with the connecting information "fits into". Note the predictive interactions and the consistencies between the illustrated active encodings. The visualized encodings include top-down predictions about the visual appearances, spatial predictions about where, relative to the observer (and relative to the other object), one object may be present, as well as what size it may be. Temporal predictive encodings predict interaction consequences as well as potential motion dynamics. In the concept compositions these may annihilate each other, indicating the thought of the ball lying stably inside the suitcase with unlikely current motion dynamics.

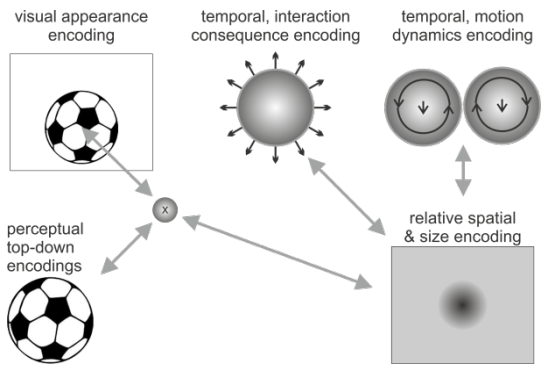


**Figure 1 a:** Sketch of predictive coding network for "the ball".

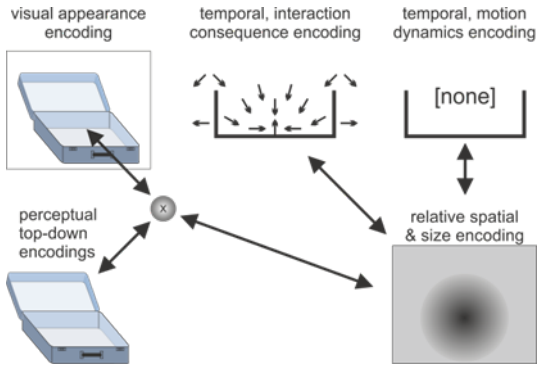


**Figure 1 b:** Sketch of predictive coding network for "the suitcase".

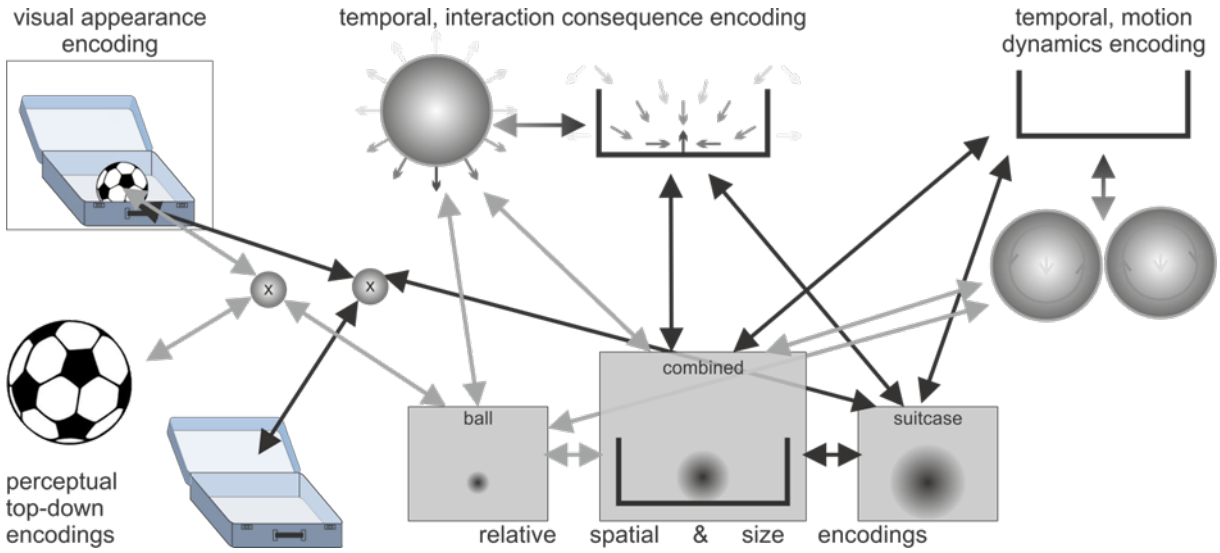


**Figure 1 c:** Concept composition: "The ball fits into the suitcase."

## References


Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577–600.

———(2008). Grounded cognition. *Annual Review of Psychology, 59*, 617-645.

Botvinick, M., Niv, Y. & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition, 113* (3), 262 - 280. https://dx.doi.org/10.1016/j.cognition.2008.08.011.

Buckner, R. L. & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11*, 49-57.

Butz, M. V. (2002). *Anticipatory learning classifier systems.* Boston, MA: Kluwer Academic Publishers.

———(2008). How and why the brain lays the foundations for a conscious self. *Constructivist Foundations, 4* (1), 1-42.

———(2014). Rubber hand illusion affects joint angle perception. *PLoS ONE, 9* (3), e92854. Public Library of Science. https://dx.doi.org/10.1371/journal.pone.0092854.

———(2016). Towards a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology, 7* (925). https://dx.doi.org/10.3389/fpsyg.2016.00925.

Butz, M. V. & Hoffmann, J. (2002). Anticipations control behavior: Animal behavior in an anticipatory learning classifier system. *Adaptive Behavior, 10*, 75-96.

Butz, M. V., Swarup, S. & Goldberg, D. E. (2004). Effective online detection of task-independent landmarks. In R. S. Sutton & S. Singh (Eds.) *Online proceedings for the ICML'04 workshop on predictive representations of world knowledge* (pp. 10). Online. http://homepage.mac.com/rssutton/ICMLWorkshop.html.

Butz, M. V., Herbort, O. & Hoffmann, J. (2007). Exploiting redundancy for flexible behavior: Unsupervised learning in a modular sensorimotor control architecture. *Psychological Review, 114*, 1015-1046.

Butz, M. V., Thomaschke, R. Linhardt, M. J. & Herbort, O. (2010a). Remapping motion across modalities: Tactile rotations influence visual motion judgments. *Experimental Brain Research, 207*, 1-11.

Butz, M. V., Shirinov, E. & Reif, K. L. (2010b). Self-organizing sensorimotor maps plus internal motivations yield animal-like behavior. *Adaptive Behavior, 18* (3-4), 315–337.

Chikkerur, S., Serre, T. Tan, C. & Poggio, T. (2010). What and where: A Bayesian inference theory of attention. *Vision Research, 50*, 2233-2247. https://dx.doi.org/10.1016/j.visres.2010.05.013.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Science, 36*, 181-253.

———(2016). *Surfing uncertainty: Prediction, action and the embodied mind.* New York: Oxford University Press.

Ehrenfeld, S. & Butz, M. V. (2013). The modular modality frame model: Continuous body state estimation and plausibility-weighted information fusion. *Biological Cybernetics, 107*, 61-82. https://dx.doi.org/10.1007/s00422-012-0526-2.

Ehrenfeld, S., Herbort, O. & Butz, M. V. (2013). Modular neuron-based body estimation: Maintaining consistency over different limbs, modalities, and frames of reference. *Frontiers in Computational Neuroscience, 7* (148). https://dx.doi.org/10.3389/fncom.2013.00148.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*, 127- 138. https://dx.doi.org/10.1038/nrn2787.

Friston, K., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics, 104* (1-2), 137-160. https://dx.doi.org/10.1007/s00422-011-0424-z.

Friston, K., Levin, M. Sengupta, B. & Pezzulo, G. (2015). Knowing one's place: A free-energy approach to pattern regulation. *Journal of The Royal Society Interface, 12* (105). https://dx.doi.org/10.1098/rsif.2014.1383.

Gallese, V. & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2* (12), 493–501.

Giese, M. A. & Poggio, T. (2003). Neural mechanisms for the recogniton of biological movements. *Nature Reviews Neuroscience, 4*, 179-192.

Goodale, M. A. & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15* (1), 20-25. https://dx.doi.org/10.1016/0166-2236(92)90344-8.

Gumbsch, C., Kneissler, J. & Butz, M. V. (2016). *Learning behavior-grounded event segmentations* (pp. 1787-1792). Cognitive Science Society.

Hoffmann, J. (1993). *Vorhersage und Erkenntnis: Die Funktion von Antizipationen in der menschlichen Verhaltenssteuerung und Wahrnehmung. [Anticipation and cognition: The function of anticipations in human behavioral control and perception.].* Göttingen, GER: Hogrefe.

———(2003). Anticipatory behavioral control. In M. V. Butz, O. Sigaud & P. Gérard (Eds.) *Anticipatory behavior*

*in adaptive learning systems: Foundations, theories, and systems* (pp. 44-65). Berlin/Heidelberg, Springer-Verlag.

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

Holmes, N. P. & Spence, C. (2004). The body schema and multisensory representation(s) of peripersonal space. *Cognitive Processing, 5*, 94-105.

Hommel, B., Müsseler, J. Aschersleben, G. & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849-878.

Hsiao, K. Y. & Roy, D. (2005). In C. Castelfranchi, C. Balkenius, M. V. Butz & A. Ortony (Eds.) *A habit system for an interactive robot* (pp. 83-90). Menlo Park, CA: AAAI Press.

Iriki, A. (2006). The neural origins and implications of imitation, mirror neurons and tool use. *Current Opinion in Neurobiology, 16*, 660-667.

Jackendoff, R. (2002). *Foundations of language. Brain, meaning, grammar, evolution.* Oxford University Press.

Janczyk, M., Pfister, R. Crognale, M. A. & Kunde, W. (2012). Effective rotations: Action effects determine the interplay of mental and manual rotations. *Journal of Experimental Psychology: General, 141*, 489-501. https://dx.doi.org/10.1037/a0026997.

Knauff, M. (2013). *Space to reason. A spatial theory of human thought.* Cambridge, MA: MIT Press.

Kneissler, J., Stalph, P. O. Drugowitsch, J. & Butz, M. V. (2014). Filtering sensory information with XCSF: Improving learning robustness and robot arm control performance. *Evolutionary Computation, 22*, 139-158. https://dx.doi.org/10.1162/EVCO_a_00108.

Kneissler, J., Drugowitsch, J. Friston, K. & Butz, M. V. (2015). Simultaneous learning and filtering without delusions: A Bayes-optimal combination of predictive inference and adaptive filtering. *Frontiers in Computational Neuroscience, 9* (47). Frontiers Media S.A. https://dx.doi.org/10.3389/fncom.2015.00047.

Knott, A. (2012). *Sensorimotor cognition and natural language syntax.* Cambridge, MA: MIT Press.

Koffka, K. (2013). *Principles of Gestalt psychology.* Abingdon, UK: Routledge.

Levesque, H. J. (2014). On our best behaviour. *Artificial Intelligence, 212*, 27–35.

Levesque, H., Davis, E. & Morgenstern, L. (2012). The Winograd schema challenge. *Knowledge Representation and Reasoning Conference. Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning.* http://www.aaai.org/ocs/index.php/KR/KR12/paper/view/4492.

Liesefeld, H. R. & Zimmer, H. D. (2013). Think spatial: The representation in mental rotation is nonvisual. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39* (1), 167–182.

Lohmann, J., Rolke, B. & Butz, M. V. (2016). In touch with mental rotation: Interactions between mental and tactile rotations and motor responses. *Experimental Brain Research*.

Maturana, H. & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living.* Boston, MA: Reidel.

Mishkin, M., Ungerleider, L. G. & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences, 6*, 414-417. https://dx.doi.org/10.1016/0166-2236(83)90190-X.

Pavlova, M. A. (2012). Biological motion processing as a hallmark of social cognition. *Cerebral Cortex, 22* (5), 981–995. https://dx.doi.org/10.1093/cercor/bhr156.

Prinz, W. (1990). A common coding approach to perception and action. In O. Neumann & W. Prinz (Eds.) *Relationships between perception and action* (pp. 167-201). Berlin Heidelberg: Springer-Verlag.

Rao, R. P. & Ballard, D. H. (1998). Development of localized oriented receptive fields by learning a translation-invariant code for natural images. *Computational Neural Syststems, 9*, 219-234.

Rochat, P. (2010). The innate sense of the body develops to become a public affair by 2-3 years. *Neuropsychologia, 48* (3), 738 - 745. https://dx.doi.org/10.1016/j.neuropsychologia.2009.11.021.

Schrodt, F. & Butz, M. V. (2015). *Learning conditional mappings between population-coded modalities* (pp. 141-148).

Simsek, Ö. & Barto, A. G. (2004). Using relative novelty to identify useful temporal abstractions in reinforcement learning. *Proceedings of the Twenty-First International Conference on Machine Learning (ICML-2004)*, 751-758.

Sommer, M. A. & Wurtz, R. H. (2006). Influence of the thalamus on spatial visual processing in frontal cortex. *Nature, 444*, 374-377.

Tomasello, M. (2014). *A natural history of human thinking.* Cambridge, MA: Harvard University Press.

von Holst, E. & Mittelstaedt, H. (1950). Das Reafferenzprinzip (Wechselwirkungen zwischen Zentralnervensystem und Peripherie.). *Naturwissenschaften, 37*, 464-476.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive*

*Sciences, 7* (11), 483–488. https://dx.doi.org/10.1016/j.tics.2003.09.002.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology, 3* (1), 1–191.

Wolpert, D. H. & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1* (1), 67-82.

Zacks, J. M. & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127* (1), 3–21. https://dx.doi.org/10.1037/0033-2909.127.1.3.

Zacks, J. M., Speer, N. K. Swallow, K. M. Braver, T. S. & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin, 133* (2), 273–293. https://dx.doi.org/10.1037/0033-2909.133.2.273.

# Folk Psychology and the Bayesian Brain

Joe Dewhurst

Whilst much has been said about the implications of predictive processing for our scientific understanding of cognition, there has been comparatively little discussion of how this new paradigm fits with our everyday understanding of the mind, i.e. folk psychology. This paper aims to assess the relationship between folk psychology and predictive processing, which will first require making a distinction between two ways of understanding folk psychology: as propositional attitude psychology and as a broader folk psychological discourse. It will be argued that folk psychology in this broader sense is compatible with predictive processing, despite the fact that there is an apparent incompatibility between predictive processing and a literalist interpretation of propositional attitude psychology. The distinction between these two kinds of folk psychology allows us to accept that our scientific usage of folk concepts requires revision, whilst rejecting the suggestion that we should eliminate folk psychology entirely.

In section 1 I will introduce predictive processing, giving a quick summary of the framework that focuses on the details most relevant for my comparison with folk psychology. I will also introduce folk psychology and define the distinction between propositional attitude psychology and folk psychological discourse. In section 2 I will consider the relationship between predictive processing and propositional attitude psychology, and in section 3 I will consider the relationship between predictive processing and folk psychological discourse. Finally, in section 4 I will argue that the distinction between propositional attitude psychology and folk psychological discourse makes space for us to revise our scientific usage of folk psychological concepts without thereby eliminating folk psychology altogether.

**Keywords**

Belief | Cognitive ontology | Desire | Eliminativism | Folk psychology | Instrumentalism | Predictive processing | Propositional attitude psychology | Realism | Revision

**Acknowledgements**

I would like to thank Dan Calder, Jonny Lee, Suilin Lavelle, Dave Ward, Chris Michel, Alessio Bucci, Adrian Downey, and two anonymous reviewers, all of whom provided detailed comments on earlier drafts of this paper.

## 1 Introduction

This paper will focus on Andy Clark and Jakob Hohwy's presentations of the Bayesian brain hypothesis, which are not entirely identical, but I will note when they differ. I will use the term predictive processing to refer to their versions of this hypothesis. Other versions of the Bayesian brain hypothesis exist (see Spratling in press, for an overview), and these differ from Clark and Hohwy's in many important ways, but I will not be discussing them here. Below I present a very brief introduction to predictive processing (see Hohwy 2013, or Clark 2016, for a more detailed overview). I will also introduce what I mean by folk psychology, which includes both the classical understanding of folk psychology as propositional attitude psychology, and a more general phenomenon that I call 'folk psychological discourse'.

### 1.1 Predictive Processing

Predictive processing inverts conventional assumptions about the flow of information in the brain. Rather than starting with raw perceptual inputs that are gradually processed into refined models of the world, it begins with a rich, internally generated model that predicts incoming sensory data. These predictions are then compared with the actual inputs, and the model is updated accordingly. Overall

the system aims to minimise prediction error, which can be accomplished in two distinct ways. The model can be revised so as to more accurately predict incoming stimuli (passive inference), or the system can act on its environment in order to make its own predictions more accurate (active inference). Which kind of inference is performed (active or passive) will depend on higher-level predictions of the best way to reduce error in the current situation. Thus Clark summarises predictive processing as positing "core perception-attention-action loops in which internal models of the world and their associated precision expectations play key action driving roles" (Clark 2016, p. 71). By uniting action and perception in this way, predictive processing aims to provide a general account of cognition.

There are a few further features of predictive processing that are especially relevant to my discussion of folk psychology. Predictions can be regarded as more or less precise by the system, with the level of precision being taken into account when updating the predictions. For example, a less precise prediction will be expected to generate some error, and so may not need to be modified too much when an error signal is received. Changes in precision weighting can also drive the system to attend more or less to different sources of stimuli (Clark 2016, chapter 2). Finally, the predictive processing systems described by Hohwy and Clark are hierarchical; they consist of a nested hierarchy of precision/error units, with each level of the hierarchy predicting the current state of the unit below, which is then compared to the actual state of that unit and updated (in the next iteration) in response to any error signals that it receives. This hierarchy bottoms out in units that predict inputs received via sensory transduction, and tops out with a very abstract model, perhaps just predicting general causal laws or regularities. I will describe some further features of predictive processing in more detail as I go on to compare it with folk psychology.

## 1.2 Folk Psychology

In philosophy of mind and cognitive science, 'folk psychology' is typically taken to be synonymous with 'propositional attitude psychology', i.e. a theory of the mind based on a folk understanding of propositional attitudes such as belief and desire. According to this interpretation of folk psychology, our everyday understanding of how other people (and ourselves) think is constituted by a theory of how propositional attitudes interact with one another to cause behaviour, along with a capacity to attribute propositional attitudes to other people. A propositional attitude is an attitude, such as belief or desire, towards a proposition, such as 'the sky is blue'. When I see you pick up your umbrella, I may attribute to you the belief that it is going to rain and the desire not to get wet, which together produce the behaviour of picking up your umbrella. One can be committed to this interpretation of folk psychology with or without being committed to the additional claim that it is a true account of how the mind actually works. Whilst Fodor (Fodor 1975) and Churchland (Churchland 1979, Churchland 1981) both agree that folk psychology consists in the attribution of propositional attitudes, they disagree about whether or not there are actually propositional attitudes 'in the head'. Historically the debate between realists and eliminativists about folk psychology has shared a foundational assumption that folk psychology is equivalent to propositional attitude psychology, and that propositional attitude psychology aims to literally describe the structure of human cognition. For this reason I refer collectively to both realists and eliminativists about propositional attitude psychology as 'literalists'. There are a number of alternative positions that one can take towards propositional attitude psychology, including Dennett's intentional stance (Dennett 1989), Davidson's anomalous monism (Yalowitz 2014), and various kinds of dualism or non-physicalism. In this paper I will be focusing primarily on the literalist interpretation of propositional attitude psychology, as it has the most immediate consequences for cognitive scientific practice, contrasted with the other interpretations that typically have less to say about what sorts of things should populate our scientific ontology. As such, in section 2 I will be assessing whether predictive processing is compatible with the existence of propositional attitudes 'in the head'.

The philosophical literature on propositional attitudes has tended to focus fairly exclusively on beliefs and desires, but these are not the only propositional attitudes that could exist. As well as believing or desiring that it is raining, we could also hope that it is raining, or expect that it is raining, or suspect that it is raining, and so on indefinitely (cf. Stich 1983, p. 217). Stich suggests that what many theorists do at this point is simply to redefine 'belief' and 'desire' such that they cover any potential propositional attitude, but notes that this may just distort the terms so far that they cease to bear any resemblance to their folk usage (Stich 1983, p. 218). Like Stich, I will not take sides in this debate — what I have to say about belief and desire will generalise to other propositional attitudes.

Whilst it is typically equated with propositional attitude psychology, folk psychology is in fact a far more complex phenomenon, consisting of (at least) behavioural predictions, mental state attributions, narrative competency, and normative constraints (cf. Ratcliffe 2006). I will refer to this complex phenomenon collectively as 'folk psychological discourse', in order to capture the sense in which our everyday descriptions of behaviour consist of more than just theoretically motivated ascriptions of propositional attitudes. We also describe and predict behaviour in non-mentalistic terms, situate our descriptions and predictions in an on-going narrative structure, and make explicit normative judgements about how one ought to behave — where "ought" is understood as carrying both ethical and rational weight. The traditional characterisation of folk psychology as the attribution of propositional attitudes is only one component of this discourse. My claim that folk psychology ought to be understood in this broader sense is somewhat controversial, but I do not intend to defend it here. Those who want to restrict the use of the term 'folk psychology' to refer only to propositional attitude psychology might want to refer to these broader phenomena in some other way, such as 'mentalising' or 'common sense psychology'. Section 3 will focus on the relationship between predictive processing and folk psychology in this broader sense.

Folk psychology in both senses is closely related to social cognition, i.e. the cognitive mechanisms that facilitate social interaction and interpersonal understanding. Quadt (Quadt 2017) discusses social cognition in the context of predictive processing, and I will briefly touch on how predictive processing could influence our understanding of social cognitive mechanisms, insofar as those are relevant to folk psychological discourse, but this is not the main aim of my paper.

## 2 Predictive Processing and Propositional Attitude Psychology

Both Clark and Hohwy have suggested in informal discussion that predictive processing might be incompatible with the folk psychological conception of cognition. Clark has described the content of the predictions as "alien" and "opaque"[1], and Hohwy has acknowledged the challenge posed by predictive processing to "folk psychological notions of perception, belief, desire, decision (and much more)".[2] Clark has also written that predictive processing "may one day deliver a better understanding even of our own agent-level experience than that afforded by the basic framework of 'folk psychology'" (Clark 2013, p. 17, repeated in Clark 2016, p. 82). I take it that what both of them have in mind when they refer to folk psychology is propositional attitude psychology. This is the interpretation of folk psychology that has traditionally been of most interest to philosophers, and as such it is a good place to begin my assessment of predictive processing and folk psychology. In this section I will focus on belief and desire, although as I have already noted the issues raised here will generalize to other propositional attitudes.

1 Comment made during BPPA (British Postgraduate Philosophy Association) Masterclass on Action-Oriented Predictive Coding, University of Edinburgh, 26th-27th October 2013.

2 From the comments section of a Brains Blog featured scholar post: http://philosophyofbrains.com/2014/06/22/is-prediction-error-minimization-all-there-is-to-the-mind.aspx. See also (Hohwy 2013, p. 2) for a statement of how predictive processing might lead us to "radically reconceptualize who we are".

## 2.1 Belief

According to the conventional account of propositional attitudes, a belief is a state consisting of a proposition coupled with a positive epistemic attitude, i.e. one that regards it as true, and it will interact with other mental states so as to generate actions in accordance with the state of affairs captured by the proposition being true. It is sometimes said that beliefs have a 'mind-to-world' direction of fit — that is to say, a belief should be modified in response to how the world is, and not vice versa.

On the face of it this kind of mental state seems to fit nicely into the predictive processing story. It is natural to interpret predictions as beliefs about the world, albeit ones that are first generated and then tested, rather than being generated in response to sensory input. Indeed, some researchers have described predictions as beliefs, including Friston (see e.g. Hobson and Friston 2014), Hohwy (Hohwy 2012), and occasionally Clark himself (Clark 2016, p. 129). Given that the term 'belief' is used in a technical sense in Bayesian theory, this might be passed off as a harmless usage; however, simply equating predictions with beliefs in the everyday sense would be to ignore a crucial difference between folk psychological beliefs and the predictions invoked by the predictive processing story. The former are usually understood as determinate (you either believe something or you do not), whereas the latter are inherently probabilistic. Rather than simply believing that it is raining, a predictive processing system will assign a level of probability to it raining, and act in accordance with this probability. As Clark puts it,

> Instead of simply representing 'CAT ON MAT', the probabilistic Bayesian brain will encode a conditional probability density function, reflecting the relative probability of this state of affairs (and any somewhat-supported alternatives) given the available information. (Clark 2016, p. 41)

As a consequence of this, adopting the predictive processing framework will require either a reconception of the relationship between folk psychological beliefs and the brain, or an acceptance that beliefs are in fact probabilistic rather than determinate. I explore the first option later on in this paper, where I will argue that it is a mistake to think that folk psychology has ever been in the business of describing the structure of cognition at the same level of detail as a cognitive scientific theory like predictive processing does. Something like the second option has been explored by Pettigrew (Pettigrew 2015), who considers the epistemological implications of adopting a probabilistic notion of belief alongside the more conventional determinate notion. In section 4 I will also consider the possibility that we should modify our understanding of 'belief' rather than eliminating it from our scientific ontology.

The predictions involved in predictive processing may also be individuated at a much finer level of detail than folk psychological belief attributions usually allow for. Whilst a paradigmatic belief might be about whether or not it is raining, the content of the predictions at some levels of the hierarchy are more likely to be cashed out in terms of fine-grained details of the external world, predicting features such as edges and light gradients rather than the 'middle sized dry goods' that populate the folk ontology. Even at higher levels of the hierarchy, the content of the predictions are still somewhat unusual, as they incorporate multi-modal, emotional, bodily, and other contextual associations. Combined with the probabilistic nature that I described above, the predictions posited by predictive processing begin to look less like the everyday notion of a belief. Clark expresses something like this view himself when he writes that the "looping complexities" involved in predictive processing "will make it hard (perhaps impossible) adequately to capture the contents or the cognitive roles of many key inner states and processes using the terms and vocabularies of ordinary daily speech" (Clark 2016, p. 292).

However, Hohwy (Hohwy 2013, p. 60) describes how the relationship between a predictive processing system and a dynamically evolving world could give rise to higher-level regularities that might come to resemble something more like folk psychological contents. He gives the example of perceiv-

ing a partially occluded cat, but forming a prediction of a whole cat based on feedback from seeing different parts of this cat at different points in time as it moves behind the occluder. The content of the whole-cat prediction is relatively coarse-grained, but it would in turn predict lower-level perceptions of parts of cats that change over time. The system is thus able to account both for the diachronic appearance of rapidly changing parts of cats, and the more abstract notion of a whole cat who stands behind the occluder and is temporally extended. So, if Hohwy is correct, we might expect to see something resembling folk psychological states at the higher levels of a predictive processing hierarchy, even if these states are different to how we usually conceive of them (i.e. their content is non-linguistic, abstract, and probabilistic, rather than consisting of linguistic propositions with determinate content).

Finally, it is important to recognise the dual nature of predictions. Predictions function both as representations of the world and of ways that the system can act in the world. Via the mechanism of active inference, predictions can be used to motivate and generate actions, a feature that is usually associated more with desires than beliefs. Clark likens this feature of high-level hypotheses to Millikan's (Millikan 1996) "pushmi-pullyu" representations, which have "both descriptive and imperative content" (Clark 2016, p. 187). At this point there is a sense in which beliefs, if they were to be identified with predictions, would begin to blur into what we might more naturally characterise as desires. Hohwy himself suggests that perception and belief might both be reconceived as a single notion of expectation (Hohwy 2013, p. 72), which could go some way towards reconciling predicting processing with propositional attitude psychology, although it would require that we adopt a revisionary approach towards folk psychology. Taken a step further this revisionary approach could also involve collapsing desire in with perception and belief, leaving us with a single kind of mental state that encompasses all aspects of cognitive processing.

Belief, understood as a positive epistemic attitude towards a proposition, does not straightfowardly fit in to the ontological framework of predictive processing. Whilst proponents of predictive processing have occasionally described predictions as beliefs, they have in mind something quite different to the traditional propositional attitude interpretation of folk psychology. Folk psychological beliefs are typically determinate and take linguistic propositions as their argument, whilst predictions are probabilistic and refer to a wide range of distinct contents, most of which are likely to be non-propositional. Nonetheless, it is plausible that we might find something closer to the folk psychological notion of belief at higher levels of the predictive processing hierarchy, where coarse-grained predictions about stable features of the environment are to be found. In the next section I will turn to desire, which in the context of action-oriented predictive processing is also constituted by predictions about the world.

## 2.2 Desire

Much like a belief, a desire consists of a proposition coupled with an attitude; only this time the attitude has a world-to-mind direction of fit, and will function accordingly. If I desire that it is raining, I will not pick up my umbrella (as I would if I believed it was raining), but I might sigh deeply and complain about the heat, or invest in experimental cloud seeding technologies.

As I mentioned above, the predictions posited by Clark and Hohwy's versions of predictive processing are action-oriented. This means that as well as providing a model of the world, they also serve to motivate the system to act via the mechanism of active inference. In this latter capacity they seem to fulfil a role very much like that played by desires in the traditional account of folk psychology. They represent how the system would like the world to be, and coupled with beliefs about the current state of the world, they generate the appropriate actions to bring about this desired state of affairs. Understood in this way we might conclude that it is viable to adopt a mild revisionism, where it turns out that beliefs and desires are both instantiated by a single kind of state, an 'action-oriented prediction'.

However, as Clark draws attention to, there is another sense in which predictive processing seems to do away with desire entirely. Friston et al. 2011 write, "crucially, active inference does not invoke any

'desired consequences'" (Friston et al. 2011, p. 157), which Clark interprets as "a world in which value functions, costs, reward signals, and *perhaps even desires* have been replaced by complex interacting expectations that inform perception and entrain action" (Clark 2016, p. 129, emphasis added). The key issue here is that predictive processing inverts the conventional ordering of action causation assumed by folk psychology. Rather than a desire generating behaviour that leads to an expected outcome, a prediction of an expected outcome is generated first, which then goes on to cause behaviour that brings about that outcome. Desire seems to be relegated to a phenomenal sensation associated with this sequence of events, and does not seem to play any causal or functional role in generating either the behaviour or the outcome.

Clark argues that there need not be any contradiction here. Instead of eliminating desire from our ontology, we can reconceive of it as a consequence of the interaction between predictions and the environment (Clark 2016, p. 129). Insofar as it allows us to recognise the differences between predictive processing and folk psychology without simply eliminating the latter, I will go on endorse something like this position, but first I want to mention a further issue that it raises. Reconceiving desire as a consequence rather than a cause of action has the potential for a deeply counterintuitive picture of personal level agency. Rather than being a distinct source of actions, agency (in the guise of active inference) turns out to be nothing more than a tool used by the system to minimise prediction errors.[3] We do not do things because we want to do them; we *feel* like we want to do things *because* doing them will minimise prediction error. As Hohwy puts it, "[w]hat drives action is prediction error minimisation […] rather than what the agent wants to do" (Hohwy 2013, p. 89). Hohwy presents this as a positive result, unifying perception and action under one single mechanism (Hohwy 2013, p. 76), but for many this will seem like a sleight of hand, akin to Dennett's attempts to reconcile free will with a deterministic cognitive architecture (see Dennett 1984, Dennett 2004). Perhaps this is just a symptom of a mistaken folk conception of agency, or perhaps it points towards confusion between two distinct modes of explanation — either way, the folk concept of desire would turn out not to be doing any significant work in our cognitive scientific explanations of action generation.

The folk psychological concept of desire as an action-motivating attitude is encompassed by the predictive processing notion of an action-oriented prediction, which via the mechanism of active inference is able to act on the world in order to make itself come true. Thus, predictive processing differs from the folk notion of desire in two crucial respects: firstly, beliefs and desires are implemented by a single kind of state, an action-oriented prediction; secondly, desire is relegated to a secondary status, as it is prediction-error minimization, rather than any personal goals of the system, that drive action. Hohwy presents this as a positive result, offering the possibility of unifying perception and action under one single mechanism. I think instead that what it indicates is that the project of trying to naturalise folk psychology by identifying propositional attitudes with the theoretical posits of our best cognitive science is a mistaken one, as it misconstrues the aim and purpose of folk psychology. In the next section I will consider how a broader interpretation of folk psychology fits with the predictive processing framework.

## 3 Predictive Processing and Folk Psychological Discourse

Propositional attitude psychology is only one small part of folk psychological discourse. In our everyday lives, we are often able to make predictions about how someone is likely to behave, both in the short and long term. Sometimes we may also attribute mental states when predicting behaviour, although more typically such ascriptions are used in an attempt to explain or justify either past or future behaviour. Another type of explanation or justification can be given by folk psychological narratives, which are distinct from either behavioural or mentalistic language. All three of the above kinds of folk psychological discourse can also be used to impose normative constraints on behaviour, i.e. by

3 Colombo (Colombo 2017) makes a similar point, in the context of challenging the empirical foundations of the Humean theory of motivation.

using them in an imperative rather than descriptive mode. I will now describe each component of folk psychological discourse in more detail, before considering what impact the success of predictive processing would have on this discourse.

### 3.1 Folk Psychological Discourse

At a very basic level, we are able to avoid bumping into strangers on the street and can make use of physical cues to understand what someone is about to do, or what they expect us to do. We can improve on this basic capacity for behaviour prediction by engaging in explicit theorising. Rather than just predicting future behaviour on the basis of current behaviour, I can supplement my prediction with a model of the kind of situation that I am observing, and how it normally plays out. This is distinct from mental state attribution or propositional attitude psychology in that it may not require anything other than a simple grasp of behavioural regularities, something that even non-human animals without a full-blown theory of mind seem to be able to do (see e.g. Rosati and Hare 2010).

When I see my colleague stand up from her desk and head empty-handed towards the fridge that stands in the corner of our office I can safely predict that she will open it and take something out. Of course, predictions of this kind only go so far — without any additional information I probably couldn't predict what she was going to get out of the fridge, although once I knew what she had got out I could probably predict what she was going to do with it. The additional information required to predict what she might get out is precisely what is provided by the other components of folk psychological discourse. For example, if I had seen her put some chocolate in there earlier in the day, and if she had just told me that she fancied a snack, I might be able to successfully predict not only that she would open the door, but also that she might take out the chocolate, break off a piece, and eat it. If I was further aware of her kindly nature, and that she knew I liked chocolate, I might predict that she would offer me some as well. So behavioural predictions that go beyond very simple and immediate circumstances seem to typically require further, non-behavioural information.

There is a thin line between behavioural predictions of this kind and more complex attributions of mental states as hidden causes of behaviour. The latter are what have typically been emphasised in previous philosophical discussions of folk psychology, normally under the more specific guise of propositional attitude attributions. It is important to distinguish between mental state attributions in general and the particular case of propositional attitude attributions, if only to make room for the in-principle possibility that there could be non-propositional mental states attributed by the folk. It is extremely natural to supplement behavioural descriptions with mentalistic language, to the point where not doing so can in fact feel somewhat artificial. Consider again my prediction of what my colleague will do when she stands up from her desk and walks towards the fridge. Based only on behavioural assumptions, I can predict that she will open it and take something out, and perhaps even predict what she will take out if I saw her put something there earlier, or if she only ever uses the fridge to store one item, but my predictions immediately become much more powerful if I have access to mentalistic data. Now I can predict that she will take out some chocolate and eat, because I know that she is hungry, and that she believes there to be some chocolate in the fridge. I can also predict that she will offer me some, because I know that she is kind, and I know that she knows that I am hungry. This is only a very simple case, but the complexity begins to increase rapidly as we add in details, especially once we get to recursive attributions ("I know that she knows that…").

Note that there are what appear to be two kinds of attributions in the above vignette, at least prior to further analysis. We have propositional attitude attributions: "she believes *x*" and "she knows *x*". We also have something like emotional or dispositional attributions: "she is hungry" and "she is kind". Whilst these can easily be reinterpreted as propositional attitude attributions — "she desires food" and "she wants to please others" — I think that doing so mischaracterises the nature of folk psychological discourse, as we typically interpret dispositional attributions as having a wider remit than proposi-

tional attitude attributions. "She is hungry" implies not only that she desires food, but also that she might be somewhat irritable, and that she might use food-related examples when making philosophical arguments, for instance.

Whilst I can predict my colleague's behaviour by engaging in crude behaviourism or by attributing mental states, it is often easier to simply situate her actions in an on-going narrative structure, one that I have built up over the weeks, months, and years that I have known her (see Bruner 1990, and Hutto 2008). This narrative allows for predictions, as if I am familiar with the narrative then I know what comes next, but it also provides a contextual justification for her behaviour. Narrative competency is distinct from mental state attribution because it does not rely on any particular theory of how the mind works, or even an explicit awareness of other people as mental agents at all.

Folk psychology can also serve as a normative or regulative discourse. Morton (Morton 1980) first articulated the suggestion that folk psychology might be partially normative, and McGeer (McGeer 2007), Zawidzki (Zawidzki 2013), and Andrews (Andrews 2015) have all explored it more recently. Folk psychological discourse can be considered to be playing a regulative role whenever it causes us to adjust our behaviour in some way. Zawidzki lists several different forms that this can take, "including imitation, pedagogy, norm cognition and enforcement, and language based regulative frameworks, like self- and group-constituting narratives" (Zawidzki 2013, p. 29). Note that 'mindshaping', as Zawidzki calls it, spans the whole range of folk psychological discourse, from "self- and group-constituting narratives" right down to the basic, perhaps pre-folk psychological, imitation of the behaviour of conspecifics. One particularly interesting case that Zawidzki explores in some detail is the way in which our explicit attributions of mental states to ourselves and to others ends up becoming self-fulfilling prophecies, as we consequently feel social pressure to conform to those attributions and thus to maintain at least a kind of surface level consistency (Matthews 2013, p. 111, makes a similar suggestion). This reversal of the usual way in which folk psychology is thought to operate, from mindreading to mindshaping, resembles the relationship between passive and active inference under predictive processing accounts. An explicit mental state attribution that leads to conformative behaviour could be seen as a kind of active inference, whilst folk psychology understood as mindreading is more akin to a form of passive inference.

Folk psychology in the more complex sense that I have described above is less obviously threatened by predictive processing. Regardless of whether or not folk psychological discourse paints a literally correct picture of how the brain works, it is self-evidently capable of predicting behaviour across a wide range of everyday situations, and it would continue to be able to do this even if the way in which it reached these predictions appealed to an implicit theory of cognition that turned out to be false. It is not even clear that folk psychological discourse is primarily in the business of explaining cognition, as the likes of Fodor and Churchland seem to take it to be. Rather it can be characterised as being primarily in the business of predicting behaviour under normal conditions, much as folk physics is able to make successful predictions under a range of everyday circumstances despite making several false assumptions and failing to accurately explain the dynamics of physical systems (cf. Churchland 1979).

There is a potential concern here, that folk psychology's predictive success might be sufficient reason to say that it does in fact give correct explanations of behaviour, and thus constitutes an accurate theory of cognition. According to some theorists in philosophy of science, the predictive success of a theory gives us good reason to be realists about that theory (see e.g. Putnam 1975; cf. Chakravartty 2015, sec. 2.1). If this were true then the predictive success of folk psychological discourse would be a reason to adopt a realist attitude towards it. I am sympathetic to this concern, but I think that it risks conflating two ways in which folk psychology could be successful. Rather than denying that folk psychological discourse describes real entities and processes, I want to deny that these entities and processes are of the sort that should serve an explanatory role in our scientific theories of cognition.

## 3.2 The Impact of Predictive Processing

With all that in mind, we can consider how well the predictive processing story aligns with folk psychological discourse in this more general sense. Friston and Frith 2015 have explored behaviour reading and prediction, understood in the context of Bayesian inference. They argue that predicting the behaviour of another requires synchrony between the two brains in question (that of the predictor and the predicted), allowing predictions to be made based entirely on the current state of one's own brain. This sounds somewhat like the simulation theory in social cognition (see e.g. Goldman 2006), which claims that we understand other minds by analogy with our own mind, although Friston and Frith 2015 do not make this connection. Given that predictive processing can be interpreted as generating a simulation of the target domain, this similarity is perhaps unsurprising, although there's also a sense in which the heavy emphasis on inference puts predictive processing closer to the theory-theory (see Ravenscroft 2010, sec. 2.1), which posits a literal theory of how minds work as the main mechanism for social cognition. One possibility here is that adopting the predictive processing framework would contribute to the development of a hybrid theory that includes elements of both theory-theory and simulation theory. (See Quadt 2017, for further discussion of predictive processing and social cognition.)

Narratives and social norms both seem to fit very comfortably into the predictive processing framework. Clark writes that individuals may "actively constrain their own behaviours so as to make themselves more easily predictable by other agents" (Clark 2016, p. 286), a suggestion that fits very neatly into the mindshaping account of social cognition presented by Zawidzki (Zawidzki 2013). Clark also suggests that personal narratives might "function as high-level elements in the models that structure our own self-predictions, and thus inform our own future actions and choices" (Clark 2016, p. 286). This is very close to the role envisioned for narratives in personal and social cognition by Hutto (Hutto 2008), and thus entirely consistent with my broader characterisation of folk psychological discourse. Hohwy (Hohwy 2013, p. 163) describes how difficult even simple behavioural predictions can be, and suggests that a failure to take into account broader contextual features might help explain the social cognitive deficit found in people with autism. So understanding folk psychological discourse within the predictive processing framework might involve telling a story about how high-level predictions of behaviour will involve complex models spanning not only individual agents but also the social and cultural environments that contribute to their behaviour.

The predictive processing account of cognition might even be reliant upon folk psychological narratives and social interaction more generally. Clark has written elsewhere about the importance of niche construction and cognitive scaffolding for human cognition (Clark 2008), and he devotes a chapter of his book on predictive processing to discussing these issues (Clark 2016, chapter 8). By conceiving of folk psychological discourse as a form of cognitive scaffolding we can retain an important space for it in our explanations of cognition, even if folk psychological explanations themselves are sometimes hard to reconcile with predictive processing. For example, by helping to regulate human behaviour via the enforcement of social norms, folk psychological discourse might serve as a form of active inference, changing the social environment so as to make it easier to predict. It also provides shared narratives that can help make sense of the behaviour of others, as well as exerting a regulative influence in their own right (see Andrews 2015). So even if propositional attitude psychology turns out to be a bad model of the cognitive architecture required for predictive processing, it might continue to be pragmatically useful to conceive of people as the kinds of systems that have beliefs and desires. We are then left with the further question of whether the failure of folk psychological discourse to match up precisely to our current best theories of cognition gives us reason to eliminate it, or whether being pragmatically useful is enough to escape this fate.

## 4 Revision Without Elimination

Predictive processing seems to be at least partially incompatible with the model of cognition given by propositional attitude psychology. The action-oriented nature of predictions means that it is hard to maintain a principled distinction between beliefs and desires, and the fine-grained content of some levels of the predictive processing hierarchy means that the content of those predictions will be hard to describe in everyday terms. One response to this partial incompatibility, originally offered by Lycan (Lycan 1988) in response to Stich's (Stich 1983) eliminativism, is that we should simply reinterpret folk psychological terms such as belief and desire in line with empirical discoveries. The motivation for this response is given by the causal theory of reference advocated by Putnam and Kripke, under which the reference of a term is determined by its initial use, rather by a definite description. So it would turn out that propositional attitude psychology was referring to fine-grained, action-oriented predictions all along, and that the folk were simply confused about what it was that they were trying to refer to when they attributed beliefs and desires to each other.

Without wanting to enter into the details of this debate, I am perfectly happy to concede that this is a viable option when presented with evidence of incompatibility between a folk theory and a scientific theory. However, as I have tried to argue, applying this argument to folk psychology misconstrues the nature of what the folk are trying to do in the first place. Propositional attitudes were never meant to refer to fine-grained mental states, but are instead intended to pick out and predict coarse-grained behavioural patterns and dispositions. I should emphasise that I am not advocating an anti-realism with regard to the posits of propositional attitude psychology — I think that whilst they are perfectly real, they just aren't the kind of thing that we should look for inside people's heads. This position is not so different from that advocated by Dennett 1989, and seems to more accurately capture the nature of our everyday understanding of other people.

Accepting that there is an incompatibility between folk psychology and cognitive science does not necessitate any radical change to our everyday folk psychological practices. Much of folk psychological discourse, such as behavioural predictions, normative constraints, and narrative competency, need not give an entirely accurate picture of the mechanisms of cognition in order to contribute to our understanding of other people. Whilst the literalist interpretation of folk psychology as propositional attitude psychology is at least partially incompatible with predictive processing, this is only a problem for those (such as Fodor) who insist that our theories of cognition should be structured along the same lines as propositional attitude psychology. Similarly, the eliminative materialism advocated by Churchland and others relies upon a commitment to folk psychology *either* accurately describing the structure of cognition *or* being eliminated. Without this dichotomy we can accept that our scientific usage of folk psychological terms should be revised without feeling at all compelled to eliminate it from everyday usage.

Nonetheless, it is plausible that a popularised version of predictive processing could partially modify our folk psychological self-conception, in ways that could be hard to predict ahead of time. Consider, for instance, the impact that the psychoanalytic notion of the unconscious had on the popular conception of the mind (Richards 2000). Prior to the 20th century the idea that our behaviours might be guided by unconscious motivations was not widely held at all, but now it has become common to refer to 'Freudian slips' and so on in everyday descriptions of behaviour. Similarly, if predictive processing catches on we might one day find ourselves referring to high-level predictions in order to explain why someone mistook a disguised mule for a zebra. This would be an entirely natural process that is (mostly) outside of the control of philosophers and cognitive scientists, so unless we thought that there was some ethical imperative to change the way that people think of themselves,[4] we should not let this concern us.

4 Michael L. Anderson suggested that someone like George Lakoff might think that we do have an ethical imperative of this sort – I am content to leave this as a question for the ethicists.

Quite aside from threatening to eliminate folk psychology, predictive processing presents us with an opportunity to revise our scientific usage of folk psychological concepts. We should aim to develop a novel conceptual taxonomy that more accurately reflects the structure of cognition and allows us to move beyond the limitations of folk psychological discourse. Understood in this way folk psychology could be used to identify interesting target phenomena and inspire scientific research (cf. Turner 2012), but should not be used as a source of technical cognitive scientific concepts. Some suggestions as to how to accomplish this transition can be found in recent literature on cognitive ontology revision (see e.g. Price and Friston 2005, Klein 2012, Poldrack 2006, Poldrack 2010, and Anderson 2014, Anderson 2015). Here the basic idea is that by comparing data from a large number of studies we can work out what the functionally relevant states and processes are, and design new terminology that better reflects these states and processes. In the case of predictive processing this might mean acknowledging the bidirectional nature of predictions, which in a sense serve both as beliefs and desires, and identifying the relevant differences between the way that these states interact with one another and the way that propositional attitudes are thought to interact with one another. I will not discuss the details of this proposal any further in this paper, although I do think that a concerted effort will be required in this regard if we are to move beyond the conceptual taxonomy offered by folk psychology.

## 5 Conclusion

Hohwy and Clark are both correct to suggest that folk psychology, interpreted as propositional attitude psychology, is likely to turn out to be incompatible with the account of cognition presented by predictive processing. As a consequence of this incompatibility the adoption of predictive processing will require the development of a novel conceptual taxonomy that more accurately describes the states and processes involved in prediction error minimisation. Whilst it might be okay to continue referring to predictions as 'beliefs' when one is clear that this is a technical usage, it would be better to use a new term so as to avoid confusion with the folk psychological sense of 'belief' (much as the term 'surprisal' has been adopted to avoid confusion with the common-sense understanding of 'surprise'). One option would be to use only the term 'predictions', although even here there is an implication of personal level agency that could perhaps be misleading. Better yet would be to coin an entirely new term that does not carry any unwanted associations.

Regardless of how we go about revising our conceptual taxonomy, folk psychology will not be threatened with elimination. This is because the traditional interpretation of folk psychology as propositional attitude psychology mischaracterises the nature of folk psychological discourse, which is actually a far more complex phenomenon consisting of behavioural predictions, narrative competency, and normative constraints, in addition to mental state attributions. Whilst a literalist reading of propositional attitude attributions is incompatible with predictive processing, there is no incompatibility between predictive processing and folk psychological discourse as a whole. If predictive processing turns out to be a successful theory of cognition (which is largely an empirical question), it may consequently have some impact on how we conceive of ourselves, and thus indirectly influence folk psychological discourse. The extent of this influence, however, is hard to predict, and in any case is not the sort of thing that philosophers or cognitive scientists should be in the business of attempting to regulate. We should content ourselves with ensuring that the language used to describe predictive processing is clear and accurate, and that the framework itself is coherent and supported by empirical evidence.

## References


Anderson, M. L. (2014). *After phrenology.* Cambridge, MA: MIT Press.

———(2015). Mining the brain for a new taxonomy of the mind. *Philosophy Compass, 10*(1), 68-77.

Andrews, K. (2015). The folk psychological spiral. *Southern Journal of Philosophy, 53*, 50-67.

Bruner, J. S. (1990). *Acts of meaning.* Cambridge, MA: Harvard University Press.

Chakravartty, A. (2015). Scientific realism. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*.

Churchland, P. M. (1979). *Scientific realism and the plasticity of mind.* Cambridge: Cambridge University Press.

———(1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy, 78*(2), 67-90.

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension.* New York: Oxford University Press.

——— (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(03), 181-204.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

Colombo, M. (2017). Social motivation in computational neuroscience: Or if brains are prediction machines then the Humean theory of motivation is false. In J. Kiverstein (Ed.) *Routledge handbook of philosophy of the social mind*. Abingdon, OX / New York, NY: Routledge.

Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting.* Cambridge, MA: MIT Press.

———(1989). *The intentional stance.* Cambridge, MA: MIT press.

———(2004). *Freedom evolves.* London: Penguin UK.

Fodor, J. A. (1975). *The language of thought.* Cambridge, MA: Harvard University Press.

Friston, K. & Frith, C. (2015). A duet for one. *Consciousness and Cognition, 36*, 390-405.

Friston, K. Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics, 104*(1-2), 137-160.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading.* New York: Oxford University Press.

Hobson, J. A. & Friston, K. J. (2014). Consciousness, dreams, and inference: The Cartesian theatre revisited. *Journal of Consciousness Studies, 21*(1-2), 6-32.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3*. https://dx.doi.org/10.3389/fpsyg.2012.00096.

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

Hutto, D. D. (2008). *Folk psychological narratives: The sociocultural basis of understanding reasons.* Cambridge, MA: MIT Press.

Klein, C. (2012). Cognitive ontology and region-versus network-oriented analyses. *Philosophy of Science, 79*(5), 952-960.

Lycan, W. G. (1988). *Judgement and justification.* Cambridge / New York / Melbourne: Cambridge University Press.

Matthews, R. J. (2013). Belief and belief's penumbra. *New essays on belief* (pp. 100-123). Springer.

McGeer, V. (2007). The regulative dimension of folk psychology. In D. D. Hutto & M. M. Ratcliffe (Eds.) *Folk psychology re-assessed* (pp. 137-156). Dordrecht, NL: Springer.

Millikan, R. (1996). Pushmi-pullyu representations. *Philosophical Perspectives, 9*, 185-200.

Morton, A. (1980). *Frames of mind: Constraints on the common-sense conception of the mental.* Oxford: Oxford University Press.

Pettigrew, R. (2015). Pluralism About Belief States. *Aristotelian Society Supplementary Volume, 89*(1), 187-204.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, 10*(2), 59-63.

———(2010). Mapping mental function to brain structure: How can cognitive neuroimaging succeed? *Perspectives on Psychological Science, 5*(6), 753-761.

Price, C. J. & Friston, K. J. (2005). Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology, 22*(3-4), 262-275.

Putnam, H. (1975). *Mathematics, matter, and method.* London / New York: Cambridge University Press.

Quadt, L. (2017). Action-oriented predictive processing and social cognition. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Ratcliffe, M. (2006). 'Folk psychology' is not folk psychology. *Phenomenology and the Cognitive Sciences, 5*, 31-52.

Ravenscroft, I. (2010). Folk psychology as a theory. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*.

Richards, G. (2000). Britain on the couch: The popularization of psychoanalysis in Britain 1918 – 1940. *Science in Context, 13*(02), 183-230.

Rosati, A. G. & Hare, B. (2010). Social cognition: From behavior-reading to mind-reading. In G. F. Koob, M. Le Moal & R. F. Thompson (Eds.) *Encyclopedia of behavioral neuroscience*. London, UK / Burlington, MA / San Diego, CA: Academic Press.

Spratling, M. W. (in press). A review of predictive coding algorithms. *Brain and Cognition*. https://dx.doi.org/10.1016/j.bandc.2015.11.003.

Stich, S. P. (1983). *From folk psychology to cognitive science: The case against belief.* Cambridge, MA: MIT Press.

Turner, R. (2012). The need for systematic ethnopsychology: The ontological status of mentalistic terminology. *Anthropological Theory, 12*(1), 29-42.

Yalowitz, S. (2014). Anomalous monism. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*.

Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition.* Cambridge, MA: MIT Press.

# Moderate Predictive Processing

Krzysztof Dolega

Recent developments in the predictive processing literature have led to the emergence of two opposing positions regarding the representational commitments of the framework (Hohwy 2013; Clark 2015; Gładziejewski 2016; Orlandi 2015). Proponents of the *conservative* approach to predictive processing claim that the explanatory power of the framework comes from postulating a rich nesting of genuine representational structures which can account for many cognitive functions (Gładziejewski 2016). Supporters of the more *radical* interpretation of predictive processing, on the other hand, postulate that not all elements of the computational architecture should be interpreted as full-blown representations, stressing the framework's connection to ecological and embodied approaches to cognition instead (Clark 2015, Orlandi 2015). Surprisingly, despite defending opposing positions, both camps seem to adopt William Ramsey's representational 'job description challenge' (Ramsey 2007) as a standard for genuine ascriptions of representational function.

The aim of this paper is to evaluate competing approaches and show that both sides of the debate must overcome additional challenges with regard to determining the representational commitments of the predictive processing framework. Following the discussion of the opposing views and the way they employ the representational job description in their arguments, I raise a worry that Ramsey's criterion may be ill suited for establishing a strong distinction between them. In light of this I propose to frame the debate between the two camps as a disagreement over the contents of generative models, rather than the functional roles fulfilled by the structures under investigation. Finally, presented with the problem of content determination and the lack of framework constrains strong enough to help resolve the disagreement, I call for moderation in making claims about predictive processing's representational status.

**Keywords**

Conservative predictive processing | Internal models | Markov blanket | Predictive coding | Radical predictive processing | Representational job description | Representations | Structural representations

Probabilistic modeling of perception and cognition is quickly becoming the leading trend in cognitive psychology and neuroscience. Hierarchical predictive coding or processing (henceforth predictive processing or PP for short) is one of the more prominent and widely discussed frameworks emerging from the recent developments in applying statistical methods to machine learning and computational neuroscience (Rao and Ballard 1999; Friston 2008; Friston 2010; Hohwy 2013; Clark 2013; Clark 2015; Clark 2016).

The framework is an attempt at formulating a unified explanation of the processes underlying perception, cognition, and action, by postulating that the brain's neural populations are organized into multiple hierarchies performing cascading statistical inferences in order to predict future inputs. This is done through the workings of an internal generative model, which aims "to capture the statistical structure of some set of observed inputs by tracking (one might say, by schematically recapitulating) the causal matrix responsible for that very structure" (Clark 2013, p. 185). The model is used to generate hypotheses (predictions or expectations) about the states of the external world and the corresponding activations of sensory peripheries, which are then tested against the actual states of the sensorium, driving perception and action in a top-down manner.

Dolega, K. (2017). Moderate Predictive Processing.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 10. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573116

In this paper I discuss two leading positions regarding the representational commitments of the predictive processing approach to cognition. So called conservative predictive processing (cPP) (Clark 2015), claims that the explanatory power of the framework comes from postulating a rich nesting of genuine representational structures which come to serve as a model of the organism's external and internal milieu (Hohwy 2016; Gładziejewski 2016). Radical predictive processing (rPP), on the other hand, postulates that not all elements of the computational architecture should be interpreted as full-blown representations (Clark 2015). Instead, proponents of this approach argue that the relevant computational level descriptions (Marr 1982) serve merely as abstract schemata aimed at capturing the dynamics of processes, which are embedded into neural structures by evolutionary selection (Orlandi 2013; Orlandi 2014; Orlandi 2015) and do not depend on manipulating genuine representations (Clark 2016; Downey 2017; Bruineberg 2017).

Surprisingly, both approaches are motivated by the adoption of William Ramsey's representational 'job description challenge' (Ramsey 2007), according to which appeals to representational posits must be supported by a convincing demonstration of the relevant elements or structures playing a representational role within the functioning of a wider cognitive system. Gładziejewski follows Ramsey in order to show that PP's generative models perform a representational function by acting as detachable, information bearing structures which stand-in for the features of the environment, in order to enable capacities such as action-guidance and error detection. Clark and Orlandi, on the other hand, claim that only higher levels of the PP hierarchy exhibit such capacities, while lower (e.g. perceptual) levels should be construed as *model-free* structures governed by biases acquired through reinforcement learning or phylogenetic development.

The main aim of this paper is to evaluate the competing approaches in relation to the underlying computational architecture and show that the disagreement between the two sides of the debate does not cut as deep as it has been presented in the literature. The main reason for this is that both sides agree on how to understand Ramsey's representational challenge and the functionally distinct notions of representation it introduces. However, an investigation into the roles played by the structures posited by PP reveals that the job description challenge may be insufficient to differentiate the competing views. Although it does not support the radical claim that peripheral layers of PP systems consist in solely non-representational elements, it also poses a serious problem for the conservative side of the debate by inviting ambiguities with regard to the ascription of representational function to more nested structures. Proponents of cPP are tasked with distinguishing internal models targeting external, environmental features from structures functioning as meta-representations modelling the behavior of other parts of the system. I propose that the difference between the competing positions must come down to the question about the content of PP's internal models. Supporters of rPP are faced with the task of introducing additional criteria which would help distinguish their position from the conservative one. Members of the cPP camp, on the other hand, can secure their interpretation by providing conditions for determining the contents of representational and meta-representational elements. Still, this is not an easy task due to the informational encapsulation of different layers of the system and the unclear conditions for identifying cases of misrepresentation.

Because this paper is part of a larger collection beginning with a primer aimed at elucidating the main tenets of PP to an uninitiated audience (Wiese and Metzinger 2017), I will skip a typical introduction to the framework. Instead, I will start by articulating Ramsey's representational job description requirement for non-vacuous ascription of representational function (section 1), followed by a brief presentation of the two competing interpretations of the framework (section 2). Having delineated the available positions, I will move on to evaluate which of the elements posited by PP should be the target for the debate over the framework's representational status (section 3.1). From there, I will argue that the representational job description challenge does not offer sufficient ground for distinguishing rPP from its conservative counterpart, by showing that it fails to secure a non-representational interpretation of the system's peripheral layers (section 3.2). This, however, does not mean that cPP cannot be

contested, as proponents of this reading must face the opposite problem of functional indeterminacy in models removed from the sensory periphery (section 3.3). While, in principle, it is possible to resolve these issues by appealing to the contents of such posited representations, in practice this solution faces further difficulties, relating to a lack of clear conditions for content determination (section 3.4). I close the paper with a call for moderation in making claims regarding PP's representational status, and point to two strategies for solving the problem at the heart of the debate (section 4).

## 1 William Ramsey on the Confusion About the two Aspects of Mental Representations

The representational job description challenge (Ramsey 2007) is central to much of the recent literature concerning the representational status of probabilistic approaches to cognition. The aim of the job description challenge is simple — to provide a condition for genuinely explanatory ascriptions of representational function in cognitive science. Ramsey motivates the need for such a criterion by pointing out that our notion of mental representation must somehow be rooted in the everyday conception of what a representation is (otherwise calling such theoretical posits 'representations' would make no sense), but that our common usage fails to provide us with a clear grasp of how to individuate such entities or how to identify the role they should play in naturalistic theories of cognition. Thus, the goal is to formulate a condition that yields a scientifically valuable, yet intuitively recognizable notion of representation, one which will offer an explanation of how, or in virtue of what, particular posits function as representations. However, it is important to stress that it is not a challenge to define or describe what content is in naturalistic terms.[1]

According to Ramsey, there has been a long standing confusion between "understanding how a physical structure actually functions as a representation" as opposed to "[…] understanding the nature of the relationship in virtue of which it represents one sort of thing and not something else" (Ramsey 2016, p. 5). To clarify this confusion, he proposes to distinguish two dimensions, or aspects, of mental representation, both of which have to be fulfilled "for something to actually qualify as a full blown representation […]" (p. 6). The representational job description is set out to elucidate the first dimension — one of a representation's functional role, by providing a "set of conditions that make it the case that something is functioning as a representational state. In other words, it is the set of relations or properties that bestow upon some structure the role of representing" (p. 4). The other aspect of representations concerns *what* such structures or states represent. This dimension can be understood as "[…] a set of relations or features that bestow upon some representational structure its specific representational content" (p. 4). With this distinction in place, I will now focus on the question of what the functional dimension challenge exactly *is*.

For Ramsey, the 'job' of a representational state or entity is to *stand-in* for something external to the wider consumer system in which it is employed. Therefore, the challenge for any scientific theory is to explain how particular posits fulfill the functional role of standing-in for external features of the world within that theory or model.[2] Although this account of representational function may seem simple, it proves to be a powerful tool for assessing applications of the notion within cognitive science and philosophy.

As Gładziejewski (Gładziejewski 2016) points out, the way Ramsey develops his challenge relies heavily (though not exclusively, see section 4) on an argumentative strategy of 'comparing-to-prototype' — one examines the everyday, pre-theoretical uses of the notion in order to find a widely ac-

1 In a similar manner, the central issue of Ramsey's investigation should not be confused with the question about the possibility of describing a purely physical process as intentional (this would be a question about the possibility of adopting the intentional stance — Dennett 1987), nor should it be confused with the question concerning the indispensability of intentional ascriptions (since, following Dennett, every physical system can be described in purely causal/structural terms — a description or stance which may be more relevant for some tasks).

2 It should be noted that this is only a sufficient condition for ascriptions of the representational role to cognitive systems. See Gładziejewski's discussion of the challenge for more detail.

cepted and uncontroversial application, which then serves as a prototype to which particular uses of that notion in scientific literature are compared. For Ramsey, cartographical maps serve as an intuitive prototype for ascriptions of representational function. Following his lead, Gładziejewski (Gładziejewski 2015, Gładziejewski 2016) sets out to distinguish several features of all map and map-like devices (e.g. GPS navigation), which make them suitable for fulfilling the representational role on Ramsey's account. Firstly, maps are deemed useful in virtue of the relation (a mapping function — see footnote 3) obtaining between the physical properties or features of the map and the properties of the environment which is its representational target. It is this relation that allows for the physical artifact to stand-in for the environment during navigation and action planning, even when the depicted location is not immediately present to the user (detachability). The mapping relation also makes it possible for the map user to control her performance and adjust her actions in accordance with the available information (action guidance). Finally, it also accounts for the possibility of the representational item failing to fulfill its role, e.g. in cases of low fidelity or error in the map's structure, therefore allowing for misrepresentation and the presence of user detectable error.

Having established the prototypical case for non-vacuous representational ascription, Ramsey proceeds with the second step of his argumentative strategy — comparing different scientific applications of the concept to that prototype. It is here that his strategy proves its mettle by offering interesting and unexpected results, challenging some of the established applications of the notion of representation in philosophy and cognitive science. For the sake of brevity, I will focus on just two key examples.

The so-called 'cognitive maps' discovered in the rat's hippocampus (O'Keefe and Dostrovsky 1971; O'Keefe and Nadel 1978) and entorhinal cortex (Hafting et al. 2005) are among the favorite study cases of the proponents of the job description challenge (see e.g. Miłkowski 2015). The view that the activations of place cell neurons encode information about the two-dimensional structure of a rat's environment has been largely accepted by the scientific community, earning the scientist who made the discovery a Nobel prize. This claim has been strengthened by recent findings showing that the same neural mechanism is also processing information about the animal's future location in the environment, thus pointing to the place cells' involvement in action anticipation and planning (Van der Meer and Redish 2010; Gupta et al. 2013). Together, these developments suggest that such 'maps' are the right kind of structures to pass the representational job description challenge, as they not only have elements that correspond to features of the world, but also allow for exploiting this relation in action guidance and error correction (Pfeiffer and Foster 2013).

The notion of representation which emerges from the job description challenge and the discussion of cognitive maps is one where representation can act as a model of the world in virtue of having an internal structure which maps onto the structure of the world[3] — a notion of structural representation or S-representation.[4] As Ramsey elaborates:

> What S-representation has going for it […] is a distinctive role within a cognitive system that is recognizably representational in nature and where the state's content is relevant to that role. […] With S-representation, the fact that a given state stands for something else explains how it functions as part of a model or simulation, which in turn explains how the system performs a given cognitive task (Ramsey 2007, p. 126).

3 Importantly, this does *not* mean that the target properties must be represented by the same kind of properties in the mental representation, i.e. that the representation must be first-order isomorphic with its target. Structural representations, in the sense discussed above, are in fact standing in a second-order, functionally isomorphic relation to their representational targets (see Palmer 1978, for a detailed discussion), meaning that, for example, distance can be represented by frequency, frequency by magnitude, etc. Similarly, the analogy with cartographical maps (which *are* first-order representations) does not mean that structural mental representation must be static and cannot change in accordance with changes in the world or in the needs of the consumer system. See, for example, Wiese 2016, for a discussion of Rick Grush's emulator representations (Grush 2004) as structural representations (see also Bartels 2005, Ch. 3.).

4 Note that S-representation is not the only notion of representation that passes the representational job description challenge, see also the discussion of Wiese 2016, in section 3.4.

However, not all established uses of the notion of representation fare equally well when confronted with Ramsey's challenge. On the tradition of information theoretic accounts (most notably Dretske 1981), which were later supplemented with teleological accounts of function (e.g. Millikan 1984; Dretske 1988), biological mechanisms can be considered representational when they have a function of responding to certain environmental conditions and there is a lawful-like dependence between the signaling of the system and the behavior of the organism. This account allows for error and misrepresentation because such indicator systems can be triggered by environmental factors or properties different from the ones they are supposed to react to. It also "[…] provides Dretske with a way of showing how informational content can be explanatorily relevant. Structures are recruited as causes of motor output because they indicate certain conditions" (Ramsey 2007, p. 130). Despite its appeal, the notion of representation employed by teleofunctionalists does not meet the representational job description challenge. Ramsey points out that "there are several non-representational internal states that must, in their proper functioning, reliably respond to various states of the world […]". For example, the immune system reacts to infections, "yet no one suggests that any given immunological response (such as the production of antibodies) has the functional role of representing these infections. While nomic dependency may be an important element of a workable concept of representation, it clearly is not, by itself, sufficient to warrant viewing an internal state as a representation" (Ramsey 2007, p. 125).

Thus, Ramsey argues the notion of *detector* or *indicator* representations employed by Dretske and his followers, when treated as a criterion for ascribing the functional role for representations, threatens to trivialize the representational theory of mind. Instead he proposes (Ramsey 2016) to treat the teleological project as an account of representational content. He notes that:

> […] there is a notorious problem of content indeterminacy for any account of representation based upon structural similarity. The problem is that isomorphisms are cheap — any given map or model is going to be structurally similar to a very wide range of different things. While elements of models may function as representational proxies during various sorts of cognitive operations, exactly what they represent is impossible to determine by merely focusing on the 'structural' properties of the model or map itself (Ramsey 2016, p. 8).

This means that, even though structural properties are sufficient to determine whether a state or part of a system *could* function as a representation, a mere mapping relation is not enough to determine whether or not it carries any content. Ramsey suggests that, in order to postulate that theoretical posits are, in fact, full blown representations, constraints additional to the representational job description challenge must be presented and fulfilled.

## 2 Two Flavours of PP

The distinction between *conservative* and *radical* predictive processing has recently been introduced in to the literature by Clark (Clark 2015), who aims to distinguish between two different approaches to the framework's philosophical and scientific significance.

### 2.1 Conservative Predictive Processing

According to the conservative understanding, PP is similar to *reconstructive* views of perception, which "[…] depict our cognitive contact with the world as rooted in a kind of neuronally-encoded rich inner recapitulation of an observer-independent reality" (Clark 2015, p. 12). In this take on the framework, the cognitive system is reliant on an inner model that encodes the structure of the complex relationship between stimulations of the sensory peripheries and their distal causes, effectively becoming what Hohwy has called "an internal mirror of nature" (Hohwy 2013, p. 220). Clark describes the inner models of cPP as being employed by the system "[…] to stand-in for the external world for

the purposes of planning, reasoning, and the guidance of action" (Clark 2015, p. 12). This is similar to Gładziejewski's defense of the representational reading of predictive processing as relying on S-representations.

In his 2016 article, Gładziejewski employs Ramsey's compare-to-prototype strategy in order to argue that the generative models employed in architectures postulated by PP play a role akin to probabilistic maps of the world. By analyzing the role such models must fulfill in order to coordinate perception and behavior in ways that will minimize prediction error and keep the organism within homeostatic bounds, he comes to the conclusion that parts of the PP system responsible for generation of predictions must function as information bearing structures with features which correspond to the structure of the external environment.

To support this claim, Gładziejewski imagines a toy example of a two level PP system, consisting of a sensory periphery and a generative model. Because the system is assumed to approximate Bayesian reasoning, the prior probabilities of certain environmental occurrences taking place (e.g. of encountering a particular kind of object) should be stored in the generative model. Gładziejewski points out that the model in question must encode a set of, so called, 'hidden' or 'latent' variables which act as parameters for generating predictions about the lower stage of the system by corresponding "[…] to different likelihoods of potential patterns of activity at the lower sensory level" (Gładziejewski 2016, p. 571). Moreover, because "[…] the hidden variables are not only related to lower-level, sensory patterns, but to each other (intra-level) as well, their […] values evolve over time in mutually-interdependent ways […]" (Gładziejewski 2016, p. 572), which allows them to have a structure mirroring the dynamics of their target domain.

Together, these properties allow the PP structures to fulfill the functional role of genuine representations which work as detachable models of the environment during action guidance and error detection. Such generative models must be detachable since, on this framework, action involves an off-line computation of multiple possible strategies for minimizing error (see also section 3.1 and 3.2). This process, in turn, leads to a deployment of predictions which can drive behavior. Finally, predictions which are sub-optimal for a given context produce error signals that can be corrected in the next cycle of hypothesis testing. All this leads Gładziejewski to conclude that generative models qualify as S-representations and that the framework belongs to the tradition of computational-representational theories of mind.

## 2.2 Radical Predictive Processing

Proponents of the radical approach are not opposed to the claim that *some* levels of the probabilistic architecture do fulfill a genuinely representational role. They are, however, against viewing the framework as another iteration of the computational-representational theory of mind by stressing the embodied and embedded nature of many cognitive processes.

Once again, this is most visible in Clark's 2015 article where he claims that predictive processing should not be construed as being solely dependent on internal models of the environment. Although rPP proponents do not aim to deny that higher cognitive functions (e.g. abstract reasoning, planning, language) most likely depend on manipulating some kind of mental representations, they do stress that the appeal of the PP framework lies in reconciling such structures with 'fast and frugal' heuristics for action which emerge from the agent's dynamical coupling with the environment. Thus, the radical version of the framework is meant to highlight that "[…] sensing delivers an action-based grip upon the world, rather than a rich reconstruction apt for detached reasoning […]" (Clark 2015, p. 15). Similarly, Clark opposes conceptualizing successful behavior only as an "[…] outcome of reasoning defined over a kind of inner replica of the external world" and prefers to see it as an "[…] outcome of perception/action cycles that operate by keeping sensory stimulations within certain bounds" (Clark 2015, p. 15).

In his 2016 book, Clark adds detail to the rPP position by appealing to the distinction between 'model-based' and 'model-free' strategies for behavior selection and guidance (Dayan 2012; Dayan and Daw 2008; Wolpert et al. 2003). On this picture, which resembles the distinction present in dual-system/process literature (Frankish 2010), 'model-based' reasoning involves "[…] the acquisition and the (computationally challenging) deployment of fairly rich bodies of information concerning the structure of the task-domain, while 'model-free' approaches […] implement pre-computed 'policies' that associate actions directly with rewards, and that typically exploit simple cues and regularities while nonetheless delivering fluent, often rapid, response" (Clark 2016, p. 252). When applied to PP, this distinction boils down to the difference between the kind of posits defended by Gładziejewski, as capable of simulating and comparing multiple action plans, and reflex-like responses acquired through reinforcement learning.

In clarifying the relationship between these two kinds of processes, Clark follows (Daw et al. 2011) and suggests that different strategies can be flexibly combined together in accordance with contextual information. Acquisition and deployment of more rigid routines can be guided by information rich internal models, such as the expected precision optimization scheme associated with attention (Feldman and Friston 2010; Hohwy 2012). In such a case, "[…] a kind of meta-model (one rich in precision expectations) would be used to determine and deploy what ever resource is best in the current situation […]" (Clark 2016, p. 253).

Unfortunately, Clark remains vague about the criterion by which genuine representational processes are to be differentiated from the 'fast and frugal' ones. He stipulates that model-free processing should be associated with larger reliance on the bottom-up information, while model-dependent one would rely on top-down influence of prior knowledge. This suggests a kind of gradation from response based processes to genuinely representational ones. Unfortunately, this interesting proposal remains underdeveloped. By claiming that some, but not all, PP structures act as representations, Clark raises a challenge for proponents of the radical interpretation, who are now charged with presenting a criterion by which representational structures can be distinguished from non-representational ones. Failing to provide such a criterion, threatens the rPP view with collapsing into the conservative one.

This issue has recently been taken up by Nico Orlandi (Orlandi 2013; Orlandi 2014; Orlandi 2015), who builds on Clark's proposal by advocating the use of Ramsey's representational job description challenge for the purpose of demarcating embodied and embedded processes from representational ones. Though her focus is placed mostly on the long-standing disagreement between the ecological (Gibson 1979) and inferential (reconstructive) theories of vision (Gregory 1980; Marr 1982; Friston et al. 2012), the representational status of probabilistic models of perception occupies a prominent place in her treatment of that debate.

The main claim of Orlandi's *embedded seeing* (Orlandi 2013; Orlandi 2014) project is that vision is not an inferential process relying on manipulating intermediate states or tokens that qualify as representations. Rather, vision is a process embedded in to the biological structure of the organisms' visual apparatus, which has been molded to reliably respond to the presence of certain environmental properties through evolution and development. Computational theories of vision are merely re-descriptions aimed at capturing the dynamics of causal interactions between the physical elements of the visual system. As Orlandi explains, what follows from her Gibsonian assumptions is that probabilistic theories of vision mistakenly re-describe "[…] biased processes that operate over non-representational states" (Orlandi 2015, p. 1) as inferential. She further clarifies that on her interpretation "priors and likelihoods rather look like built-in or evolved causal intermediaries of perception that incline visual systems toward certain neuronal configurations (or certain 'hypotheses')" (Orlandi 2015, p. 25).

Orlandi motivates her radical departure from the probabilistic orthodoxy in several ways (Orlandi 2015), at least two of which seem to be relevant for understanding what her position is and how it relates to Clark's. The crux of her argument relies on questioning the restricted role of bottom-up inputs in PP schemes. By presenting a simple example of an ambiguous stimulus, such as a circle which

can appear convex or concave depending on the assumed position of the light source, she builds the case that that priors and hyper-priors alone are unable to restrict the hypothesis space of probabilistic models of perception to a single prediction. Hyper-priors, here understood as domain-general assumptions guiding the acquisition and deployment of more domain-specific hypotheses, are presented as too general (Orlandi 2015, p. 9). An appeal to a likelihood function (the probability of evidence — here the inverse of error — given the hypothesis) is similarly ruled out as, according to Orlandi, it would not restrict the hypotheses space enough to favor a single prior. This reasoning leads her to conclude that it is the sensory signal which drives and constrains the system by conveying information about the statistics of natural scenes, activating or pre-selecting the "[…] priors that should be employed even in contexts of high noise where the signal is compatible with multiple hypotheses" (p.10).

An important caveat to the above points is that they do not line up with most of the empirical evidence supporting PP's success in modeling perceptual and cognitive phenomena (see e.g. Rao and Ballard 1999; Spratling 2016).[5] Orlandi fails to appreciate the fact that many versions of the framework assume that the system's internal states come to mimic the statistics and dynamics of the environment (see also section 3.4), precisely for the purpose of disambiguating stimuli in conditions where the sensory signal is corrupted by noise (Hohwy 2012). By not paying attention to this crucial assumption, Orlandi effectively rejects the problem of perceptual inference, echoing Clark's idea that 'model-free' strategies are heavily reliant on prediction errors, which can act as a constraint on the process of hypothesis selection.

What is crucial for the present treatment is that Orlandi uses these points to motivate distancing her position from Gładziejewski's. Interestingly, she applies the same criterion he has previously used to defend the representational status of PP's generative models. According to her the early stages of the visual system are best conceived of as "mere detectors or mediators" (Orlandi 2015, pp. 23-24), meaning that they do not pass Ramsey's representational job description challenge. For example, she argues that levels directly involved in predictive coding of information about the states of the retina (e.g. cells in primary visual cortex sensitive to discontinuities in patterns of retinal stimulation traditionally associated with edge perception) are not sufficiently detached from their inputs and do not play a robust role in driving and guiding action.[6] She defends the position that, in the case of such low levels, neither predictions, nor error signals seem to have the function of carrying information about things external to the system. Errors convey only information regarding "[…] the need to adjust its own states to reach an error-free equilibrium", while predictions "are states produced for checking the level below them", which "exhausts their function" (Orlandi 2015, pp. 23-24). Following this reasoning, Orlandi claims that only the outputs of the whole visual system can pass the representational job description challenge, because they are the only part of the visual system that can be said to play the role of standing-in for the world in consumer systems realizing higher cognitive functions. This and other claims held by the rPP camp, however, may not be supported by the underlying assumptions of the framework.

5 Admittedly, Orlandi draws attention to the fact that PP models differ from the traditional constructivist views of perception in that they do not postulate early and intermediate visual states to explicitly track well-defined elements (Jehee and Ballard 2009). This point, however, does not undermine a representational reading of PP in any way. PP is a probabilistic modeling framework and it is consistent with the idea that, for example, lower-levels of the system track different targets, depending on the predictions they receive from the higher layers of the hierarchy. Orlandi fails to appreciate such a possibility as she does not engage with a large body of empirical work on top-down modulation in early visual areas (e.g., Petro et al. 2014). Nothing in the present article hinges on this point and it will not be discussed further.

6 The editors have pointed out the possibility of augmenting the rPP view by developing a graded account of representation in which the degree to which something serves as a model or a representation is, at least partially, determined by the degree of its detachment from the target. This is an intriguing suggestion, but it faces significant obstacles which deserve a separate full length treatment. For example, one of the requirements for such a view is to provide a set of conditions that would allow for defining the degree of detachability independently from the level of analysis at which the system is decomposed. Moreover, I would like to point out that this would not absolve the proponents of rPP from the task of providing a clear and unambiguous distinction between the representational and non-representational ends of the spectrum.

## 3 Three Obstacles on the Way to Determining PPs Representational Status

By now the tension between the conservative and radical treatments of PP should be visible. What is of special interest here is that the conflict at hand stems from the employment of the same conceptual tool — the representational job description challenge, in order to defend opposite positions regarding the representational commitments of the framework. One way to resolve the qualm between cPP and rPP would be to show that one of the sides misunderstands or misconstrues Ramsey's challenge. However, little in the discussion of these positions suggests such a solution, especially since the members of both camps seem to agree on how the distinction between S-representations and detector representations should be understood. Rather, it seems that the point of contention consists either in the way the challenge is applied to PP (i.e. which parts of the frameworks' computational description are submitted to Ramsey's test), or in some disagreement independent from the issue of the representations' functional role, such as the kind of contents that the relevant parts of the PP system trade in (e.g. rich or poor).[7]

In what follows, I begin by focusing on the functional dimension of PP, starting with the problem of clarifying what structures pass the representational job description. However, as the discussion progresses I will show that Ramsey's challenge is not sufficient for establishing the two positions as competing alternatives. The difference between these two views must come down to the question about the content of posited structures, rather than the question about their functional role.

### 3.1 Which Elements of PP Fulfill the Representational Job Description?

From the discussion in section 2 it can be seen that each side of the debate puts stress on different aspects of PP's computational architecture, opening possibilities for miscommunication. Clark and Gładziejewski focus on the functional role played by internal models, but it is not always clear which parts of system constitute a model and whether or not such models are coextensive with levels. Friston and Hohwy apply the notion quite loosely, easily changing between talking about a single, over-arching model of the world spanning the whole predictive hierarchy and multiple, restricted models of different cognitive domains. Finally, Orlandi seems to be preoccupied with the representational status of top-down and bottom-up messaging pathways communicating priors and errors respectively. This is why it is important to clarify which parts of the PP system are the target of the discussion.

As has been mentioned in the introduction and the supplied exposition of PP (Wiese and Metzinger 2017), the framework posits probabilistic systems organized in a hierarchical manner. Since these architectures are supposed to model the behavior of brain structures, they are usually implemented as artificial neural networks. Although competing implementations may have different assumptions about the wiring and number of functionally distinct types of nodes (compare e.g. Rao and Ballard 1999, with Spratling 2008), the general blueprint is shared by all versions of PP. It is the schema of a 'stacked' hierarchy with two distinct feed-back and feed-forward passageways connecting 'prediction estimation' (PE) groups which are assumed to correspond to specific cortical regions (Spratling forthcoming). It is these PEs that are usually referred to as 'levels' in the subject literature, and are assumed to encode the parameters of predictive models.

The PE levels are collections of different types of units themselves, including ones encoding the hidden variables used for generating predictions. Notably, when describing the make-up of these modules, Rao and Ballard include the input and output units as their components. Each level in the

7 It is also possible that the disagreement is about issues orthogonal to the problem of representation altogether. For example, Hohwy and Clark seem to disagree about the normative commitments of the framework. The first author is genuinely interested in an epistemic interpretation of PP, according to which its Bayesian roots offer a cognitive system aimed at truth, whereas the latter views the system as aimed at ecologically and evolutionarily bounded optimality. In the present treatment I am steering away from this debate for several reasons: a) the relationship between the cognitive (PP) and normative (Bayesian rationality) components of the framework is not entirely clear (see the treatment of Fink and Zednik 2017, for some ideas about that); and b) the issues discussed in this paper do not hinge on whether one assumes that the representations manipulated by the PP system are truth-capturing or just truth-approximating.

hierarchy (let's call it the *n-th* level in an ordering starting from the lowest to the highest level) is specified as receiving two kinds of inputs. The first are the descending predictions from the level above or upstream (*n + 1*), which constrain the activity on the level in question (*n*) by fulfilling a role which is equivalent to that of priors in Bayesian inference. The second kind of input received by any level is the bottom-up prediction error signal, which carries the information about the difference between the actual and predicted state of the level below or downstream (*n – 1*). Each of the PE levels is also producing two kinds of output: constraining predictions about the expected activity on the lower level (*n – 1*), and an unresolved residue of the error signal fed to the level above (*n + 1*).[8]

It is possible to restrict the discussion of PP's representational commitment to the question about the functioning of particular units. However, this would not be productive, as placing too much focus on the components of PEs can lead to a false conclusion about the functioning of the wider architecture. First of all, network implementations of PP do not differ from other artificial neural nets, which have been extensively discussed as being composed of simple detector-like elements rather than full blown representations (Ramsey 2007, p. 145). Since the nodes encoding causes of the lower level patterns of activation must reliably react to incoming inputs as well as be able to trigger appropriate activations in populations which communicate predictions, it is possible to label them as mere causal relays. Orlandi is correct in arguing that, on their own, error and prediction units do not fulfill the representational job description challenge. It does not, however, mean that more complex, representational systems cannot be constructed out of such simple elements. For example, the employment of S-representations in explanations of cognitive maps' functioning is not threatened by such maps being composed of detector-like elements. The fact that the firing rates of place cell neurons co-vary with the rat's location does not undermine the functioning of the whole structure as a detachable representation employed in controlling behavior and navigation.

Somewhat similarly, the success of certain connectionist architectures employing generative models stems exactly from the organization of their simple elements (Hinton 2007). These systems exhibit complex behavior, which cannot be fully accounted for by an appeal to mere biasing. The aptly named Helmholtz machine (Hinton et al. 1995), for example, is said to construct a generative model of its own input in a manner analogous[9] to the PP systems — i.e. by approximating Bayesian inference and learning the complex statistical regularities in a current data set in order to predict possible future inputs of the same kind. By using the so called 'wake-sleep' algorithm (Dayan et al. 1995), it operates in two alternating modes — processing inputs in a bottom-up manner during the wake phase, and optimizing its internal dynamics by a top-down generation of possible inputs in the off-line 'fantasy' phase. While it is possible to give a purely causal, adaptationist account of a passive learning process in connectionist models (Ramsey 1997), the ability to generate inputs off-line for the purpose of adjusting the internal parameters and optimizing performance seems to go beyond such simple accounts. It implies that the system can not only recapitulate the internal structure of the target domain, but also deploy it in a manner which is detached from the inputs, additionally assessing and correcting its own performance.

A non-representational account falls short in cases where top-levels of a hierarchical architecture come to encode different models of regularities present in its input sets, which are then used to guide lower levels to produce outputs that are best at capturing the structure of these input sets. A good

8 In most hierarchical predictive coding algorithms (e.g. Rao and Ballard 1999; Friston 2008) this means that the bottom-up signal communicates only the difference between the actual and predicted states of the target level, ignoring the redundant information. However, it is worth noting that in Spratling's PC/BC-DIM (Predictive Coding/Biased Competition using Divisive Input Modulation) algorithm, due to a different mapping of levels onto cortical regions, bottom-up signal carries estimates about the predicted input's causes (see also footnote 5 and Spratling forthcoming, p.5).

9 It is important to note that, although both approaches depend on the use of generative models, there are some significant differences between them. Clark 2016, p.309, stresses that Hinton's algorithms employ self-generated prediction during learning and/or optimization, not during online processing. It is also significant that Hinton's work represents the connectionist tradition of back propagation algorithms which, despite belonging to the same wider category of recurrent networks as predictive coding algorithms, can differ significantly in computational detail (even if both can be seen as approximating Bayesian inference). However, there is a growing interest in bringing these two approaches together, see e.g. Rezende et al. 2014, and Domingos 2015.

example of this is Charles Kemp & Joshua Tenenbaum's unsupervised, hierarchical Bayesian model, which can learn different forms of two dimensional graphs and then produce outputs in a form best matching the relations between the members of a given data set (see Kemp and Tenenbaum 2008, and Griffiths et al. 2010, for more details). If we agree that such outputs are structural maps of the input data, then it seems that the same should be said about the system's acquired internal models consisting of extracted input regularities and graph forming rules used to generate such outputs. After all, the main difference between them is that of format and not of function — the internal model guides the behavior of the lower levels of the system in the same way that the external graph drives the behavior of human users. Even though Kemp and Tenenbaum's model lacks the feedback-like error correction capacity that is exhibited by the Helmholtz machine and the PP models, it puts stress on views aiming to explain the behavior of probabilistic models in terms of mere biasing relations between their elements. While such a perspective can be useful in trying to decompose the system into its parts, placing sole focus on particular components of PEs could obscure how the interactions between different types of units and sub-systems contribute to the functioning of the wider system.

What is crucial for the present discussion of PP's status is that, from a formal perspective, each PE level in the hierarchy is performing the same kind of basic function — carrying out probabilistic inferences aimed at producing hypotheses, which are best at accommodating the currently available data (incorporating the information fed from the level downstream into the next estimation about the activity on that level). This is done by modelling (via hidden variables) of possible causes responsible for the obtained data, which, together with top-down information, is used for generating the most probable states of hidden variables on the level below (downward projecting prediction units), and are updated in response to the actual states of the modeled variables (backward error connections). What is crucial in this picture is that each PE is predicting the activity on the level below by building an estimate of the causes of that activity. By interacting in such a way, each level of the hierarchy is minimizing prediction error by effectively acting as a model of the level below, although this does not mean that the system *only* represents itself as Anderson 2017, suggests.

### 3.2 The Functional Role of Low-Level Prediction Estimators — A Puzzle for rPP

The above discussion of PEs calls into question Orlandi and Clark's arguments against treating low levels of the hierarchy as fulfilling a representational role.

Firstly, although composed of simple elements, PEs gain their computational prowess from the interactions of their parts and the fact that they are estimating possible causes of the states which they are supposed to model. This is especially problematic for Orlandi's claim about the non-representational nature of early perceptual stages. Levels directly predicting the states of the sensory organs do so by employing latent variables, which act as estimates of external states responsible for activations of sensory periphery. These variables play the role of model parameters for generating predictions (or 'mock inputs' as Orlandi calls them) which are tested against the actual states of the sensorium. Peripheral levels of the prediction error minimization hierarchy cannot act as mere error detectors, even if they employ simple non-representational units. To function properly they must be able to update their internal estimates in response to the error signals and the information from levels above. In other words, they must be capable of error correction in the same way that levels further up the hierarchy are — by updating a model of the hidden causes from which predictions are to be generated. This means that, despite their proximity to the periphery, PEs exploit estimates of states which are not immediately available to them and which are removed from the structures that are doing the modeling. Therefore, even if the task of the lowest levels consists in generating patterns of 'mock-stimulus' activations which will be compared against the states of sensory detectors (e.g. rods and cones), the manner in which this is fulfilled goes beyond mere error detection. In other words, the lowest levels act as models, not because they can generate 'mock inputs', but in virtue of how such inputs are generated.

Furthermore, though predictions on the lowest perceptual levels may not play a direct role in causing behavior, similar parts of the hierarchy terminating in the ventral horn of the spinal cord play a crucial role in initiating and controlling bodily actions. For example, Friston and colleagues propose that such prediction passageways carry information about the expected proprioceptive inputs (Friston et al. 2010). In cases of discrepancy between actual and predicted bodily states such predictions can, upon further processing in the spinal cord and the peripheral nervous system, effectively act as motor commands which bring bodily actuators (such as muscle spindles) into their expected states. Importantly, they do not have such an effect in virtue of being simple activation commands themselves, but by carrying information about the desired state of particular actuators, which is then translated into commands that can bring it about (this proposal can explain somatic reflex arcs and offers an outlook on developing it into a full-fledged story about goal-directed action; see also Burr 2017; Vance 2017; Limanowski 2017, for a more detailed discussion of PP's treatment of action).[10]

A similar point holds for Clark's claim regarding model-free processing. Since early processing stages are supposed to act, not only as models of stages downstream, but also as models of the environment (not to mention relying on upstream stages for the control of their internal estimates), there can never be a truly model-free process within the hierarchy. The initial proposal that some models may rely heavily on bottom-up input is equally difficult to defend, since all PP algorithms assume that bottom-up channels carry information about the difference between the predicted and actual states on the level below[11]. This signal is informative only in context of current estimates of the causes used to generate the predictions. Admittedly, Clark does motivate his position by pointing out that PP systems strive for efficient coding by simplifying and reducing the complexity of their models (FitzGerald et al. 2014). However, this can be understood, for example, in terms of generating predictions using less hidden variables or variables with fewer degrees of freedom. Such a solution does not undermine the fact that PP's processing stages are functioning as models. Instead, it would relegate the problem of differentiating rPP from cPP to the side of implementation details or questions about content of different PEs. For example, the issue would now be to explicate the number of hidden variables at lower levels or to explain which environmental features they correspond to, rather than to describe the way in which they are employed by the system. However, as I will try to show in the penultimate section of this paper, the issue of content determination in PP can be a problem even for proponents of cPP.

### 3.3 The Functional Role of High-Level Prediction Estimators — A Puzzle for cPP

The discussion of PEs, as presented so far, has confirmed Gładziejewski's claim that the PP systems' generative models do fulfill a representational role, thereby meeting the representational job description challenge. But the cPP camp should not celebrate victory just yet, as this is not the end of problems for a representational reading of PP.

Recall, that what follows from the modeling details of PP is that, from a formal perspective, each level is acting as a model of the level below. Meaning that, while producing predictions about the behavior of the nearest stage downstream, each level is also modeled by the level immediately above. This is supposed to be accomplished by each PE generating predictions from estimates of causes responsible for the behavior of the stage below. In the previous section we saw that this creates a problem

10 Orlandi claims that equating action with error correction for the purpose of explaining how internal models can fulfill the function of action guidance required by S-representations is question-begging (Orlandi 2015, p. 23). This accusation seems to lean heavily on a different claim made by Orlandi, namely, that low-level PEs are not sufficiently detached from their representational target — the body. I hope that this and the previous section are successful at elucidating why such sentiment is misguided. I would also like to point out that the view of motor control offered by PP can be traced back to popular forward-model/efference copy accounts on which the motor control system optimizes motor performance using the predicted sensory outcomes of action commands (e.g. Wolpert and Miall 1996, and Grush 2004). As Clark himself has argued, PP accounts of action production belong to the wider category of forward-models (Pickering and Clark 2014). The main difference being, that most forward-model based theories assume the models to be implemented in separate prediction modules, rather than be integrated into the parts of the system issuing motor commands.

11 Note that, although Spratling's PC/BC-DIM does forward propagate estimations of causes, its lowest levels are still in business of comparing predictions to states of sensory peripheries (see Spratling 2008, Fig.2).

for proponents of rPP as it suggests that, in levels terminating in the sensory peripheries, PEs model external causes responsible for activation patterns of sensory receptors. However, this feature of PP architectures may also be a problem for proponents of cPP. Consider, once again, an *n-th* level of the system, one which is at least 2 steps away from the periphery ($n > 2$). The PE on the *n-th* level is supposed to model the behavior of the level below in virtue of estimating the causes behind the 'observed' activity patterns of the lower level's estimator units. The question here is: what exactly are the causes of $n - 1$'s behavior, which the hidden variables of $n$ are estimations of? Do the hidden variables of $n$ compute distal environmental causes, just like in the levels terminating in the sensory receptors, or do they estimate the input $n - 1$ receives from $n - 2$?

This issue has recently surfaced in the subject literature. According to Spratling's review of PP algorithms, only some of the computational models "[…] are concerned with finding the coefficients which encode the underlying causes of the sensory data […], while others are concerned with finding these coefficients only for the purpose of calculating, and transmitting, the residual error […]" (Spratling forthcoming p.5). The reason behind Spratling's assessment seems to be that the processing stages of some PP algorithms are said to be covered by a 'Markov blanket' (Pearl 1988), where each level is said to be 'inferentially encapsulated' from all but the immediately neighboring stages, the informational states of which completely determine the total informational state of the level under investigation (see Friston 2008, for a detailed overview of the mathematical properties of his models, and Hohwy 2016, for a philosophical discussion of the consequences of such encapsulation). In cases of algorithms with such properties, discovering what exactly is being tracked by levels deep in the hierarchy can be problematic, since determining their states does not require taking any system external context into consideration. This, in turn, implies an ambiguity between levels which perform the function of standing-in for features of the environment and those which act only as meta-representations by standing-in for the features of levels downstream.

This is a serious problem for proponents of the S-representational reading of PP, since the very notion rests on the assumption that the function of representational structures is to exploit a mapping relation obtaining between said structures and the world. Proponents of cPP could attempt to relax this condition and simply claim that features of a representational structure have to map onto any other physical structure, regardless of whether it is internal or external to the larger system in which the representation is employed.[12] However, taking this route would significantly weaken the structural view and would not solve the problem at hand. It would obscure how S-representations fulfill their function, since we could no longer refer to prototypes in order to guide the ascription of representational roles (unless someone can point to an uncontroversial, commonsense example of such a structure performing a map-like function). Moreover, such a solution would inflate the problem of content indeterminacy, as it would extend the representational structures' target domain to include not only external, but also internal states of the system.

## 3.4 Structural Ambiguity of Prediction Estimators — A Puzzle for PP

Proponents of cPP may push back against the accusation of functional indeterminacy by calling upon the mapping relation ingrained into the notion of S-representation. After all, the newly presented problem of functional ambiguity questions what the levels (or rather the currently active models encoded in relevant PEs) stand-in for and not whether they perform the function of standing in for anything at all. The cPP proponents can, therefore, try to solve this problem by refocusing on the issue of content and showing that PEs do refer to the worldly causes and therefore satisfy the requirement for being S-representations. One way of accomplishing this is by pointing out that functional ambiguity

12 Interestingly, this move would bring the account of representation espoused by proponents of cPP closer to Grush's notion of emulator representation. Still, it is worth noting that supporters of S-representation take the representational job description to be merely a sufficient condition while Grush holds it (or something very close to it) as a jointly necessary and sufficient one: "[…] something is a representation if and only if it is used by some system to stand for something else, and the "stand for" is explained in terms of use" (Grush 2004, p.428).

does not mean that the levels do not come to represent some statistically relevant features of the world in a transitive way. Any level can be said to process information about the system's peripheral inputs in virtue of being fed prediction error from a lower level $n - 1$, which receives inputs from a yet lower level, and so on until level $n - x$ which receives inputs directly from the sensory periphery (which under current taxonomy should not be considered a 'level' of the predicative hierarchy). Therefore, a possible line of defense is that, each and every level is ultimately modelling changes in the sensory states. In order to predict these changes the processing stages must come to resemble the sources of the occurring stimulations. This seems to be an assumption held by Friston and Hohwy, who claim that the dynamics of the higher levels in the hierarchy come to mirror the causal dynamics of their target domain by employing the empirical Bayes method, which allows the system to extract estimates of prior probabilities from the input data through learning (Friston et al. 2012).

However, this reply seems to face at least two major problems. The first, more general worry, is that Friston's answer to the problem of content determination does not clarify how representations modeling different worldly regularities can be distinguished. The typical answer here seems to be that the structure of such PEs will somehow resemble the structure of the represented domain (this is, for example, how Gładziejewski interprets this claim, but see also Friston 2013). However, as has been pointed out by Ramsey, resemblance is a very weak constrain which cannot solve the problem of content determination on its own.

This issue has been recently addressed by Wanja Wiese who suggested that, by focusing on particular computational models and algorithms, it is possible to identify unique constrains they place on the kinds of cognitive contents that are employed in PP (Wiese 2016). As Wiese points out, computational models make specific predictions about the mathematical contents (Egan 2014) they operate on (e.g. employing continuous vs. discrete values), the kind of relations obtaining between them (e.g. the relationship between different levels in the hierarchy), and the ways they interact with each other (e.g. in the expected precision system, which modulates the bandwidth or gain on the error signal pathways, see Hohwy 2012). This allows for a higher degree of specificity with regards to the structures which are supposed to underpin different cognitive processes, introducing constraints specific to PP, which can help in identifying cognitive representations. Additionally, since the models themselves are meant to serve as functional descriptions of such processes, they must offer some initial specification and constraints on the kinds of cognitive contents involved. Finally, some PP theorists have postulated that the computational descriptions on offer have implications for subjects' phenomenology. By placing constraints on the viable phenomenological descriptions, Wiese argues, PP introduces yet more constraints on the system's contents, e.g. by making predictions about the dynamics and possible manipulations of perceptual experience.

Owing to constraints in space, I cannot do justice to Wiese's nuanced position, I will merely gesture at one possible objection to his proposal, before moving onto a more burning problem facing all of PP. The initial argument that by focusing on the mathematical details of particular algorithms, PP can offer a nuanced story about the implementation and individuation of representational structures, is compelling, However, it is not entirely clear that this will provide a grip on the environmental properties to which different parts of the system correspond. Cognitive contents used in descriptions relating cognitive functions to their underlying computational processes can be seen as serving the role of identifying the mapping between the inputs/outputs of cognitive functions and the inputs/outputs of an abstract computational description. William Ramsey has argued that such input-output representations or IO-representations pass the representational job description by playing a crucial explanatory role in computational theories of cognition. Firstly, they are used for defining the theory-independent explananda, with different competing computational descriptions acting as their explanans (Ramsey labels such IO-representations as 'external'). Secondly, they are meant to guide the process of task decomposition by specifying the start- and end-points of intermediate processing stages postulated by a theory or computational description (in which case they describe what Ramsey calls 'internal' IO-rep-

resentations, specific to the computational description under investigation). What is important here is that this notion of representation is different from S-representation, as it is not meant to accord to our everyday use of the concept, and the explanatory role it plays in our theories does not presuppose the existence of a mapping relation between the representational structure and the world (rather the mapping is supposed to obtain between the computational description and its physical vehicles manipulated by the computational mechanism). Therefore, a possible worry is that it is not clear whether this type of representation can offer an account of how S-representational contents are to be individuated, as opposed to presenting an account of implementation, explaining how we can adjudicate between adequate and inadequate computational descriptions of cognitive functions. Wiese seems to postulate that the constraints placed by the mathematical contents of the system's internal IO-representations will be sufficient to guide ascriptions of cognitive contents to corresponding S-representations. This is certainly an interesting proposal that warrants further investigation. Still, it is not clear whether it will help us determine the cognitive contents of particular structures beyond providing a general description of the kinds or types of content that portions of the hierarchy are supposed to process in order to fulfill their function (e.g. by labeling several levels as responsible for shape perception).

The second general, but closely related worry, is that it is not clear to what extent the learning process can yield variables with structural features resembling the dynamics of properties existing in the environment. Living organisms can only achieve bounded optimality. It is entirely possible that some organisms could acquire models which are useful under certain environmental and ecological constraints but which do not, in fact, correspond to easily discernible features of the world. Properties which are organism-relative, such as 'attractiveness' or 'tastiness', could serve as an example.[13] Although such properties do depend on the physical make-up of the world and the organism, they do not directly pick out any fixed features of the environment, but rather rely on an abstraction from a set of complex interdependencies obtaining between such features. This is problematic for determining contents of PP models because such properties can vary with changes in the environment or in the organism itself. For example, a model of a desirable mating partner can not only vary significantly within one species, but change due to seemingly unrelated factors (e.g. climate, access of food, presence of predators, etc.), further complicating the task of determining what exactly is the target of the system under investigation.

One way to tackle this problem, as Ramsey himself proposes, is to consider the contents of S-representations to be, at least partially, determined by more than structural similarity. Augmenting his view with a teleological account of function could yield a more restricted account of content by appealing to the representational system's etiology and the relevance of its contents for action guidance. In a similar vein, Gładziejewski suggests that action based accounts of content, such as success semantics (Blackburn 2005) or interactionist accounts of representation (Bickhard 1999), which broadly claim that contents are determined by conditions of successful action, could also provide a way out of this problem. Both of these proposals are interesting and worth further exploration within the PP framework.

However, as Gładziejewski himself notes, there may still exist cases where such augmented account would break down, especially in situations where contents do not affect organismic actions directly, but only guide perceptual processes. Therefore, a more problematic possibility is the existence of models which are adaptive, but operate on variables which do not correspond to any properties in the environment. It is possible that a PP system could acquire a false model of hidden causes which is 'good enough' at minimizing error under certain often encountered conditions. Such pragmatically driven misrepresentations could be a result of reducing internal models' complexity, e.g. in cases where regular co-occurrence of two environmental states or features is wrongly inferred as being caused by a common coefficient. In such scenarios, the model could operate on variables which do not correspond to any worldly counterpart, but are successful at generating error minimizing predictions as long as the organism is under the conditions that shaped them (see, e.g. Pliushch 2017, for a more detailed

13 I am indebted to one of the reviewers for this suggestion.

account of how this can occur in cases of self-deception). Thus, we could talk of the system being stuck in a local prediction error minimum until it is affected by an unexpected perturbation.

The above examples do not present an insurmountable obstacle for proponents of the representational interpretation of PP. As McKay and Dennett conclude in their influential treatment on misbelief: "Although survival is the only hard currency of natural selection, the exchange rate with truth is likely to be fair in most circumstances" (McKay and Dennett 2009, p. 509). Following a similar intuition, some proponents of PP point out that the long term survival chances of any organism seriously wrong about the structure of its environment will be very low. The kind of misrepresentation just discussed is likely to be an exception rather than a wide spread phenomenon. This does not mean, however, that the issue of misrepresentation is not a problem for determining the contents of systems' representations. Without a detailed account of how to differentiate the structures that map onto the features of the environment, from those that do not, cPP fails to secure a completely representational interpretation of the framework.

## 4 Conclusion — A Call for Moderation

The aim of this paper was to compare and analyze two main interpretations of the predictive processing framework with regards to its representational status. Two conclusions emerge from this discussion.

Firstly, although proponents of rPP present the discussion as a disagreement over the functional details of PP by appealing to Ramsey's job description challenge, this condition favors a representational interpretation of the framework. Therefore, if rPP is to be defined in terms of commitments about the functional role played by generative models, the position will collapse into the standard, representational reading of the framework. In order to avoid this fate, the disagreement regarding the representational status of PP should be construed as a disagreement about either the contents of PP's models, or the details of their implementation.

Secondly, applying Ramsey's job description challenge to PP architecture supports the cPP reading only if the problems stemming from its use can be resolved. Thereby, proponents of the representational view are burdened with distinguishing models which have the function of representing the world, from those that function as representations or meta-representations of the system's internal states. Defenders of cPP need to offer a detailed account of content determination, since, as Ramsey points out: an account of content and not only function is needed for something to qualify as a full blown representation.

All this calls for moderation in making claims about the representational status of PP. Yes, the framework offers a compelling account of a cognitive system composed of nested probabilistic structures which do function as models capable of self-generating their inputs. Still, it may be too early to bet money on how accurately (or if at all) different features of these models map onto the external environment. There are currently two solutions emerging in response to problems presented in the paper. The first is to double down on the search for additional (e.g. mechanistic) constrains which would help to anchor the representational claims made by cPP in implementations of particular algorithms, hoping that (as in the case of cognitive maps), once relevant physical structures are known, a strong correspondence relation between their features and some worldly domain can be established. This position is endorsed, among others, by Gładziejewski 2015, and Wiese 2016. However, not everyone is optimistic with regard to the framework's ability to provide such constraints (e.g. Miłkowski 2013, argues that PP models are at best mechanistic schemata). Those questioning the framework's explanatory scope and ability to present necessary empirical details, have called for treating probabilistic models of cognition as instrumental (Colombo and Seriès 2012), implying a similar approach to ascribing contents to entities postulated by such models. Much more work is needed to settle this debate, but regardless of its final outcome, the search for framework specific constrains on mental content is likely to exert a tremendous transformative power across the fields of cognitive studies.

## References


Anderson, M. L. (2017). Of Bayes and bullets: An embodied, situated, targeting-based account of predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Bartels, A. (2005). *Strukturale Repräsentation*. Paderborn: mentis.

Bickhard, M. H. (1999). Interaction and representation. *Theory & Psychology, 9* (4), 435–458. https://dx.doi.org/10.1177/0959354399094001.

Blackburn, S. (2005). Success semantics. In H. Lillehammer & D. H. Mellor (Eds.) *Ramsey's legacy*. Oxford: Oxford University Press.

Bruineberg, J. (2017). Active inference and the primacy of the 'I can'. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Burr, C. (2017). Embodied decisions and the predictive brain. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36* (03), 181–204. https://dx.doi.org/10.1017/S0140525X12000477.

——— (2015). Radical predictive processing. *The Southern Journal of Philosophy, 53*, 3–27. https://dx.doi.org/10.1111/sjp.12120.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York: Oxford University Press.

Colombo, M. & Seriès, P. (2012). Bayes in the brain—on Bayesian modelling in neuroscience. *The British Journal for the Philosophy of Science, 63* (3), 697–723. https://dx.doi.org/10.1093/bjps/axr043.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69* (6), 1204–1215. https://dx.doi.org/10.1016/j.neuron.2011.02.027.

Dayan, P. (2012). Instrumental vigour in punishment and reward. *European Journal of Neuroscience, 35* (7), 1152–1168. https://dx.doi.org/10.1111/j.1460-9568.2012.08026.x.

Dayan, P. & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience, 8* (4), 429–453. https://dx.doi.org/10.3758/CABN.8.4.429.

Dayan, P., Hinton, G. E., Neal, R. M. & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation, 7* (5), 889–904. https://dx.doi.org/10.1162/neco.1995.7.5.889.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Downey, A. (2017). Radical sensorimotor enactivism & predictive processing. Providing a conceptual framework for the scientific study of conscious perception. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Dretske, F. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press.

Dretske, F. I. (1988). *Explaining behavior: Reasons in a world of causes*. Cambridge, MA: MIT Press.

Egan, F. (2014). How to think about mental content. *Philosophical Studies, 170* (1), 115–135. https://dx.doi.org/10.1007/s11098-013-0172-0.

Feldman, H. & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*, 215. https://dx.doi.org/10.3389/fnhum.2010.00215.

Fink, S. B. & Zednik, C. (2017). Meeting in the dark room: Bayesian rational analysis and hierarchical predictive coding. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

FitzGerald, T. H. B., Dolan, R. J. & Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in Human Neuroscience, 8*, 457. https://dx.doi.org/10.3389/fnhum.2014.00457.

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass, 5* (10), 914–926. https://dx.doi.org/10.1111/j.1747-9991.2010.00330.x.

Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 11* (4). https://dx.doi.org/doi:10.1371/journal.pcbi.1000211.

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

———(2013). Life as we know it. *Journal of the Royal Society, Interface, 10* (86), 20130475. https://dx.doi.org/10.1098/rsif.2013.0475.

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

Friston, K., Adams, R., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments.

*Perception Science, 3*, 151. https://dx.doi.org/10.3389/fpsyg.2012.00151.

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

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

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14* (8), 357–364. https://dx.doi.org/10.1016/j.tics.2010.05.004.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences, 27* (3), 377–396.

Gupta, K., Erdem, U. M. & Hasselmo, M. E. (2013). Modeling of grid cell activity demonstrates in vivo entorhinal 'look-ahead' properties. *Neuroscience, 247*, 395–411. https://dx.doi.org/10.1016/j.neuroscience.2013.04.056.

Gładziejewski, P. (2015). Explaining cognitive phenomena with internal representations: A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric, 40* (1), 63–90. https://dx.doi.org/10.1515/slgr-2015-0004.

——— (2016). Predictive coding and representationalism. *Synthese, 193* (2), 559–582. https://dx.doi.org/10.1007/s11229-015-0762-9.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B. & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature, 436* (7052), 801–806. https://dx.doi.org/10.1038/nature03721.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11* (10), 428–434. https://dx.doi.org/10.1016/j.tics.2007.09.004.

Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science, 268* (5214), 1158–1161.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3*. https://dx.doi.org/10.3389/fpsyg.2012.00096.

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

———(2016). The self-evidencing brain. *Noûs, 50* (2), 259–285. https://dx.doi.org/10.1111/nous.12062.

Jehee, J. F. M. & Ballard, D. H. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS Computational Biology, 5* (5), 1-10. https://dx.doi.org/10.1371/journal.pcbi.1000373.

Kemp, C. & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105* (31), 10687–10692. https://dx.doi.org/10.1073/pnas.0802631105.

Limanowski, J. (2017). (Dis-)attending to the body. Action and self-experience in the active inference framework. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Marr, D. (1982). *Vision: A computational approach.* San Francisco: Freeman & Co.

McKay, R. T. & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences, 32* (06), 493–510. https://dx.doi.org/10.1017/S0140525X09990975.

Metzinger, T. (2017). The problem of mental action. Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Millikan, R. (1984). *Language, thought and other biological categories.* Cambridge, MA: MIT Press.

Miłkowski, M. (2013). A mechanistic account of computational explanation in cognitive science. In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (Eds.) *Cooperative minds: Social interaction and group dynamics. Proceedings of the 35th annual meeting of the cognitive science society* (pp. 3050-3055). Austin, Texas: Cognitive Science Society. http://csjarchive.cogsci.rpi.edu/Proceedings/2013/papers/0545/paper0545.pdf.

———(2015). Satisfaction conditions in anticipatory mechanisms. *Biology & Philosophy, 30* (5), 709–728. https://dx.doi.org/10.1007/s10539-015-9481-3.

O'Keefe, J. & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research, 34* (1), 171–175. https://dx.doi.org/10.1016/0006-8993(71)90358-1.

O'Keefe, J. & Nadel, L. (1978). *The hippocampus as a cognitive map.* New York: Oxford University Press.

Orlandi, N. (2013). Embedded seeing: Vision in the natural world. *Noûs, 47* (4), 727–747. https://dx.doi.org/10.1111/j.1468-0068.2011.00845.x.

———(2014). *The innocent eye: Why vision is not a cognitive process.* New York: Oxford University Press.

———(2015). Bayesian perception is ecological perception. http://mindsonline.philosophyofbrains.com/2015/session2/bayesian-perception-is-ecological-perception/.

Palmer, S. (1978). Fundamental aspects of cognitive representation. In E. Rosch & B. B. Loyd (Eds.) *Cognition and categorization* (pp. 259–303). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Fransisco: Morgan Kaufmann.

Petro, L. S., Vizioli, L. & Muckli, L. (2014). Contributions of cortical feedback to sensory processing in primary visual cortex. *Perception Science, 5*, 1223. https://dx.doi.org/10.3389/fpsyg.2014.01223.

Pfeiffer, B. E. & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature, 497* (7447), 74–79. https://dx.doi.org/10.1038/nature12112.

Pickering, M. J. & Clark, A. (2014). Getting ahead: Forward models and their place in cognitive architecture. *Trends in Cognitive Sciences, 18* (9), 451–456. https://dx.doi.org/10.1016/j.tics.2014.05.006.

Pliushch, I. (2017). The overtone model of self-deception. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Ramsey, W. M. (1997). Do connectionist representations earn their explanatory keep? *Mind & Language, 12* (1), 34–66. https://dx.doi.org/10.1111/j.1468-0017.1997.tb00061.x.

——— (2007). *Representation reconsidered*. Cambridge: Cambridge University Press.

———(2016). Untangling two questions about mental representation. *New Ideas in Psychology, 40, Part A*. https://dx.doi.org/10.1016/j.newideapsych.2015.01.004.

Rao, R. P. N. & 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. https://dx.doi.org/10.1038/4580.

Rezende, D. J., Mohamed, S. & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv:1401.4082*.

Spratling, M. W. (2008). Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience, 2*, 4. https://dx.doi.org/10.3389/neuro.10.004.2008.

——— (2016). Predictive coding as a model of cognition. *Cognitive Processing, 17* (3), 279–305. https://dx.doi.org/10.1007/s10339-016-0765-6.

——— (forthcoming). A review of predictive coding algorithms. *Brain and Cognition*. https://dx.doi.org/10.1016/j.bandc.2015.11.003.

Van der Meer, M. A. A. & Redish, A. D. (2010). Expectancies in decision making, reinforcement learning, and ventral striatum. *Frontiers in Neuroscience, 4*, 6. https://dx.doi.org/10.3389/neuro.01.006.2010.

Vance, J. (2017). Predictive processing and the architecture of action. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Wiese, W. (2016). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 1–22. https://dx.doi.org/10.1007/s11097-016-9472-0.

Wiese, W. & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Wolpert, D. M. & Miall, R. C. (1996). Forward models for physiological motor control. *Neural Networks: The Official Journal of the International Neural Network Society, 9* (8), 1265–1279.

Wolpert, D. M., Doya, K. & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 358* (1431), 593–602. https://dx.doi.org/10.1098/rstb.2002.1238.

# Radical Sensorimotor Enactivism & Predictive Processing

## Providing a Conceptual Framework for the Scientific Study of Conscious Perception

Adrian Downey

In this paper I outline and defend a novel approach to conscious perception, which I label *"radical sensorimotor enactivism"*. The aims of the paper are two-fold: (1) to respond to a common objection to theories like radical sensorimotor enactivism— that they are empirically vacuous— and explain why, because radical sensorimotor enactivism uses (a non-representational version of) predictive processing to operationalize its sub-personal aspects, this objection cannot be levelled at the theory; and, (2) to argue that radical sensorimotor enactivism provides a better empirical account of conscious perception than predictive processing taken as a stand-alone theory. I conclude that radical sensorimotor enactivism provides one with a strong over-arching conceptual framework for the scientific study of conscious perception which clarifies the relation between existing strands of empirical work and provides practical guidance for future research. As such, it is worthy of further development, study, and application in empirical settings.

**Keywords**

Consciousness | Non-representational theories of cognition | Perception | Predictive processing | Sensorimotor enactivism

**Acknowledgments**

This paper (or variants thereof) was presented at the Postgraduate WIP conference at the University of Sussex, an E-Intentionality seminar at the University of Sussex, at the Situating Cognition: Agency, Affect, and Extension conference held at the University of Warsaw, and the Philosophical Aspects of the Predictive Processing Framework conference held at the Frankfurt Institute for Advanced studies. I would like to thank the respective audiences of each of these talks for helpful discussions and constructive critical feedback. Thanks as well go to Sarah Sawyer and Chris Mole, both of whom provided extremely valuable feedback on the ideas contained within this paper. For helpful feedback on the manuscript itself, I would like to thank Joe Morrison, Joe Dewhurst, and two annonymous referees. Finally, and most importantly, I would like to thank Thomas Metzinger and Wanja Wiese for both: 1) providing instructive feedback on this paper, and 2) facilitating the creation of this volume and organising the conference upon which it is based. A lot of hard work goes on behind the scenes in these kind of ventures, but although it may go unseen, it does not go unappreciated – thanks very much for all the help. This research was financed with funding from the AHRC.

## 1 Introduction

In this paper I introduce a novel approach to conscious perception, which I label "radical sensorimotor enactivism" (RSE). RSE is advanced within the intellectual tradition of ecological and enactive approaches to mentality. It is commonly objected that theories advanced within this intellectual tradition are incapable of explaining the brain's role in conscious perception (Chemero 2009, p. 93; Clark 2009; Seth 2014). Consequently, it is thought that such theories provide mere descriptions of conscious perception which are empirically unilluminating. I explain how a non-representational version of predictive processing (PP) can be subsumed within RSE and used to empirically explain the brain-based aspects of this framework. By doing so, I show that one cannot reject RSE for failing to account for the brain's role in conscious perception. Furthermore, I then argue that, not only can RSE account for the brain's role in conscious perception, it actually provides a better account of the brain's role in

Downey, A. (2017). Radical Sensorimotor Enactivism & Predictive Processing – Providing a Conceptual Framework for the Scientific Study of Conscious Perception. In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 11. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573123

conscious perception than PP taken as a stand-alone theory. RSE provides a powerful over-arching conceptual framework for the scientific study of conscious perception which helps to clarify and taxomonise existing strands of empirical work whilst providing guidance for future research. As such, I conclude that it is worthy of further research, development, and application in empirical settings.

My paper is structured as follows— in section two, I outline RSE. I explain that RSE is predicated on the *sensorimotor enactive* theory of conscious perception, but that it improves upon this theory because it can provide a better account of the sensorimotor enactive concepts of "*sensorimotor knowledge*" and *"attention"*. In section three I explain why theories like RSE are often thought to ignore the brain's role in conscious perception, before outlining a non-representational version of PP and explaining how it can be subsumed within RSE. I show that this non-representational PP account can operationalise and empirically explain the brain-based aspects of RSE, and so it cannot be objected that RSE ignores the brain's role in conscious perception. Finally, in section four, I provide two reasons for preferring this conjunction of RSE and non-representational PP over a representational PP account. I argue that RSE provides a better explanation of conscious perception because: it provides a better account of the inter-relation between the sub-personal, personal, and conscious levels of explanation; and, it can account for and better categorise a wider range of empirical work in cognitive science. Thus, I conclude that not only can RSE account for the brain's role in conscious perception, it actually provides a better account of its role in conscious perception than rival theories (such as PP taken stand-alone).

## 2 What Is Radical Sensorimotor Enactivism?

In this section I outline a novel approach to conscious perception, which I label "radical sensorimotor enactivism" (RSE). I begin by outlining the *sensorimotor enactive* theory of conscious perception, upon which RSE is predicated. I explain that extant versions of this theory are problematic (from an enactive perspective) because the key concepts of "*sensorimotor knowledge*" and *"attention"* both require representation, or they are left entirely unexplained. Then, I provide a non-representational explanation of both of these concepts. Consequently, I arrive at a non-representational (or *radical)* version of sensorimotor enactivism, and thus at *radical* sensorimotor enactivism.

### 2.1 Sensorimotor Enactivism

*Sensorimotor enactivism* (O'Regan and Noë 2001; Noë 2004; O'Regan 2011) is a direct realist theory of (conscious) perception. Consequently, it takes (conscious) perception to involve a direct relation between the perceiving organism and its environment. It explains this relation to be enabled by the organism's possession of *sensorimotor knowledge*, which is knowledge of the law-like relation between sensation and movement. For example, there is a law-like relation between an organism's movements and its visual stimulation— when an organism moves closer to an object the object looms in the visual field, when it gets further away the object appears smaller, and so on. On sensorimotor enactivism an organism is thought capable of perceiving only when it *understands* this relation between sensory stimulation and movement. Finally, sensorimotor enactivism takes attention to be necessary for *conscious* perception. On this theory an organism will be conscious of its perceptual relation to the environment only when it attends.

Sensorimotor enactivism traces its intellectual roots directly from the non-representational tradition of enactive and ecological approaches to mentality (Gibson 1979; Varela et al. 1991; Ryle 1949/2000; *cf.* Chemero 2009, fig. 2.4). However, in spite of its non-representational heritage and its proponents' arguments against representational theories of conscious perception (O'Regan and Noë 2001), extant versions of the view have not done enough to distance themselves from representation (as will be explained in the following sections). In particular, the key concepts of "sensorimotor knowledge" and "attention" are both explained in a manner which makes indispensable use of the concept

of "representation". Consequently, these key terms have either received a representational explanation, or they are left underspecified. I am going to provide a non-representational account of both of them, and therefore arrive at a thoroughly non-representational version of sensorimotor enactivism.

### 2.1.1 Non-Representational Sensorimotor Knowledge

It is often emphasised that sensorimotor knowledge should be understood as a kind of practical skill, which is predicated upon the organism's possessing a certain kind of embodied know-how. However, in spite of its purported non-representational credentials, the concept itself appears to require representation. This point has been most forcefully argued for by Hutto (Hutto 2005; *cf.* Hutto and Myin 2013, pp. 23-32). Hutto argues that, in order to amount to any kind of substantive claim about the nature of perception which can play any sort of explanatory role, the concept of "sensorimotor knowledge" requires representation. Consequently, in order to provide a properly *radical* version of sensorimotor enactivism, I must provide a thoroughly non-representational and yet explanatorily substantive account of sensorimotor knowledge.

At the sub-personal level of explanation, I propose that sensorimotor knowledge be explained in a psychological behaviourist manner (*cf.* Block 2001). Psychological behaviourists took mentality to be constituted by a series of relations between sensory input and motor output, with a given sensory input related to a given motor output (or series of motor outputs, depending on the organism's phylogenetic and ontogenetic history). If we explain sensorimotor knowledge in this manner, then it can be taken to concern a series of relations between certain sensory inputs to, and certain other motor outputs from, the brain. On this sub-personal account of sensorimotor knowledge, the brain acts merely as a causal mediator between certain neural inputs and certain motor outputs, and so one avoids postulating representation at the sub-personal level (*cf.* Ramsey 2009, ch. 4).

At the personal level of explanation, I propose that sensorimotor knowledge be understood in terms of Ryle's theory of knowledge-how (Ryle 1949/2000, ch. 2). On this theory, knowledge-how is understood wholly in terms of behavioural dispositions— if an organism possesses knowledge of how to perform an activity, then this knowledge-how will be exhibited in the organism's actual behaviour and their counter-factual behavioural tendencies. For example, if an organism knows how to throw a stone, then it will exhibit stone-throwing behaviour in scenarios where stones are available and the organism deems it pertinent to throw them. If we apply this Rylean account of knowledge-how to personal level sensorimotor knowledge, then we should take the ability to perceive to be exercised when the organism is disposed to engage in perceptual behaviour. It will, for example, exhibit mating behaviour when it *sees* another member of its species of the requisite sex, when it *smells* the presence of such a member, and so on and so forth. If the organism is disposed to behave in a manner consistent with its understanding the law-like relation between sensation and movement, that organism can be considered to possess personal level sensorimotor knowledge. Once more, because this personal level account of sensorimotor knowledge is entirely concerned with dispositions, one definitively avoids representation.

By providing an account of sensorimotor knowledge upon which it is wholly constituted by non-representational causal mediation and/or dispositions at both the sub-personal and personal levels of explanation, we therefore arrive at a thoroughly non-representational account of sensorimotor knowledge. In section three, I will explain how this non-representational account can be made explanatorily substantive.

### 2.1.2 Non-Representational Attention

Although sensorimotor enactivism takes attention to be necessary for conscious perception, there has been hardly any work focused on providing a substantive account of what, exactly, is meant by "attention". The only proposal within this vicinity has been provided by O'Regan , who argues that sen-

sorimotor enactivism should adopt a *higher-order thought* approach to consciousness (O'Regan 2011). Given that attention is necessary for consciousness, it could therefore follow that attention is itself to be explained in terms of higher-order thought theory. Higher-order thought theory understands consciousness to occur when an organism possesses a higher-order thought about a lower-order mental state. For example, the organism's visual states will become conscious when it has a higher-order thought about those states. This theory indispensably requires representation (the higher-order thought is *about* the lower-order state).[1] Consequently, acceptance of a higher-order thought approach to attention requires acceptance of representation. Therefore, extant sensorimotor enactive accounts either do not explain attention at all, or they explain it in terms of representation. I am going to propose a non-representational theory of attention. First, I will outline Mole's metaphysical distinction between "*process*" and "*adverb*". Then, I will explain how Mole applies this metaphysical distinction to the case of attention. Finally, I will outline Mole's adverbial theory of attention, before providing a non-representational version of the theory and applying it to sensorimotor enactivism.

#### 2.1.2.1 Distinguishing "Process" and "Adverb"

Mole's adverbial theory of attention is predicated upon a metaphysical distinction between the concepts of "process" and "adverb". He summarises this distinction as follows:

> A taxonomy is a taxonomy on the basis of process if the taxonomy classifies events on the basis of having or gaining of a property *by an object*. A taxonomy is a taxonomy on the basis of manner if the taxonomy classifies events on the basis of the having or gaining of a property *by an event*. (Mole 2011, p. 29, *italics in original*)

In order to determine the metaphysical category of a given *x,* Mole argues that we must consider the following two questions (Mole 2011, ch. 2):

1. What is *x*?
2. What does it mean for *x* to occur?

Mole argues that *x* should be accorded the metaphysical status of process if, in determining its metaphysical status, it is most natural to answer question (1) first. If, however, it is more natural to answer question (2) first, then Mole concludes that *x* should be accorded the metaphysical status of adverb. Mole uses the examples of "combustion" and "hastily" (a process and adverb respectively) to make this point clearer.

In order to determine the metaphysical category of combustion it is natural to answer question one first. Combustion occurs when an object gains the property of burning because, when an object combusts, a chemical reaction between oxygen and fuel occurs which results in burning. We can therefore classify combustion as a process which occurs to an object and causes it to gain the property of burning. Having arrived at an answer to question one, the answer to question two becomes obvious. For combustion to occur we require that the chemical process of burning occurs. For this reason, Mole labels combustion a *process-first* phenomenon— in order to determine what combustion is we have to first understand what the *process* of combustion is.

Consider now the adverb "hastily". This adverb can be used to describe many different types of event— the publication of a newspaper can be performed hastily, a person's walk to the train station can be performed hastily, and the actions of two particular Hobbits in Middle Earth can be hastily performed (or at least, considered as such from the perspective of a disapproving Ent). Each of these events involve entirely different processes and it would be difficult, if not impossible, to find a pro-

1 Put more precisely, all extant versions of higher-order thought theory indispensably require representation. Although higher-order thought theory could perhaps be explained without representation, no such version of this theory has as yet been proposed.

cess which all of these hastily executed events have in common. Consequently, one cannot determine whether *x* was hasty solely by focusing on the process of *x*. Each of these events are similar because they are carried out in a similar manner (hastily), and not because they involve the same processes. As such, if we are to determine whether a given *x* was performed hastily, we need to understand the manner in which the process of *x* was carried out. Thus, given that it is most natural to answer question two first when determining whether an event was hasty, haste should be accorded the metaphysical status of adverb.

In summary— Mole argues that there is a metaphysical difference between certain processes and events, and he claims that this difference is captured by the metaphysical categories of "process" and "adverb". Process-first phenomena can be grouped into the same set because they all involve an object gaining a property by undergoing a particular process. Adverbial phenomena, on the other hand, can be grouped into the same set because they all involve an event gaining a property by being carried out in a particular manner. In order to determine the metaphysical category of a given *x*, one must determine whether it makes more sense to ask first of *x* what process it involves, or to instead ask first of *x* the manner in which its process is carried out.

#### 2.1.2.2 Applying the Distinction to Attention

In order to apply Mole's metaphysical taxonomy to attention, we must ask the following two questions:

1. What is attention?
2. What is it for something to be done attentively?

(Mole 2011, p. 24)

Philosophers and psychologists have traditionally answered question one first. Consequently, they tend to believe that attention is a process-first phenomenon. The consensus view that attention is a process has not, however, led to a consensus opinion on what "attention" is. Indeed, there are so many different candidates for explaining the process of attention that most psychologists do not believe there is a single set of processes which are necessary and sufficient for attention. Mole argues against process-first theories of attention (arguments which I will not rehearse here), and in their stead offers his own positive proposal about how attention should be classified. In determining the metaphysical category of attention, Mole argues that it is most natural to ask question (2) first. Thus, he concludes that attention should be understood as an adverb.

According to Mole, when an organism attends to *x*, it does so by 'attentively *x*-ing'. His *Cognitive Unison* theory proposes that attention occurs when an organism uses its cognitive resources in unison to perform a task. According to Mole, if an organism is to count as performing a task, the following three conditions must be met (Mole 2011, pp. 52-5):

1. The task must include the organism.
2. The organism must know-how to perform the task.
3. The organism must be putting their know-how to use.

Mole therefore argues that attention occurs when an organism uses its cognitive resources in unison to attend to task *x*. Organisms will only count as 'attentively *x*-ing' if it is the organism itself which performs *x*, the organism knows-how to *x*, and the organism is currently engaged in *x*-ing. If these three conditions are met, then the organism can be ascribed the adverbial property of 'attentively *x*-ing'.

#### 2.1.2.3 'Radicalising' Cognitive Unison

Mole's version of Cognitive Unison theory requires representation at both the personal and sub-personal levels of explanation. Mole takes attention to be a type of personal level cognition, and he argues that personal level cognition is representational:

> A cognitive process, in this sense, is a process that operates on representations that encode their contents *for the agent of the task*: In order for a process to count as cognitive, there must be an agent-level contentful state whose content is directly determined, at least in part, by the content of the representations on which that process operates. (Mole 2011, pp. 57-58, *italics in original*)

Furthermore, at the sub-personal level of empirical implementation, Mole argues that Cognitive Unison is best understood as implemented via *biased-competition*:

> If the cognitive unison theory gives us the correct account of *what* attention *is*, then the biased-competition model may give us the correct account of *how* many instances of this attention-realizing unison *come about*, and of how they get maintained. (Mole 2011, p. 133, *italics in original*)

Competition models of cognition take different areas of the brain to be engaged in constant competition with one another, with their end goal thought to be that of controlling neural processing. Biased-competition theory is a theory of attention which explains neural competition in terms of brain-based representations vying for control of the brain (Desimone and Duncan 1995). Consequently, biased-competition is a representational theory of attention, and so Cognitive Unison appears to require representation at the sub-personal level. Thus, in order to provide a non-representational version of Cognitive Unison, I must explain: how personal level cognition can be explained without representation; and, explain how biased-competition could occur without requiring representation.

Spatial constraints require that I do not argue for any particular non-representational theory of personal level cognition here. Instead, I am going to assume that a non-representational theory of personal level cognition *could* be given (see, for example Anderson 2014; Barrett 2011; Chemero 2009; Hutto and Myin 2013; Ramsey 2009).[2] Given that there are plenty of non-representational theories of cognition present in the literature, and given that Cognitive Unison is itself a type of cognition, it therefore follows that Cognitive Unison *can* be accepted without requiring representation. Therefore, provided one accepts a non-representational theory of cognition, one can arrive at a non-representational version of Cognitive Unison theory.

Where the empirical implementation of Cognitive Unison is concerned, although biased-competition theories are representational theories, there are non-representational alternatives to be found in the literature. One such approach is developed in the work of Anderson (Anderson 2014, esp. chs. 5 & 6; Anderson 2015). In his 2014 *After Phrenology,* Anderson argues for an extremely anti-modular, *neural-reuse* theory of cognition, upon which different neural regions are constantly altering their connections with one another in response to various task-demands. Anderson claims that brains are constantly engaged in the formation and dissolution of what he labels "transiently-assembled-local-neural-subsystems" (or TALoNS, for short). Although TALoNS can be studied via the traditional methods of systems neuroscience (for example Sporns 2010), Anderson argues that we can come to a better understanding of TALoNS if we adopt the *affordance competition hypothesis*.

*"Affordance"* is a theoretical concept which states that organisms perceive their environment in terms of the activities it affords (Gibson 1979). Humans, for example, are thought to perceive chairs

2 Note, I am not assuming the truth of such non-representational theories of cognition. Rather, I am only claiming that non-representational theories of cognition *can* be provided. Even the staunchest representational theorist has to agree on this point, on pain of making representational theories of cognition true a priori and so empirically vacuous (Ramsey 2015).

as affording sitting behaviour, and so humans will perceive them as 'sit-upon-able'. The Affordance Competition Hypothesis proposes that organisms are constantly engaged in processing multiple environmental affordances at any given time— affordances compete with one another within the brain, with the winning affordance being the one that ends up controlling behaviour (Cisek 2015).

Anderson argues that the brain's neural dynamics should be studied in terms of affordance competition. By conceptualising neural dynamics in this manner, he contends that we can come to an understanding of why TALoNS are constantly being formed and re-formed 'on the fly'.[3] TALoNS are formed 'on the fly' because different affordances are competing with one another within the brain, with the winning affordance being the one which determines the nature of the TALoNS being deployed. Anderson labels his theory "*biased-affordance-competition*", and he is quite explicit that it should not be understood in terms of representation. Because Anderson provides a non-representational version of biased-competition theory, we can apply his framework to Cognitive Unison theory and therefore arrive at a non-representational empirical implementation of it.

Cognitive Unison theory understands attention to be a type of personal level cognition which is empirically implemented via biased-competition. Although Mole's own version of the view requires representation, I have explained how the view can be accepted without requiring representation— we accept a non-representational account of personal level cognition and a non-representational account of biased-competition (biased-affordance-competition). At this point, we have arrived at a non-representational (adverbial) theory of attention.

## 2.2 'Radical' Sensorimotor Enactivism

Having explained how the key sensorimotor enactive concepts of "sensorimotor knowledge" and "attention" could both be accounted for without requiring representation, we can now provide a thoroughly non-representational version of sensorimotor enactivism. At this point, we therefore arrive at radical sensorimotor enactivism (RSE). RSE is predicated upon the sensorimotor enactive account of conscious perception. As such, it understands perception to be constituted by a direct relation between perceiver and environment which is enabled by the possession and exercise of sensorimotor knowledge. The organism becomes conscious of their perceptual relation to the environment only when they attend to it. RSE improves upon sensorimotor enactivism because: it provides a non-representational account of sensorimotor knowledge; and, it provides a non-representational account of attention.[4]

In order to better appreciate RSE, it will help to consider an example. Imagine the following scenario— two people who take the exact same route whilst walking through a park, and who carry out the same perceptual processes. One of the walkers, who I will label the "*mindful-walker*", is walking through the park whilst practising mindfulness meditation. This being the case, they are consciously aware of their various perceptual relations to the environment as they make their way through the park. They can see the bright red of the rose in front of them, feel the cool breeze on their neck, and hear the gentling rustling of leaves behind them. The other walker, although taking the exact same route through the park, is not consciously aware of their perceptual relation to the environment. This *worrying-walker* is currently in the midst of a financial crisis, and they are trying to work out how best to extricate themselves from their distressing predicament. Because both walkers take the same routes and engage in the same perceptual processing, each walker can be said to perceive. This explains why neither walker trips over branches, why both divert their gaze when they (accidently) look straight at

3 I derive my use of this phrase from Clark's discussion of TALoNS (Clark 2016).

4 Of course, my claim— that RSE improves upon sensorimotor enactivism— is reliant on the prior assumption that non-representational theories of conscious perception are to be preferred over representational ones. Although I do think that non-representational theories of conscious perception should be preferred over representational ones (see Anderson 2014; Chemero 2009; Downey 2016, ch. 3; Hutto and Myin 2013), I will not rely on this claim here. Rather, I will simply re-iterate an earlier noted point: sensorimotor enactivism has been advanced primarily as a non-representational theory and draws its intellectual roots from the non-representational tradition of enactive and ecological approaches to mentality. Consequently, a non-representational version of the theory is to be preferred.

the sun, and so on and so forth. However, although both walkers are perceiving, only one walker is conscious of their perceptual states— the mindful-walker. RSE explains the difference between each walker in terms of attention: the mindful-walker is engaged in 'attentive-perception' whereas the worrying-walker is not. Whilst the former is using their cognitive resources in unison to attend to the task of perception, the latter is using their cognitive resources in unison to attend to their financial problems. Thus the difference between the two walkers, and the reason why one is conscious of their perceptual relation to the environment whilst the other is not, is an adverbial one. It concerns the manner in which each walker engages in the task of perception, and so is not solely concerned with the perceptual processing itself. Although both walkers are engaged in the same sort of perceptual processing, only one of these walkers (the mindful-walker) carries out the perceptual processing 'attentively'.

## 3 RSE, Brains, and Predictive Processing

It is often objected that enactive and ecological accounts of conscious perception, such as RSE, are empirically inadequate. These theories are often considered to provide mere descriptions of conscious perception, whilst failing to provide any substantive empirical insights or interesting hypotheses. In particular, it is often argued that these theories are incapable of accounting for the (undoubtedly key) role of the brain in conscious perception (Chemero 2009, p. 93; Clark 2009; Seth 2014). The empirical paradigm of predictive processing, on the other hand, is generally considered to provide a good account of the brain's role in conscious perception (Clark 2013; Clark 2016; Hohwy 2013; Hohwy 2016; Seth 2014). In this section, I am going to explain how predictive processing can be subsumed within RSE. Therefore, I will conclude that RSE can account for the brain's role in conscious perception, and so RSE cannot be rejected for failing to do so.

### 3.1 What is Predictive Processing?

Predictive Processing (PP) conceives of the brain as a prediction machine whose sole role is to minimise prediction-error (Clark 2013; Clark 2016; Hohwy 2013). According to this theory, the brain uses Bayes' theorem (or an approximation thereof) to update *internal generative models*, which are used to derive a best-guess as to the external causes of its current sensory input. The brain is constantly updating these models in response to *error-signals* (which mark a divergence between predicted input and actual input) and this occurs via the process of *active inference*. In active inference, the brain can either: change its model to fit the incoming sensory stream, or, change the incoming sensory stream to fit its model (for example, by moving). Perception, on this account, is constituted by the brain's expectations as to the external causes of its perceptual input.[5] Importantly, the terms "prediction" and "expectation" are wholly sub-personal and non-conscious. We, qua personal level agents, are not aware of the Bayesian nature of our perceptual processes. Rather, these perceptual processes are carried out at, and applicable to, the sub-personal level of explanation.

Given that the world we live in is inherently uncertain and unpredictable, it is inevitable that there will be errors within the brain's internal generative models. In order to make the processing of error-signals manageable, the brain must have a mechanism for determining which error-signals should be scrutinised more carefully, and which should be ignored. PP theorists suggest that the brain can do this via assignments of *precision-weighting*. These assignments concern the amount of 'weight' or 'gain' which will be accorded to a given error-signal. The more weight or gain an error-signal receives, the more it will be able to influence the internal generative models created by the brain. Analogously, the lower the weight or gain assigned to a given error-signal, the less influence this signal is likely to have on the brain's internal generative models. Precision-weighting thus provides the means through which the brain can make its task (of minimising prediction-error) manageable.

5 Note, these expectations are considered to be essentially *action-oriented* (Wiese and Metzinger 2017).

## 3.2 Non-Representational Predictive Processing

PP has been advanced and developed as a theory which is situated firmly within the cognitivist paradigm in cognitive science (Von Helmholtz 1867; Gregory 1980). As such, it is unsurprising that PP is generally taken to indispensably require representation (Clark 2016; Gładziejewski 2016; Hohwy 2013, ch. 8). In order to make PP compatible with RSE, I must provide a non-representational account of its key posits. Orlandi has recently provided such an account, arguing that the key PP posits of "prediction-signal", "error-signal", "prior probability", "hyper-prior", and "likelihood" all fail Ramsey's *job-description challenge.* As such, she concludes that these posits do not deserve a representational status. In this sub-section, I will summarise Orlandi's argument.[6]

In his 2009 *Representation Reconsidered,* Ramsey proposes a strategy for determining whether or not a given mechanism deserves to be described in terms of representation. In order to count as representational, Ramsey argues that the mechanism in question must play a causal and functional role in a system which is recogniseably representational. He claims that we should determine whether or not its role is recogniseably representational by comparing the properties of the mechanism with the properties of an artefact which we would pre-theoretically regard as representational. If the mechanism fulfils the same functional role as a proto-typically representational artefact, then that mechanism passes the job-description challenge and so deserves to be described in terms of representation.

Orlandi applies this argumentative strategy to PP and concludes that its posits fail the job-description challenge (Orlandi 2015; *cf.* Orlandi 2014). She first argues that the concepts "prediction-signal" and "error-signal" (which are often applied to low- and mid- levels of the neural hierarchy by PP theorists) are concerned only with proximal conditions and so can be explained in terms of causal covariation. She then argues that causal covariation accounts of representation are not truly representational because they fail the job-description challenge— these accounts concern only correlation between neural events, and such correlation is more naturally described in terms of causal mediation than in terms of representation (*cf.* Hutto and Myin 2013, ch. 4; Ramsey 2009, ch. 4). Consequently, she concludes that prediction- and error- signals in the brain should be described as non-representational causal mediators. Finally, she argues that the concepts of "prior", "hyper-prior", and "likelihood" themselves fail the job-description challenge because they act only as biases which pre-dispose the brain to enter into certain neural arrangements.[7] Consider, for example, the functional role played by a pump in a typical water-fountain. When water reaches the bottom bowl of a fountain, it pools around the opening to a pump. This pump sucks a small proportion of water in and pushes it back up to the top of the fountain, where the water once more begins to trickle down. Only a small proportion of water will be pumped up to the top of the fountain at any given time, because if too much were to flow from the top at once the fountain would malfunction (e.g. it may overflow). The pump therefore plays the functional role of biasing the flow of a system of water, such that there will always be a large pool of water at the bottom of the fountain and only a small amount at the top. We would not, of course, describe the biasing function of this pump in terms of representation. According to Orlandi, higher-level aspects of the PP neural hierarchy are best described as playing exactly the same sort of biasing role within the brain. Just as the pump in a water-fountain plays the role of pre-disposing the fountain to instantiate a certain water-cycle when presented with water, so too do higher-level posits of PP play the non-representational function of pre-disposing neural systems to enter into certain arrangements whenever presented with a given environmental stimulus. Consequently, Orlandi concludes that PP processing is entirely non-representational.[8]

6 My thanks to Zoe Drayson , for bringing Orlandi's work to my attention.

7 Orlandi provides a number of other arguments against treating higher-level PP processing in terms of representation, but spatial constraints require that I do not rehearse these arguments here.

8 Orlandi does, however, conclude that the results of this processing (the 'winning hypothesis') should be described in terms of representation. I explain why I think this conclusion should be rejected, in the context of RSE, in footnote 11 of this paper.

Although this conclusion may appear merely cosmetic or superficial— we replace the word "representation" with "causal-mediation" or "non-representational bias" when describing PP processing— it in fact has profound empirical consequences. If low- and mid- level aspects of PP processing involve mere causal mediation, then empirical work on PP has much more in common with pre-cognitivist paradigms in the mind sciences (such as psychological behaviourism) than is often realized (*cf.* Ramsey 2009). Furthermore, if higher-level PP posits are to be understood as non-representational biases, then this should lead to an emphasis on empirical frameworks which conceive of neural processing in such terms. We should therefore be led to emphasise empirical approaches such as *visual scence statistics* (championed by Orlandi) or Anderson's biased-affordance-competition framework (Anderson 2014; Anderson and Finlay 2014) because these approaches can be used to study and explain how non-representational biases within the brain are created and maintained. Thus, accepting a non-representational account of PP processing has consequences not only for our understanding of current empirical work, but also for the kind of empirical work we subsequently engage in.

In short, although PP is generally assumed to indispensably require representation, a closer look at the actual causal role its posits play within the neural hierarchy should lead one to the conclusion that PP does not require representation. Prediction- and error- signals should be understood in terms of non-representational causal mediation, whilst priors, hyper-priors, and likelihoods should be understood as non-representational neuronal biases. This conclusion has empirical consequences because it requires a re-conceptualisation of extant empirical work and favours certain approaches to future research. Having arrived at a non-representational account of PP,[9] I am now going to explain how this account could be subsumed within RSE.

### 3.3 RSE and Non-Representational PP

In section two of this paper, we saw that RSE explains perception to be predicated on an organism's possession of sensorimotor knowledge and conscious perception in terms of its ability to 'attentively perceive'. Having stated that PP does account for brain-based processing, and having argued for a non-representational version of PP, I am now going to explain how this account can be subsumed within RSE. I will explain how non-representational PP can be used to provide an empirical explanation of the sub-personal aspects of RSE. Consequently, I will show that RSE *can* account for the brain's role in conscious perception.

PP describes perception as a sub-personal process concerning action-oriented expectations about perceptual stimulation. As such, it matches exactly the description of sub-personal sensorimotor knowledge. PP can therefore be used to provide an operationalisation of brain-based sensorimotor knowledge (*cf.* Seth 2014). The reader should recall that sub-personal sensorimotor knowledge was explained to concern a series of relations between certain sensory inputs and certain other neural outputs. And we have already seen that PP processing should be understood in terms of causal mediation and biased neural processing. PP can therefore be used to explain the specific relations between certain neural inputs and certain other neural outputs. By using PP to explain a certain facet of neural perceptual processing and providing a non-representational account of its posits, one could therefore arrive at a fully worked out empirical explanation of the brain-based, sub-personal aspects of sensorimotor knowledge (as that concept is understood on RSE).

In addition to explaining sensorimotor knowledge, PP can also be applied to, and used to improve, Anderson's biased-affordance-competition framework (Clark 2016, ch. 5). Clark notes that PP fits extremely well with Anderson's framework because it too offers up an action-oriented view of perception which allows for extreme neural plasticity. Clark argues that PP improves upon Anderson's framework

9 It is worth noting that Orlandi's argument is controversial. Although I find her general line of argument convincing, many proponents of PP are likely to object to it (especially those who advocate so-called *conservative predictive processing* (Gładziejewski 2016; Hohwy 2013; *cf.* Clark 2015). Spatial constraints dictate that I do not delve further into the details of this debate here. However, the interested reader can consult (Bruineberg & Rietveld 2014; Chemero 2009; Downey unpublished; Ramsey 2009) for arguments germane to the one endorsed here.

because the concept of "precision-weighting" can be used to explain how the brain manages to configure (and re-configure) TALoNS on extremely rapid time-scales. He does so by noting that the influence a given neural area has on any other given neural area can be determined via precision-weighting. If signals from a particular area are given high precision, then they are likely to be propagated to other areas of the brain. If, however, they are given a low weighting, it is likely they will have very little influence within the neural system. Importantly, the precision-weighting assigned to a given signal is itself constantly in flux and can be rapidly altered. As such, precision-weighting allows for the rapid creation, and dissolution, of various neuronal coalitions. It provides the means through which different neural areas can be allowed to influence, or be prevented from influencing, one another.[10] Thus, not only is PP compatible with Anderson's framework, it can in fact improve upon it.[11]

At the sub-personal level of empirical implementation, RSE explains perception in terms of sensorimotor knowledge and attention in terms of biased-affordance-competition. In this sub-section, we have seen that non-representational PP can be used to operationalise sensorimotor knowledge and that it can be used to improve Anderson's biased-affordance-competition framework. Non-representational PP can, therefore, be used to provide an empirical explanation of the sub-personal aspects of RSE. By subsuming non-representational PP within RSE, we arrive at an empirically satisfactory explanation of the brain's role in conscious perception. Thus, it cannot be objected that RSE ignores the brain's role in conscious perception. In short— PP is ideally suited to play the (extremely important) role of explaining the sub-personal, brain-based aspects of RSE.

## 4 RSE Improves Upon Cognitivist PP

Thus far, I have outlined RSE and explained how applying a non-representational version of PP to the framework allows for an empirically adequate explanation of the brain's role in conscious perception. In this section I am going to go further, and argue that RSE provides a better account of the brain's role in conscious perception than PP taken as a stand-alone theory. Consequently, I conclude that not only *can* a non-representational version of PP be used to explain the sub-personal aspects of conscious perception on RSE. Such a use of PP is in fact *preferable*, because one arrives at a better account of conscious perception than that which can be provided by PP alone.

### 4.1 Benefit One— Explaining Levels of Explanation

One of the chief benefits of accepting RSE and combining it with non-representational PP is that one arrives at a straight-forward account of the relation between the sub-personal, personal, and conscious levels of explanation. RSE explains personal level perception in terms of behavioural dispositions. The categorical basis of these behavioural dispositions is then explained to be brain-based— organisms are capable of perceiving only if they possess brain-based sensorimotor knowledge.[12] This brain-based sensorimotor knowledge is itself understood to be constituted by a series of relations between certain sensory inputs and certain other motor outputs, and can be operationalised via a non-representational version of PP. Therefore, on RSE, the brain is considered to 'give rise' to personal level perception by controlling and driving behaviour. Conscious perception is then, similarly, explained entirely in terms

10 On a non-representational account of PP, precision-weighting should itself be explained as a (particularly fast-acting) neurochemically mediated bias.

11 By conjoining biased-affordance-competition and PP, we are able to reject Orlandi's conclusion that the results of PP processing are representational. Orlandi concludes that the 'winning hypothesis' should be considered representational because it fulfils her three criteria for representation: it is concerned with distal events; it is de-coupleable; and, it is used for the planning of organismal action. A conjunction of biased-affordance-competition and PP will identify the 'winning hypothesis' with the 'winning affordance'. Because affordances are directly perceived by organisms, they neither concern distal events nor are they de-coupleable (at least, not in any interesting sense which would require representation, see Anderson 2014; Chemero 2009). Although the 'winning affordance' is used for the planning of organismal action, meeting this criterion alone is not sufficient for the 'winning affordance' to be ascribed a representational status (unless one deflates the meaning of representation so much that the concept becomes empirically vacuous, see Ramsey 2009; Ramsey 2015). Thus, on RSE, the 'winning hypothesis' should not be described in terms of representation.

12 This account of the relation between sub-personal and personal levels of description is consonant with that outlined in (McDowell 1994). My thanks to an anonymous referee for helping me to clarify this point.

of behaviour. On this theory, conscious perception occurs when the perceiving organism's behavioural interaction with the environment is carried out 'attentively', with 'attentive perception' empirically implemented in the brain via biased-affordance-competition. Thus, RSE attributes to the brain the role of controlling behaviour, and it explains both perception and conscious perception entirely in terms of behaviour.

There is absolutely nothing mysterious or naturalistically unacceptable about the idea that a brain could control behaviour. Therefore, there is absolutely nothing mysterious about the relation between the sub-personal and personal levels of explanation on RSE. Furthermore, on this framework there is nothing mysterious about the relation between conscious and unconscious personal level perception. Personal level perception itself occurs when an organism exercises sensorimotor knowledge. If the exercise of this knowledge is performed attentively, then the organism's perceptual relation to the environment will become conscious. As such, the difference between conscious and unconscious perceptual processing is explained to be adverbial in nature, and there is nothing metaphysically suspicious or naturalistically awry with the existence of adverbial behaviour. Consequently, RSE provides a clear distinction between the sub-personal, personal, and conscious aspects of perception. Moreover, RSE makes sense of their existence and inter-relation without requiring any leaps of imaginative faith or speculative metaphysical theorising.

Importantly, RSE can provide this metaphysically innocuous construal of the sub-personal, personal, and conscious levels of explanation whilst providing a phenomenologically *compelling* account of (conscious) perception (Noë 2004; Ward 2012).[13] Consider, in this vein, the sensorimotor enactive explanation of the difference between certain perceptual modalities. Each modality is considered to possess its own specific set of sensorimotor 'laws', which concern the law-like relation between movement and stimulation specific to the sensory modality in question. Visual sensorimotor 'laws', for example, concern the fact that objects will loom as we get closer to them, appear smaller as we move further away, and disappear from view if we close our eyes. In the case of audition, however, sound gets louder as we move closer to its source, it gets quieter as we move away, and closing one's eyes will have little (or no) effect on hearing. We can, therefore, provide a distinction between different modalities of perception by explaining the relevant modality-specific sensorimotor 'laws'.

Although these sensorimotor 'laws' are described entirely in terms of an organism's perceptual behaviour, they are also phenomenologically intuitive:

> When it is brought to our attention that certain sensorimotor contingencies are characteristic of vision, others of hearing, others of touch, there is an 'aha!' response. (Hurley and Noë 2003, p. 146)

RSE is simply a non-representational version of sensorimotor enactivism. As such, accepting RSE allows one to explain (conscious) perception to be constituted entirely by behavioural interactions with the environment, whilst providing a phenomenologically intuitive account of (conscious) perception.

This presents an improvement upon PP, if taken as a stand-alone theory of conscious perception, because that theory faces the (quite familiar) problem of having to provide a phenomenologically compelling explanation of how sub-personal, non-conscious, brain-based processing could 'give rise' to personal level, conscious perception. Although this is an active area of research within PP, barring a complete conceptual or scientific revolution, it is difficult to see how any particular brain-based

13 It is on this point that RSE departs from, and improves upon, other theories which take conscious perception to be constituted by behaviour (Dennett 1993; Ryle 1949/2000; Wittgenstein 1953/2009). Unlike these behaviourist theses, which (arguably) fall prey to the problem of verificationism, RSE can provide a phenomenologically compelling account of conscious perception whilst taking conscious perception to be constituted by behaviour. The main reason that RSE avoids the problem of verificationism is that it presents one with a *dynamic* view of conscious perception, upon which the environment itself constitutively plays an *active* role in proceedings. For an extended argument for this point, the interested reader can consult (Downey 2016, ch. 3; *cf.* Clark and Chalmers 1998; Hurley 1998, p. 420-22; Hurley 2001).

account could provide a phenomenologically compelling explanation of the inter-relation between sub-personal, personal, and conscious levels of explanation. As Hurley and Noë explain:

> By contrast, if it is brought to our attention that activity in a certain brain area is correlated with vision, we do indeed still want to ask: "But why does brain activity there go with what it is like to see, rather than to hear or touch?" (Hurley & Noë 2003, p. 147)

The problem with brain-based accounts is that, regardless of the specifics of the account in question, the identification of a certain brain-based process with a certain experience is always going to appear arbitrary. Furthermore, there is always going to be an air of mystery surrounding how the phenomenology of conscious perception could be contained within, or identified with, the brain. Even the most enthusiastic proponents of PP admit that this is a problem for the view, and they are quite forth-right in conceding that there has, as yet, been no concrete proposals as to how PP could provide a novel solution to it (Hohwy 2013, p. 202; Clark 2016, ch. 7, §16).[14]

In summary, RSE explains (conscious) perception entirely in terms of behavior (both organismal and environmental), and it does so whilst accounting for the phenomenology of conscious perception. Moreover, it provides this phenomenologically plausible account whilst providing an explanation as to how the sub-personal, personal, and conscious levels of explanation inter-relate without invoking any naturalistically mysterious or metaphysically suspect properties or posits. PP, when taken as a stand-alone theory of conscious perception, cannot provide such an account. RSE provides a more elegant and phenomenologically plausible account of the relation between different levels of explanation in conscious perception than PP taken alone. Therefore, RSE should be preferred as a theory of conscious perception on this basis.

## 4.2 Benefit Two— Accounting for Empirical Data

The second major benefit accrued by accepting RSE concerns its empirical consequences. RSE is advanced within the intellectual tradition of enactive and ecological approaches to conscious perception. PP, however, is advanced from within the intellectual tradition of cognitivist approaches to conscious perception. Although each of these traditions boasts their own empirically successful research programmes, they are generally thought to directly conflict with one another. Consequently, it is usually thought that acceptance of one approach requires a wholesale rejection of the other (Chemero 2009; Hohwy 2016). RSE does not, however, require a wholesale rejection of cognitivist insights. Quite the contrary, in fact— not only is PP compatible with RSE, it actually constitutes a crucial component of the overall RSE framework (as we have seen). Thus, by accepting RSE and taking non-representational PP to provide an explanation and implementation of the sub-personal aspects of that framework, we arrive at a theory which can take advantage of the empirical work carried out on behalf of both enactive/ecological and cognitivist traditions.

This is beneficial because both of these paradigms have given rise to empirically productive research programmes, which have led to numerous novel and predicted empirical results. Scientific research programmes survive largely on the basis of their empirical productivity, with empirical productivity itself generally thought to require an inference to the best explanation— the theory is empirically productive because it accurately describes its domain of study.[15] Although one can reject even empirically successful research programmes (for example, because one thinks they do not correctly describe a given domain of study), in order to do so one must explain how the science can be successful despite presenting us with an incorrect model of the world. Because RSE can accept empirical work carried

14 For an extended argument for this point, the interested reader can consult (Downey 2016, ch. 3).

15 Of course, there is a large literature on inference to the best explanation (and the related topics of realism, instrumentalism, and anti-realism) within the philosophy of science. I do not intend to presume a definitive answer to questions within this topic here. Rather, I am simply making the point that, by and large, the success of a scientific research programme usually gives us (defeasible) reasons to accept realism about its posits.

out in both traditions, it can simply side-step this problem. There is no need to reject, or otherwise eliminate, the vast swathes of empirical work carried out in either scientific tradition. Thus, on RSE, we have no need to reject an inference to the best explanation in either domain of study.

In fact, not only can RSE accept empirical work carried out on behalf of two traditionally opposed scientific frameworks, it actually can be used to help illuminate the distinction between the two research traditions and provide guidance for future empirical research. RSE equates the sub-personal level with the brain, and explains the brain's role to be that of controlling behaviour. Furthermore, it champions non-representational PP as the theory which should be used to study the brain. Thus, if we are interested in investigating the sub-personal aspects of RSE, we can do so by applying the conceptual and empirical tools of non-representational PP to the study of the brain (see, for example, Hohwy et al. 2008).[16] If, however, we are more interested in investigating the personal level of explanation, then we can do so by using the methods of enactive and ecological approaches to study the interaction, and inter-relation between, the organism and its environment (see, for example, Chemero 2009). Finally, if we wish to study conscious perception, we can simply study the manner in which the perceiving organism is able to perceive 'attentively' (Anderson 2014; Mole 2011). In short, RSE helps to demarcate between the different levels of explanation within the science of conscious perception and therefore provides guidance as to which tools and techniques are appropriate for a given area of study.

The second benefit of accepting RSE, then, is that one arrives at an empirical integration of enactive/ecological and cognitivist approaches to conscious perception. This is beneficial because theorists can then take advantage of the excellent empirical work carried out within both traditions. Furthermore, this conjunction provides a clear and clean conceptual distinction between empirical work on the sub-personal, personal, and conscious levels of perception that provides guidance for the methods and frameworks which scientists should be using to study a given aspect of conscious perception. Thus, not only does RSE help to simplify and clarify the conceptual terrain of empirical work, it also provides empirical guidance.

## 5 Conclusion

In this paper I have outlined RSE, explained how a non-representational version of PP can be used to empirically explain its sub-personal aspects, and argued that the resulting account of conscious perception is to be preferred over PP taken as a stand-alone theory. I began by outlining the sensorimotor enactive theory of conscious perception, and explaining that this theory is problematic because two of its key posits ("sensorimotor knowledge" and "attention") either require representation or are left explanatorily vacuous. I argued for an account of sensorimotor knowledge in which it is taken to be constituted entirely by non-representational causal mediation and/or behavioural dispositions. Then, I outlined a non-representational and adverbial theory of attention and argued that it should be applied to sensorimotor enactivism. As such, I arrived at a thoroughly non-representational version of sensorimotor enactivism, and so at *radical* sensorimotor enactivism.

It is often objected that theories such as RSE are empirically vacuous. In particular, it is often argued that these theories are incapable of accounting for the brain's role in mentality. I outlined a non-representational version of PP and explained how it could be used to empirically explain the sub-personal, brain-based aspects of RSE. Therefore, I concluded that RSE cannot be objected to on the basis that it ignores the brain's role in conscious perception. Then, I explained why RSE should in fact be preferred as an account of conscious perception over rival cognitivist theories (such as representational PP). I argued that RSE provides a better account of the inter-relation between the sub-personal, personal,

16 In their account of binocular rivalry Hohwy *et al.* accept a representational version of PP. In order to subsume this account within RSE, we would therefore have to accept a non-representational account of PP posits. We have already seen that such an account can be given. In the case of rivalry, it is the PP framework itself (and not its representational posits) which plays a key explanatory role (Anderson and Chemero 2013). Therefore, accepting a non-representational account of PP's posits would not weaken the explanation of rivalry provided by Hohwy *et al.*

and conscious levels of explanation than cognitivist theories, and that it can account for and guide a larger amount of empirical research.

RSE is a novel theory of conscious perception which provides an over-arching conceptual framework for the scientific study of conscious perception. It promises to unite a number of (seemingly incompatible) strands of empirical cognitive science whilst demystifying the very existence of conscious perception. In addition to taxonomising different areas of extant research and clarifying their scope and inter-relation, RSE provides guidance for the direction of future empirical work. The phenomenon of conscious perception has only recently been submitted to sustained scientific scrutiny. Although there has of late been an explosion of empirical work on the topic, the empirical evidence accrued vastly out-weighs our ability to taxonomise and understand it. RSE provides a framework which can help to simplify this task substantially. I therefore conclude, on this basis, that it is worthy of further research, development, and critical scrutiny.

## References


Anderson, M. (2014). *After phrenology: Neural reuse and the interactive brain.* Cambridge MA: MIT Press.

——— (2015). Précis of after phrenology: Neural reuse and the interactive brain. *Behavioural and Brain Sciences*, 1-22.

Anderson, M. L. & Chemero, T. (2013). The problem with brain GUTs: conflation of different senses of "prediction" threatens metaphysical disaster. *Behavioral and Brain Sciences, 36* (3), 204–205.

Anderson, M. & Finlay, B. (2014). Allocating structure to function: The strong links between neuroplasticity and natural selection. *Frontiers In Human Neuroscience* (7), 918.

Barrett, F. (2011). *Beyond the brain: How body and environment shape animal and human minds.* New Jersey: Princeton University Press.

Block, N. (2001). Behaviourism revisited. *Behavioural and Brain Sciences, 24* (5), 977- 978.

Bruineberg, J. & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience,* http://dx.doi.org/10.3389/fnhum.2014.00599.

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

Cisek, P. (2015). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society of London B (Biological Sciences), 362* (1485), 1585–1599.

Clark, A. (2009). Spreading the joy? Why the machinery of consciousness is (probably) still in the head. *Mind, 118* (472), 963-993.

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

——— (2015). Radical predictive processing. *The Southern Journal of Philosophy, 53* (1), 3-27.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

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

Dennett, D. (1993). *Consciousness explained.* London: Penguin Classics.

Desimone, R. & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*, 193-222.

Downey, A. (2016). *Radical sensorimotor enactivism.*

——— (unpublished). Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Manuscript*.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Gregory, R. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London B (Biological Sciences)* (290), 181-197.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese* (193), 559.

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

———(2016). The self-evidencing brain. *Noûs, 50* (2), 259–285.

Hohwy, J., Roepstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition, 108* (3), 687–701.

Hurley, S. (1998). *Consciousness in action.* Cambridge, MA: Harvard University Press.

———(2001). Perception and action: Alternative views. *Synthese, 129*, 3-40.

Hurley, S. & Noë, A. (2003). Neural plasticity and consciousness. *Biology and Philosophy, 18* (1), 131-168.

Hutto, D. (2005). Knowing what? Radical versus conservative enactivism. *Phenomenology and the Cognitive Sciences, 4*, 389–405.

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

McDowell, J. (1994). The content of perceptual experience. *The Philosophical Quarterly, 44* (175), 190-205.

Mole, C. (2011). *Attention is cognitive unison: An essay in philosophical psychology.* Oxford: Oxford University Press.

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

O'Regan, J. K. (2011). *Why red doesn't sound like a bell: Understanding the feel of consciousness.* Oxford: Oxford University Press.

O'Regan, J. K. & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioural and Brain Sciences, 5*, 939-73.

Orlandi, N. (2014). *The innocent eye: Why vision is not a cognitive process.* Oxford: Oxford University Press.

———(2015). Bayesian perception is ecological perception. *BrainsBlog*, 1-19.

Ramsey, W. (2009). *Representation reconsidered.* Cambridge: Cambridge University Press.

———(2015). Must cognition be representational? *Synthese*, 1-18.

Ryle, G. (1949/2000). *The concept of mind.* Chicago: Chicago University Press.

Seth, A. (2014). A predictive processing theory of sensorimotor contingencies. *Cognitive Neuroscience, 5* (2), 97–118.

Sporns, O. (2010). *Networks of the brain.* Cambridge, MA: MIT Press.

Varela, F., Thompson, E. & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience.* Cambridge, MA: MIT Press.

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik.* Leipzig: Leopold Voss.

Ward, D. (2012). Enjoying the spread: Conscious externalism reconsidered. *Mind, 121* (483), 731-751.

Wiese, W. & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing.* Frankfurt am Main: MIND Group.

Wittgenstein, L. (1953/2009). *Philosophical investigations.* London: Wiley-Blackwell.

# Modularity and the Predictive Mind

Zoe Drayson

Modular approaches to the architecture of the mind claim that some mental mechanisms, such as sensory input processes, operate in special-purpose subsystems that are functionally independent from the rest of the mind. This assumption of modularity seems to be in tension with recent claims that the mind has a predictive architecture. Predictive approaches propose that both sensory processing and higher-level processing are part of the same Bayesian information-processing hierarchy, with no clear boundary between perception and cognition. Furthermore, it is not clear how any part of the predictive architecture could be functionally independent, given that each level of the hierarchy is influenced by the level above. Both the assumption of continuity across the predictive architecture and the seeming non-isolability of parts of the predictive architecture seem to be at odds with the modular approach. I explore and ultimately reject the predictive approach's apparent commitments to continuity and non-isolation. I argue that predictive architectures can be modular architectures, and that we should in fact expect predictive architectures to exhibit some form of modularity.

**Keywords**

Bayesian computation | Cognitive architecture | Cognitive penetration | Information encapsulation | Modularity | Predictive processing

**Acknowledgements**

This paper was presented at MIND 23: The Philosophy of Predictive Processing at the Frankfurt Institute for Advanced Studies in May 2016. An earlier version was presented at the University of Bergen in 2015. Many thanks to both audiences for helpful feedback and questions. Particular thanks to Christopher Burr, Max Jones, and two anonymous readers for the MIND Group's "Philosophy and Predictive Processing" project.

## 1 Introduction

The aim of this paper is to explore the relationship between two approaches to the architecture of the mind: modular architectures, in which at least some mental mechanisms operate in functionally-isolated special-purpose modules; and predictive architectures, in which mental mechanisms are integrated into a hierarchy of Bayesian computational processes. The two approaches are often discussed as though they are in tension with each other, but I aim to show that predictive approaches are consistent with modular approaches, and that predictive architectures may in fact entail some form of modularity.

Claims about the modularity of mind take various forms, but all proponents of modular architectures agree that there are special-purpose sensory subsystems which are functionally isolated from other mental processes. The more modest versions of modularity propose that only this subset of mental processes are modular, while amodal reasoning and thought are non-modular. Proponents of 'massive modularity', on the other hand, propose that all mental processes are modular. In this paper, I'll be focusing on a commitment that is shared by proponents of both modest and massive modularity: the claim that there is some stage of sensory processing which is isolated from amodal thought processes. This seems to be the feature of modular architectures that is most in tension with the predictive approach to the mind.

Predictive approaches to the mind propose that our mental architecture should be understood as a hierarchy of Bayesian computational mechanisms. Cognitive information higher in the hierarchy is used to generate predictions about lower-level sensory information, and information about prediction

Drayson, Z. (2017). Modularity and the Predictive Mind.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 12. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573130

errors is passed back up the hierarchy. Proponents of the predictive approach, most notably Clark and Hohwy, claim that there is no obvious boundary or distinction to be drawn between the kinds of processing found at different levels of the hierarchy, which seems to be at odds with the idea of modular decomposition. Additionally, the Bayesian priors for each level in the hierarchy are provided by the level above, which makes it difficult to see how a lower-level sensory process could be informationally isolated from higher-level cognition. These features of the predictive approach are thus *prima facie* in tension with the modular approach to the architecture of the mind.

In this paper, I explore the apparent tensions between the predictive and modular approaches to the mind, and suggest that they can be reconciled. I start by introducing modular architectures (Section 2) and predictive architectures (Section 3). I outline the predictive approach's commitments to the continuity of perception and cognition, and to the non-isolation of sensory processing from higher level informational influence (Section 4). I then challenge these commitments (Section 5). I argue that continuity at one level of a mental architecture is compatible with discontinuity at a different level, and that continuity alone cannot provide an argument against modularity. I also argue that the predictive approach's commitment to top-down influences on lower-level processes does not entail the lack of functional isolability that would be inconsistent with modularity. Furthermore, I show that there are ways of understanding modularity on which predictive approaches seem to require a sort of modularity.

## 2 The Modular Mind

To claim that the mind is modular is to posit a functional distinction between different kinds of mental processes. The modular approach to the mind claims that certain of our mental processes are functionally distinct, in the sense of being independent or isolable from other mental processes. Sensory input processes are often cited as an example of modules in virtue of exhibiting the relevant sort of functional independence: if the data on one's relevant sensory receptors are processed without drawing on information elsewhere in the system, such as data at one's other sensory receptors or one's beliefs more generally, then that sensory process is functionally independent from those other mental processes. To the extent that a sensory process relies only on its own proprietary store of information, it operates in functional isolation from other mental processes. There is much debate over precisely how to specify the sort of independence and isolation that matter for modularity, some of which I'll touch on later in this section. It is important to notice, however, that these notions of independence and isolation associated with modular architectures are functional rather than structural. In particular, modular approaches to the mind are not claims about neural localization: whether a functionally-isolated information process is also neurally isolated is a matter for empirical investigation.

Modular architectures come in different strengths. 'Modestly modular' views, such as those of Fodor (Fodor 1983), propose that modularity is largely a feature of sensory input processes (e.g. vision, audition), and sometimes extend this claim to include mental processes such as language comprehension. Each of these modules, however, is assumed to output into the same central cognitive system: a set of non-modular processes which combine the products of the modules with stored beliefs and memories, and which is responsible for amodal thought processes such as general reasoning, belief-revision, planning, and decision-making.

'Massively modular' views, such as those of Peter Carruthers (Carruthers 2006) reject the idea of a central cognitive system in favour of further modular processing. Proponents of massive modularity tend to deny, therefore, that we have general-purpose capacity for reasoning. Instead, they propose that there are distinct modules for the kinds of reasoning related to distinct domains: the reasoning processes that allow us to detect unfair social behavior, for example, might be functionally independent from the reasoning processes that we go through in assessing potential mates.

Whether the architecture in question is modestly or massively modular, the most characteristic feature of modular processes is their functional independence or isolation from other information processes. Beyond that, there is much disagreement over the exact definition of modularity which need not concern us here. And even specifying the precise kind of functional independence exhibited by modules is difficult. One way to characterize it is in terms of information encapsulation, whereby a process is informationally encapsulated to the extent that it lacks access to information stored elsewhere in the architecture (Fodor 1983). Information encapsulation seems to explain why a mental process can be tractable and fast, in virtue of not being able to take the mind's vast amounts of information into consideration. But to use this as the defining feature of a module is somewhat restrictive: we might have reasons to think that a process is functionally independent in some interesting respect even where it does not meet the condition of being informationally encapsulated. As an example, consider sensory input processes such as vision. Visual processes have long been considered to be informationally encapsulated, on the assumption that each different feature of the environment (shape, colour, etc.) is extracted by a dedicated and functionally independent process. Empirical evidence, however, shows interactions within and across sensory modalities to an extent that challenge their claim that informational encapsulation. But visual processes still seem to exhibit degrees of functional independence, in the sense that they are functionally detachable or seperable from other mental processes. Daniel Burnston and Jonathan Cohen, for example, argue that visual processes still exhibit distinct perceptual strategies: while these strategies interact with each other, each perceptual strategy is sensitive to a restricted range of informational parameters. This sensitivity means that each process interfaces with other mental processes in a limited number of ways. This, they suggest, gives us a different way to characterize the sort of functional independence involved in modular processes:

> [W]hat makes modular processes modular, detachable, and in some sense separable from the rest of mentation is that they interface with other aspects of mental processing in a circumscribed number of ways. That is, modular processes are modular just because, and in so far as, there is a delimited range of parameters to which their processing is sensitive. (Burnston and Cohen 2015, p. 132)

In what follows, I won't assume that modular processes need to be informationally encapsulated by definition. I will assume that there needs to be some form of functional independence and isolation, and I'll come back to Burnston and Cohen's characterization of modularity later in the paper.

## 3 The Predictive Mind

The predictive approach to the architecture of the mind goes by a number of names, such as 'the hierarchical predictive processing perspective' (Clark 2013) and the 'the prediction error minimization framework' (Hohwy 2013). There are two fundamental ideas on which it rests: a hierarchical architecture of Bayesian computational processes, and a data-compression strategy called 'predictive coding'. I'll outline each out of these commitments in turn.

Proponents of predictive mental architectures propose that the brain implements Bayesian computational processes: it generates hypotheses or expectations about the world and updates these in light of new evidence in accordance with Bayes' Rule. Bayes' Rule states that the probability of a hypothesis, given the evidence, is updated by considering the product of the likelihood (the probability of the evidence given the hypothesis) and the prior probability of the hypothesis. In the case of visual perception, for example, the claim is that the visual system generates hypotheses about the external world on the basis of its existing information and estimates their prior probability: the probability of their being true before the sensory data has been taken into account. If I walk into my office, for example, the hypothesis that the furniture is located where I left it will generally have a higher prior probability than the hypotheses that the furniture is on the ceiling. The system also has existing information about

what sort of sensory input is caused by certain objects, e.g. what sort of retinal data are associated with viewing certain items of furniture from certain angles. This is what enables the system to calculate the likelihood of the hypothesis once the actual sensory data are known: the probability of that particular sensory input, given the hypothesis in question. The product of the hypothesis' likelihood and its prior probability is what generates the posterior probability of the hypothesis. According to the predictive approach, the hypothesis with the highest posterior probability will determine what I perceive.

This example is oversimplified. In fact, the predictive approach posits many hypothesis-testing computational processes, arranged hierarchically. The hypotheses and predictions at lower levels of the hierarchy tend to be spatially and temporally precise, while those at higher levels are more abstract. The higher-level hypotheses act as Bayesian priors for lower-level processes: each level tries to predict the input to the level below, and then updates the hypothesis accordingly. The predictive approach adopts an empirical Bayes method, on which the priors are estimated from the data rather than fixed pre-observationally, so they are shaped over time from the sensory data. This allows the predictive approach to account for priors without circularity, because the hypotheses determining perceptual experience are not themselves based directly on perceptual experiences but extracted indirectly from higher-level hypotheses. For more on empirical Bayes and the predictive approach, see (Hohwy 2013, p. 33) and (Clark 2013, p. 185).

Predictive mental architectures combine this hierarchy of Bayesian processing with the second key element of the approach: the computational framework of predictive coding. Predictive coding is a way of maximizing the efficiency of an information system, by ensuring that it doesn't process any more information than it needs to. Predictive coding was developed as a data compression strategy by computer scientists to allow more efficient storage and transmission of large files, and works by using the information system's existing information to predict its own expected inputs, so that the system only needs to account for deviations from its predictions. This means that more of its resources can be allocated to novel information. Neuroscientists have suggested that the brain might use a similar strategy to deal with the massive amounts of sensory information it receives: if it can correctly predict at least some of the information received by the sense organs, then it can direct its resources to accounting for novel or unexpected sensory information. (Further details of predictive coding as a neurocomputational strategy can be found in Friston and Stephan 2007.)

Proponents of predictive mental architectures combine the Bayesian computational hierarchy with the predictive coding strategy to argue that the brain doesn't first process all the sensory input available, then construct hypotheses about the probable worldly causes of the sensory input. Instead, hypotheses generated at higher levels of the hierarchy are used to predict the inputs received from the level below. At the lowest level of the hierarchy, the hypotheses are used to predict the data on the sense receptors. If a prediction is correct, there is no need to update the corresponding hypothesis: the input is explained away by it. The corresponding hypothesis only needs to be updated if its prediction is incorrect, that is, if there are discrepancies between the generated prediction and the actual input at a level. In this way, the predictive architecture only has to account for unexpected or novel inputs.

## 4 The Apparent Tension between Modular and Predictive Architectures

There are two features of predictive architectures of the mind that seem to be directly in tension with modular architectures of the mind. First, the predictive approach is often described as lacking any obvious boundary between cognitive processes and perceptual processes in the predictive hierarchy, and therefore as committed to the continuity of cognition and perception. Second, information-processing at each level of the predictive hierarchy incorporates information from the level above, which suggests that no level of processing is isolated from the information elsewhere in the system. In this section, I'll explore these features in greater detail, and show where the *prima facie* tension is supposed to arise.

## 4.1 The Continuity Claim

The continuity claim is the claim that there is no clear divide between cognitive processes and non-cognitive (particularly perceptual) processes in the Bayesian hierarchy. The claim is that the entire predictive hierarchy uses the same Bayesian computational processes, so there is no distinction to be made between the kinds of processes that underlie cognition, on the one hand, and the kinds of processes that underlie perception, on the other hand.

Clark makes the continuity claim when he proposes that predictive architectures depict perception and cognition as "profoundly unified and, in important respects, continuous" (Clark 2013, p. 187). Predictive architectures, he claims, "appear to dissolve […] the superficially clean distinction between perception and knowledge/belief […] we discover no stable or well-specified interface or interfaces between cognition and perception" (Clark 2013, p. 190). Predictive architecture "makes the lines between perception and cognition fuzzy, perhaps even vanishing" (Clark 2013, p. 190).

Hohwy makes the continuity claim when he observes that the predictive approach to mental architecture seems to deny our standard distinction between cognition (understood as involving conceptual thoughts such as beliefs) and perception (understood as involving the processing of sensory experience):

> It [the predictive framework] seems to incorporate concepts and thinking under a broader perceptual inference scheme. […] On this view, concepts and beliefs are fundamentally the same as percepts and experiences, namely expectations (Hohwy 2013, p. 73).

Proponents of predictive architectures thus suggest that our standard tendency to distinguish between believing and perceiving is not supported by their account of mental architecture. They generally allow that processes lower in the hierarchy are doing something akin to sensory processing, in the sense that their predictions are often spatially and temporally precise, and may even refer to these as perceptual processes. They likewise allow that the predictions of higher-level processes are increasingly abstract and more akin to amodal reasoning, and may even refer to these processes as cognitive. But the core of the continuity claim is that there are many levels in the middle of the hierarchy that are not recognizable as standard cases of either perception or cognition: there is no point in the hierarchy at which the processes stop being perceptual and start being cognitive (see also Vetter and Newen 2014).

There is a *prima facie* tension between the continuity claim associated with predictive architectures and the commitments of modularity. Modular architectures seem to respect our standard distinction between perceiving and believing: perception and cognition are assumed to be different kinds of processes (often relying on different forms of representation) with a clear boundary between them. In particular, at least some part of perceptual processing is assumed to be functionally isolated and independent from cognitive processing.

## 4.2 The Non-Isolation Claim

The second relevant commitment of the predictive approach is the non-isolation claim: the claim that there is no part of perceptual processing that is informationally isolated from higher-level cognitive processing. This commitment seems to arise from the hierarchical structure of the Bayesian computational mechanisms, in which the priors at each level in the hierarchy are provided by the level above. This suggests that there is top-down influence at every level. (Recall that the lower down the hierarchy, the more spatiotemporally precise the predictions are; the higher up the hierarchy, the more abstract the predictions are.)

Proponents of predictive architectures seem to assume that this role of the Bayesian priors within the hierarchy entails that no part of the hierarchy is isolated from higher-level information. Hohwy, for example, claims that there is "no theoretical or anatomical border preventing top-down projec-

tions from high to low levels of the perceptual hierarchy" (Hohwy 2013, p. 122). Clark emphasizes that the sorts of abstract predictions made at higher levels can influence the more spatiotemporally precise predictions at lower levels: he claims that perception is "theory-laden" and "knowledge-driven" on the predictive approach, and that "[t]o perceive the world just is to use what you know to explain away the sensory signal" (Clark 2013, p. 190).

In philosophy, perception that is subject to top-down effects is often characterized as 'cognitively penetrable'. Predictive architectures are sometimes described in these terms: Hohwy, for example, suggests that predictive architectures "must induce penetrability of some kind" (Hohwy 2013, p. 120); while Gary Lupyan claims that "[p]redictive systems are penetrable systems" (Lupyan 2015, p. 547) and that we should expect penetrability whenever we have higher-level information processing making predictions about lower-level information processing; and Petra Vetter and Albert Newen claim that the top-down influences on perception suggest that "[p]redictive coding can thus be regarded as an extreme form of cognitive penetration" (Vetter and Newen 2014, p. 72).

If predictive architectures are committed to the cognitive penetration of perception, this seems to put them in tension with modular architectures. Proponents of modular architectures claim that at least some informational processing is isolated from other informational processing. In particular, the relevant claim here is that sensory processing (at least in its early stages) is not influenced by cognitive information in the form of beliefs or memories. Modularity thus provides a mechanism for avoiding cognitive penetration. If proponents of predictive architectures are right that higher-level processes are able to influence lower-level processes in the relevant way, then this suggests that predictive architectures are not modular architectures.

## 5 Predictive Architectures as Modular Architectures

When proponents of predictive architectures propose versions of the continuity claim and the non-isolation claim, they don't always draw an explicit contrast between their approach and modular approaches. But it is often assumed that predictive architectures are non-modular architectures: Jona Vance and Dustin Stokes, for example, describe the predictive approach as involved in "the development of non-modular mental architectures" (Vance and Stokes forthcoming). I want to demonstrate that neither the continuity claim nor the non-isolation claim entail that predictive architectures are non-modular.

First, I'll argue that the continuity claim does not show that there is no distinction between perception and cognition. Then I'll argue against the non-isolation claim, and demonstrate that there is a kind of isolation on the predictive approach that is consistent with some form of modularity.

### 5.1 Challenging the Continuity Claim

As already demonstrated, proponents of predictive architectures argue for the continuity claim on the grounds that the same kinds of Bayesian computational processes are in play throughout the processing hierarchy. This means that at each level in the hierarchy, the same methods of hypothesis testing are used, whether the hypotheses in question are at the spatiotemporally precise or more abstract end of the scale.

This fact in itself, however, does not distinguish predictive architectures from modular architectures. Proponents of modularity can allow that perceptual processes and cognitive processes use the same kind of computational mechanisms: Fodor, for example, thinks that both perceptual and cognitive processes use classical computational inference involving rules and representations (Fodor 1983). Whether perceptual processes use precisely the same kind of format as cognitive processes, e.g. the same 'language of thought', is an empirical matter and thus isn't ruled out simply by adopting a modular architecture (Aydede 2015). When modular approaches distinguish between cognitive processes and perceptual processes, therefore, this distinction can't rely on the claim that the relevant processes

use different computational mechanisms. Since it is consistent with modularity to claim that perception and cognition use the same kinds of computational processes, the continuity of the Bayesian hierarchy on the predictive approach does not demonstrate that predictive architectures are non-modular.

It is true that modular architectures generally distinguish between higher-level cognitive processing and a certain stage of perceptual processing. But this distinction need not be found in the finer-grained details of the computational processes. By way of example, consider the difference between classical computational theories which posit a syntactic process of symbol manipulation, and connectionist theories which posit processes of activation through a network, mediated by connection weights. Connectionist networks may look non-modular in the sense that there are no obvious boundaries between one part of the network and any other part. But a system that is non-modular at one level of description can still be modular at a different level, as Martin Davies has demonstrated: once we observe the connectionist network performing a task, discontinuities can appear and functions can become dissociated. Davies suggests that we employ the idea of coarser and finer grains of modularity, to respect the fact that "[w]hat we really have is a hierarchy of levels of coarser and more detailed interpreted descriptions of the way in which the task is carried out" (Davies 1989, p. 547).

The moral here is that modularity is a matter of grain: a computational system can be modular when viewed at one level of abstraction but not when viewed at another: continuity in the fine-grained details of the information-processing is compatible with discontinuity at a coarser-grained perspective. When proponents of predictive architectures make the continuity claim, therefore, this does not rule out that the system is a modular one. And as long as modularity is not ruled out, it is possible that perceptual processes are distinct from cognitive processes at some appropriately coarse-grained level of description.

Furthermore, even if the predictive approach could demonstrate that there is no clear boundary between perceptual processes and cognitive processes, this would not entail that there is no difference between the two. The lack of clear boundary might suggest that the distinction between perception and cognitive is non-exclusive, for example, such that processes in the middle of the hierarchy are best classified as both perceptual and cognitive; or it might suggest that the distinction between perception and cognition is not exhaustive, such that processes in the middle of the hierarchy are neither perceptual nor cognitive. Both of these approaches are consistent with the claim that there are clear cases of perceptual processing which is not cognitive at the lower end of the hierarchy, and clear cases of cognitive processing which is not perceptual at the higher end of the hierarchy.

## 5.2 Challenging the Non-Isolation Claim

The non-isolation claim is the claim that there is no part of perceptual processing that is informationally isolated from cognitive processing. As already demonstrated, proponents of predictive architectures argue for the non-isolation claim on the grounds that each level in the predictive Bayesian hierarchy has its priors provided by the level above, suggesting that lower-level processes can't be wholly isolated from the influence of higher-level processes. This aspect of predictive archictectures has been interpreted by some as entailing the cognitive penetration of perception.

Care is needed when talking about cognitive penetrability, because some people use the term 'cognitive penetration' to refer to all top-down influences on perception – regardless of which stages of perceptual processes are influenced, which kinds of cognitive processes are doing the influencing, and what the type of influence is. But in philosophy in particular, the label is often reserved for a certain kind of top-down influence on perception: the phenomenon whereby particular kinds of higher-level mental states (notably beliefs and memories) exert a direct influence on the contents of conscious perceptual experience. (See Macpherson forthcoming for an overview of the different varieties of cognitive penetration.)

It is far from clear whether predictive architectures result in cognitive penetration of this sort. On the predictive approach, conscious perceptual experience is the product of the entire prediction minimization process: it is determined by the interactions between top-down and bottom-up information flow within the entire hierarchy, rather than being associated with a particular level in the Bayesian hierarchy (cf. Clark 2013, p. 185). While this might suggest that there generally will be top-down effects on perceptual experience, it doesn't entail that these top-down influences will be of the right kind to constitute cognitive penetration. As an epistemologically interesting phenomenon, cognitive penetration requires that the cognitive states involved are traditionally doxastic: they are the beliefs of the person rather than merely information represented in the cognitive system. (For further elaboration on the personal/subpersonal and doxastic/subdoxastic distinctions, see Drayson 2012, Drayson 2014.) Proponents of the predictive approach can interpret the architecture as including these doxastic states, but they also have the option of adopting an eliminativist take on beliefs so construed (cf. Dewhurst 2017). There is also a question of whether the influence from top-down processing associated with predictive architectures is appropriately direct. As a result, the predictive approach does not necessarily entail cognitive penetration. In a longer discussion of some of these issues, Fiona Macpherson reaches the similar conclusion that "mere acceptance of the predictive coding approach to perception does not determine whether one should think that cognitive penetration exists" (Macpherson forthcoming, p. 10).

If we leave talk of cognitive penetration out of the picture, however, there is still the question of top-down influence more generally on perceptual processing. The non-isolation claim merely proposes that no part of perceptual processing is isolated from cognitive processing. If we allow that the labels 'perceptual' and 'cognitive' refer respectively to the most spatiotemporally precise predictive processes and the most abstract predictive processes, as discussed with relation to the continuity claim, then predictive architecures look like clear instances of non-isolation. Predictive architectures are committed to the claim that slightly more abstract and less spatiotemporally precise hypotheses act as the priors for slightly less abstract and more spatiotemporally precise hypotheses. Since this is the case at every level in the hierarchy, then doesn't it follow that cognitive processes influence perceptual processes?

There are, I suggest, ways to avoid this conclusion. Notice that this way of reasoning seems to assume that the 'influencing' relation between two levels has the logical property of transitivity: if Level $A + 1$ influences Level $A$, and Level $A$ influences Level $A - 1$, then Level $A + 1$ influences Level $A - 1$. When Hohwy suggests that the top-down processing associated with predictive architectures "is not a free-for-all situation" (Hohwy 2013, p. 155), his argument seems to involve rejecting the transitivity of the 'influencing' relation between two levels. He claims that each pair of levels form a "functional unit", with the higher level passing down predictions and the lower level passing up prediction errors (Hohwy 2013, p. 153), and that each pair of levels is evidentially insulated:

> In this sense the upper level in each pair of levels only 'knows' its own expectations and is told how these expectations are wrong, and is never told directly what the level below 'knows'. […] For this reason, the right kind of horizontal evidential insulation comes naturally with the hierarchy. (Hohwy 2013, p. 153)

Hohwy argues that this evidential insulation prevents high-level processes from influencing low-level processes, except where there is uncertainty and noisy input. He seems to be suggesting that we shouldn't expect the 'influence' relation between any two levels to exhibit transitivity, because the relation should be understood in terms of one level 'providing evidence for' or 'justifying' another. Notice that this relies on a strongly epistemic reading of the Bayesian hierarchy in terms of knowledge and evidence, which Hohwy unpacks in terms of the confidence of a system in its judgments. Ultimately this comes down to the precision of the system's predictions, and its expectations about how good its perceptual inferences are in particular situations. If lower-level predictions are highly precise,

then it is more difficult for them to be influenced by higher-level predictions. (For an argument against Hohwy's approach, see Vance and Stokes forthcoming.)

I propose an alternative (not necessarily incompatible) way to explain why we shouldn't expect the 'influencing' relation between levels to be transitive: I suggest that the relation between any two levels is one of probabilistic causal influence. Bayesian networks are simply maps of probabilistic dependence, and probabilistic dependence is transitive: if Level *A – 1* probabilistically depends on Level *A*, and Level *A* probabilistically depends on Level *A + 1*, then Level *A – 1* probabilistically depends on Level *A + 1* (Korb et al. 2009). But predictive architectures use Bayesian computation: mechanisms that implement *causal* Bayesian networks. In causal Bayesian networks, the probabilistic dependences between variables are the result of causal processes between those variables. Metereological models used to forecast the weather provide a good example of problems that probabilistic causal models raise for the transitivity of the causal relation between two events:

> For instance, given our very coarse and only probabilistic metereological models, each day's weather may be granted to causally influence the next day's weather. But does the weather, say, at the turn of last century still influence today's weather? It does not seem so; somewhere in between the influence has faded completely, even though it may be difficult to tell precisely when or where. (Spohn 2009, p. 59).

Proponents of probabilistic causal models tend to argue that the causal influence between states of a network is weak, and that such weak causal influences will not be preserved over long causal chains (Spohn 2009). As a result, probabilistic causation is widely acknowledged to result in failures of transitivity of the causal relation (Suppes 1970). If we apply this thinking to predictive architectures, we can explain why, when it is true that Level *A + 1* causally influences Level *A*, and Level *A* causally influences Level *A – 1*, we need not expect it to be true that Level *A + 1* causally influences Level *A – 1*. The further apart the levels in the hierarchy are, the less likely there is to be causal influence from the higher level to the lower level. In this way, we can accept that each level in the predictive hierarchy is causally influenced by (i.e. gets its priors from) the level above, without having to accept that each level in the hierarchy causally influences all the levels below it, or that each level is causally influenced by all the levels above it. And so it remains plausible that there are perceptual processes (lower-level processes involved in spatiotemporally precise predictions) which are isolated from cognitive processes (higher-level processes involved in abstract predictions) in the sense that the former are not causally influenced by the latter. The non-isolation claim associated with the predictive approach seems to rely on the assumption that the relation of influence between any two levels in the predictive hierarchy is a transitive one. I have argue that the relation of causal dependence between two events fails to be a transitive relation when the causal dependence in question is probabilistic. And since Bayesian computational mechanisms are probabilistic causal networks, we shold not expect the relation of causal influence between levels to be transitive. As a result, the non-isolation claim is not entailed by predictive mental architectures.

## 5.3 Modularity or Something Like It

I have argued that, despite the appearance of tension between predictive architectures and modular architectures, adopting the predictive approach does not entail rejecting modularity. Claims about the continuity of perception and cognition seem to be compatible with modularity, and the commitment to top-down information processing doesn't entail that no part of perceptual processing is isolated from cognition.

I want to go further, and suggest that predictive architectures themselves possess a kind of modularity. Against the non-isolation claim, I argued that the causal influence of top-down information

processing is unlikely to penetrate from the higher levels of the Bayesian hierarchy to the lower levels, due to the causal intransitivity of probabilistic causal networks such as those found in a Bayesian computational hierarchy (Suppes 1970, Korb et al. 2009). Just as the causal influence of today's weather on future weather diminishes the further into the future we consider, so the causal influence of higher-level processing in the predictive hierarchy diminishes the further down the hierarchy we look: in probabilistic causal models, causal influence is not preserved over long causal chains (Spohn 2009).

As a result, we can acknowledge that a low-level process is influenced by the levels immediately above, but deny that much higher-level processes have any causal influence on it. Such a low-level process would function independently and in isolation from those high-level processes. In other words, it would posess the sort of functional features that we associate with modular architectures. The top-down effects from the level immediately above would presumably count as 'within-module' effects, and this might extend to further levels above. The important point is that not every higher level would necessarily exert a causal influence merely in virtue of being a higher level, because transitivity cannot be expected in causal probabilistic networks.

The kind of modularity involved, however, looks somewhat different from traditional approaches to the modularity of mind. It is not clear, for example, that there is genuine information encapsulation going on here: while it is very unlikely that high-level processes could extend their causal influence all the way down to low-level processes, it is still possible. On traditional approaches to modularity, encapsulated processes are generally portrayed as informationally isolated in principle, rather than merely in practice. A related concern is that, on the view I'm suggested, the boundaries of modules wouldn't be clearly defined – or at least wouldn't retain fixed boundaries over time. It even seems possible for the boundaries of modules to overlap on this view, which is not a feature of traditional modularity.

I propose that these worries should not lead us to think of the architectures in question is non-modular. Some of these features are simply the result of using Bayesian computation: probabilistic mechanisms won't yield the same clean distinctions as classically computational mechanisms. This should prompt us to explore the nature of modularity in probabilistic systems, rather than to reject the useful notion of modules as inapplicable to such architectures. At least some of the modular aspects of predictive architectures, I suggest, can be understood as informationally encapsulated in the Fodorian sense: the processes in question can access less than all of the information available to the organism as a whole (Fodor 1983). But even if it could be argued that there is no information encapsulation in predictive architectures, we can retain the claim that predictive architectures are modular architectures by returning to an alternative characterization of modules introduced earlier. Recall Burnston and Cohen's proposal that a process is modular to the extent that there is a delimited range of parameters to which its processing is sensitive; i.e. insofar as there is a circumscribed number of ways that the process interfaces with other aspects of mental processing. The processes at each level of the Bayesian hierarchy are highly circumscribed in the sense that they are generating hypotheses at different spatiotemporal grains: more precise hypotheses towards the lower levels, and more abstract hypotheses towards the higher levels. Each process is sensitive to a limited range of parameters: generally those of the processes in the levels immediately above and below. And notice that Burnston and Cohen's approach to modularity actually predicts the existence of overlapping modules. They argue that if we individuate modules by the delimited range of parameters to which their processing is sensitive, then it is likely that modular systems will significantly overlap in their associated ranges of parameters (Burnston and Cohen 2015).

My conclusion, that predictive architectures are modular architectures, is similar to Hohwy's conclusion that predictive architecture is "a kind of partially segregated architecture" (Hohwy 2013, 152). (Note that Hohwy's partially segregated architecture posits horizontal evidential insulation in addition to the vertical evidential insulation that is relevant to cognitive penetration. For further details, see Hohwy 2013, pp. 152-155.) Hohwy's argument against widespread cognitive penetration also draws

on the difference in spatiotemporal grain between hypotheses at different level: he claims that higher levels can't influence much lower levels because of the difference in abstractness of their predictions (Hohwy 2013). But it is unclear to me how Hohwy can use this difference to argue against top-down influences on sensory processing without first explaining why we should not expect the relation between levels in the predictive hierarchy to be transitive. His argument for evidential insulation between levels relies on a strong epistemological interpretation of Bayesian processing, as previously discussed, and a courtroom metaphor. My argument can be read as a way of taking Hohwy's argument (from the differences in spatiotemporal grain to the unlikelihood of cognitive penetration) and fleshing it out with the addition of two further claims: first, that causal influence in probabilistic causal networks is intransitive; and second, that the circumscribed nature of the hierarchical processes allows us to individuate them as modules.

## 6 Conclusion

Proponents of predictive architectures often emphasise the way in which their approach constitutes "a genuine departure from many of our previous ways of thinking about perception, cognition, and the human cognitive architecture" (Clark 2013, p. 187). In particular, they tend to highlight the continuity and integration of the Bayesian computational hierarchy, suggesting that it is at odds with the sorts of functionally decomposed or compartmentalized architectures that are associated with modular approaches to the mind.

I have argued that whatever continuity and integration we find in predictive architectures is consistent with modularity. There is no reason to think that the continuity of processing in the predictive hierarchy forces us to deny the distinction between perception and cognition, for example. And while the Bayesian computational processes get their priors from the level above, the limited reach of causal influences in Bayesian mechanisms prevents this from resulting in problematic cognitive penetration.

Furthermore, I have argued that the causal probabilistic networks employed by predictive approaches will actually result in the existence of sensory processes that are functionally isolated from high-level cognition. The causal intransitivity of the probabilistic computational mechanisms ensures that the top-down causal influence diminishes at each level, and suggests that there will be low-level processes which operate entirely independently of the influence from high-level processing. In other words, I have suggested that the predictive approach to the mind, in virtue of its Bayesian computational processes, leads us to expect a modular architecture – albeit perhaps a non-traditional version of modularity.

That predictive architectures are modular architectures should perhaps come as no surprise, if we reflect on the general motivations for modularity. An entirely integrated non-modular computational system, in which every process has access to all the information in the system, would be inefficient, slow, and potentially intractable. Modularity provides a way to make computational processes more efficient and fast by restricting the information to which certain processes have access. Predictive architectures operate in exactly the same way: each level of the hierarchy is restricted in the information it accesses: it uses information from the level above to predict its inputs from the level below. One way to interpret this fact is to claim that predictive architectures are rivals to modular architectures, using different processes to achieve similar results. I would suggest, to the contrary, that predictive and non-predictive architectures achieve similar results by organizing their information in a modular way. This is not to deny that there are differences between kinds of computational architectures. But modularity is a higher-level feature that can be shared by distinct computational architectures, and which allows us to categorize systems by the way they organize their information.

The focus on Bayesian computational hierarchies and predictive coding makes for a new and interesting approach to mental architecture, which is ripe for exploration and development. And while there are aspects of the continuity and integration of these architectures that deserve to be empha-

sized, we must not lose sight of the fact that they ultimately have to explain a wide array of very different skills and capacities. The appeal of the predictive approach is not that it shows our minds to be continuous and integrated, but that it shows how a set of continuous and integrated computational processes can be organized in such a way as to give rise to distinct mental capacities. It does so, I suggest, by organizing its information processes in a modular way. Moreover, it provides us with a new way to understand modularity as a flexible and dynamic feature of architectures, and to appreciate that predictive architectures are modular architecture.

## References


Aydede, M. (2015). The language of thought hypothesis. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy* Metaphysics Research Lab, Stanford University.

Burnston, D. C. & Cohen, J. (2015). Perceptual integration, modularity, and cognitive penetration. In A. Raftopoulos & J. Zeimbekis (Eds.) *Cognitive influences on perception: Implications for philosophy of mind, epistemology, and philosophy of action.* Oxford University Press.

Carruthers, P. (2006). *The architecture of the mind: Massive modularity and the flexibility of thought.* Oxford: Clarendon Press.

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

Davies, M. (1989). Connectionism, modularity and tacit knowledge. *British Journal for the Philosophy of Science, 40* (December), 541–55. Taylor & Francis.

Dewhurst, J. (2017). Folk psychology and the Bayesian brain. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. *Philosophical Perspectives, 26* (1), 1–18.

——(2014). The personal/subpersonal distinction. *Philosophy Compass, 9* (5), 338–346.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Friston, K. J. & Stephan, K. E. (2007). Free-energy and the brain. *Synthese, 159* (3), 417–458.

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

Korb, K. B., Hope, L. R. & Nyberg, E. P. (2009). In F. Emmert-Streib & M. Dehmer (Eds.) *Information-theoretic causal power* (pp. 231–265). Boston, MA: Springer US.

Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology, 6* (4), 547–569.

Macpherson, F. (forthcoming). The relationship between cognitive penetration and predictive coding. *Consciousness and Cognition*.

Spohn, W. (2009). The difficulties with indirect causation. *Causation, Coherence and Concepts: A Collection of Essays* (pp. 57–65). Springer.

Suppes, P. (1970). *A probabilistic theory of causality.* Amsterdam, North-Holland Pub. Co.

Vance, J. & Stokes, D. (forthcoming). Noise, uncertainty, and interest: Predictive coding and cognitive penetration. *Consciousness and Cognition*.

Vetter, P. & Newen, A. (2014). Varieties of cognitive penetration in visual perception. *Consciousness and Cognition, 27*, 62–75.

# Predictive Processing and Cognitive Development

Regina E. Fabry

The ability to acquire new cognitive capacities is the hallmark of the human mind. Emerging predictive processing (PP) accounts promise to offer a unified, mechanistic perspective on the neuro-functional realization of this complex process. The acquisition of any given cognitive capacity can be depicted as a specific stage of the continuous prediction error minimization process. This stage is characterized by distinct functional profiles, unique patterns of neuronal activation in crucial brain areas, and the refinement and flexible adaptation of motor programs to new processing challenges. In this paper, I will suggest that PP is a good conceptual partner for accounts of enculturation.

Enculturation is the idea that socio-culturally shaped cognitive processes emerge from an individual's scaffolded, embodied interaction with its cognitive niche. This structured engagement with specific, socio-culturally developed patterns in the niche transforms the overall cognitive capacities of an individual (Menary 2013; Menary 2015a). The process of enculturation is associated with significant changes in the functional and structural properties of the brain. Furthermore, it alters and refines the functional profiles of bodily actions and motor programs. I will argue that PP offers important conceptual tools and theoretical considerations that complement these basic principles of enculturation in a new and original way. The resulting account of the enculturated predictive acquisition of cognitive capacities (EPACC) has the conceptual resources to consider specific cases of cognitive development at multiple levels of explanation.

This new perspective on EPACC is a timely contribution to the debate about the philosophical implications of PP. On the one hand, Jakob Hohwy (Hohwy 2013) argues that PP implies an internalistic and neurocentric view of cognition. On his account, the embodiedness of cognitive systems and their flexible interaction with the local environment do not seem to play an important role in accounts of cognitive processes. On the other hand, Andy Clark (Clark 2016) defends the idea that PP is compatible with philosophical positions that emphasize the embodied, embedded, extended, or enacted (4E) dimensions of cognition. In this paper, I will discuss both interpretations of PP and assess their relationship to EPACC. I will challenge the internalistic account of PP on conceptual grounds. At the same time, I will argue that EPACC offers a concrete proposal for the complementarity of PP and enculturation. The overall argument is that the acquisition of new, socio-culturally shaped cognitive capacities is a matter of enculturation. The underlying neuronal and bodily transformation routines can be described in terms of PP. However, if we aim for a full-fledged account of the complexity and fragility of cognitive development, we also need an account of the fine-grained interactions of novices with their structured environment.

**Keywords**

Cognitive niche construction | Cognitive norms | Embodied cognition | Enculturation Neural plasticity | Neural reuse | Predictive processing | Scaffolded learning

## 1 Introduction

The temporally extended acquisition of cognitive capacities determines our repertoire of meaningful interactions with the world.[1] During ontogeny, we become thinkers and reasoners, learning to solve

1 Part of this work was supported by the Barbara Wengeler Foundation. I would like to thank two anonymous reviewers, Thomas Metzinger, and Wanja Wiese for their feedback on earlier versions of this paper.

all kinds of problems in systematic and efficient ways. Recent work on predictive processing (PP) promises to offer an intriguing account of the neuro-functional underpinnings of cognitive development (Clark 2016; Hohwy 2013). The overall purpose of this paper is to investigate the contribution of PP to a better understanding of ontogenetic cognitive development and the acquisition of cognitive capacities.

According to PP accounts, prediction error minimization is the core principle that drives the realization of developmental changes. The revision of predictions in the light of ensuing prediction error ideally leads to a step-wise modification and refinement of future predictions at multiple levels of the hierarchically organized generative model realized in the human brain. Perceptual inference and active inference are computationally similar and highly interactive means of minimizing prediction error. In both cases, the optimization of precision estimations determines the causal influence or 'weight' of bottom-up prediction errors on the prediction error minimizing process. Precision estimations also orchestrate the relationship between perceptual inference and active inference. The minimization of prediction error and the co-occurring optimization of precision estimations are pervasive features of cognitive functioning in general. The background assumption of this paper is that there are stages of increased prediction error minimization which are associated with the acquisition of particular cognitive capacities.

For ease of exposition, I will restrict my considerations to a specific type of cognitive capacities acquired, in the majority of cases, early in life: the cognitive practices that enable the completion of complex cognitive tasks (Menary 2007; Menary 2015a; Menary 2016). Cognitive practices are embodied, socio-culturally shaped interactions with epistemic resources in the cognitive niche. The acquisition of cognitive practices is a matter of enculturation. Enculturation is a developmental process that is characterized by plastic changes to neuronal circuitry and motor profiles. It is strongly influenced by the structured interaction of an individual with its cognitive niche.

In the next section, I will briefly summarize the most important features of PP. I will then, in Section 3, consider the components necessary for the ontogenetic acquisition of cognitive capacities in terms of prediction error minimization. In Section 4, I will argue that PP, taken on its own, does not have the conceptual resources and the theoretical scope to cover the entire spectrum of components that plays a decisive role in the acquisition of cognitive practices (and most likely other kinds of cognitive capacities). I will suggest that, if we are interested in a fully-fledged description of socio-culturally structured cognitive development 'in the wild', we need to give a multi-level account of the embodied interaction of a cognitive system with its cognitive niche. In Section 5, I will relate my considerations to the emerging debate on the theoretical import of PP. In particular, I will discuss how the view to be developed here relates to Jakob Hohwy's (Hohwy 2013) and Andy Clark's (Clark 2016) interpretations of PP.

## 2 A Sketch of Predictive Processing

According to recently developed accounts of PP in cognitive neuroscience (Friston 2005; Friston 2010; Seth 2015) and philosophy of cognitive science (Clark 2013a; Clark 2015; Clark 2016; Hohwy 2012; Hohwy 2013; Hohwy 2015a), our engagements with the world can be best understood as a continuous attempt to minimize prediction error. On this view, perception, action, and cognition are associated with a large-scale hierarchical generative model realized in the brain. The generative model is comprised of multiple levels, each of which interacts with the next sub-ordinate and supra-ordinate level. Top-down predictions are probabilistically assessed estimations of bottom-up signals. The discrepancy between bottom-up signals and top-down predictions is the prediction error. The prediction error serves as a bottom-up signal for the next supra-ordinate level. This basic principle is iterated at multiple levels throughout the hierarchical generative model. The function of this multi-level processing mechanism is to minimize prediction error. *Prediction error minimization* can be understood

as a functional attempt to reduce or explain away the difference between top-down predictions and bottom-up signals at multiple levels (Clark 2013a; Clark 2016; Hohwy 2013).

Prediction error minimization is a specific application of the *free energy principle (*Friston 2005; Hohwy 2015a; Seth 2015). According to this principle, "any self-organizing system that is at equilibrium with its environment must minimize its free energy" (Friston 2010, p. 127). Under some simplifying assumptions, we can apply this principle to cases of human perception, action, and cognition. This leads to the idea that "[f]ree energy is the sum of prediction error, which bounds the surprise of the sensory input to the system" (Hohwy 2015a, p. 2). In the present context, *surprise* is not meant to be a phenomenon that applies to the phenomenal experience of organisms as a whole at a personal level of explanation. Rather, it needs to be understood as an information-theoretic quantity — also known as *surprisal* — that is assessed at a sub-personal level of explanation *(*Hohwy 2015b). In this sense, surprisal reports "the sub-personally computed implausibility of some sensory state given a model of the world" (Clark 2013a, p. 186). It would not be computationally tractable for any system to assess surprisal directly. The solution to this tractability problem is to implicitly minimize surprisal and its upper bound — free energy — by minimizing prediction error (Hohwy 2013; Hohwy 2015a; Seth 2015). Prediction error minimization is thus a special case of a more general strategy for sustaining an organism's physiological integrity.

There are two distinct, yet highly interactive ways to minimize prediction error. In *perceptual inference*, top-down predictions are modified as a result of ensuing prediction error. In *active inference*, embodied actions bring about changes in the available sensory input so as to confirm the accuracy and adequacy of top-down predictions. On this construal, any type of bodily movement — from oculo-motor adjustments to locomotion — has the potential to confirm the best probabilistically generated predictions at multiple temporal and spatial resolutions. Perceptual inference and active inference are computationally similar, but they have a "different direction of fit" (Hohwy 2013, p. 178). They are complementary ways of minimizing prediction error. The on-going causal interaction of perceptual inference and active inference leads to the idea that they "unfold continuously and simultaneously, underlining a deep continuity between perception and action" (Seth 2015, p. 5).

Perceptual inference and active inference are orchestrated by the *optimization of precision estimations.* The degree of estimated precision determines the causal contribution of error signals to the prediction error minimization process in a given context (Clark 2014; Clark 2016). The optimization of precision estimation thus has the function of fine-tuning the influence of bottom-up prediction error signals and top-down predictions on each other in both perceptual and active inference. It has been proposed that the optimization of precision estimations is associated with the direction of *attention (*Clark 2016; Feldman and Friston 2010; Friston and Stephan 2007; Hohwy 2013). Precision estimation is considered to be equivalent to the regulation of the post-synaptic gain of specific prediction error units (Friston 2010; Feldman and Friston 2010; Hohwy 2013). The idea is that certain neurotransmitter systems (e.g., dopamine; Fletcher and Frith 2009) and neuro-modulatory hormones (e.g., oxytocin; Quattrocki and Friston 2014) are associated with post-synaptic gain control and thus with the optimization of precision expectations.

In sum, PP can be understood as a core mechanistic principle that is supposed to guide various cases of human perception, action, and cognition. A major consequence of this theoretical proposal is that we should alleviate any traditional clear-cut distinctions between these traditional mental kinds (Clark 2013a). This is because perceptual, active, and cognitive processes are all assumed to be realized by prediction error minimization.[2] On this construal, the differences between more perceptually basic and

2 If correct, PP has the potential to overcome traditional, unduly restrictive conceptions of the mind, which consists in the (more or less pronounced) dissociation of perceptual, active, and cognitive processes. This limitation was identified by John Dewey (Dewey 1896, p. 358) 120 years ago: "Instead of interpreting the character of sensation, idea and action from their place and function in the sensori-motor circuit, we still incline to interpret the latter from our preconceived and preformulated ideas of rigid distinctions between sensations, thoughts and acts."

more cognitively sophisticated processes, traditionally construed, would manifest themselves in the particular temporal and spatial resolution of prediction error minimization (Clark 2016; Hohwy 2013).

## 3 The Predictive Acquisition of Cognitive Capacities

Cognitive development falls naturally out of the basic principles that govern PP. The on-going attempt to minimize prediction error and to optimize precision estimations is nothing but a computationally tractable way of updating model parameters. Ideally, this leads to more accurate and precise sets of predictions. In this sense, prediction error minimization accounts for the developmental trajectory throughout the lifespan of perceiving, acting, and cognizing systems. This developmental trajectory is characterized by certain stages that show significantly increased levels of prediction errors and prediction revision. At these stages, novices acquire certain domain-specific capacities that allow them to interact with their local environment and to complete cognitive tasks in efficient and effective ways. Call this process of extensive and significant model updating the ontogenetic predictive acquisition of cognitive capacities (PACC). PACC temporally precedes the 'baseline' minimization of prediction error, which I dub the predictive improvement of cognitive capacities (PICC), in specific domains across the lifespan. The distinction between PACC and PICC is not clear-cut, but characterizes a continuous process that serves the general purpose of maintaining the bio-functional integrity of predictive systems. I will be primarily concerned with cases of PACC in this paper, with a particular focus on the acquisition of cognitive practices. However, it is important to bear in mind that acquired cognitive capacities, including cognitive practices, are refined and revised at multiple scales, and operate in specific contexts and situationally determined conditions across the lifespan (Clark 2016).

PACC offers a specific expression of developmental neural plasticity (Ansari 2012; Ansari 2015; Dehaene 2010; Menary 2014; Stotz 2010; Van Atteveldt and Ansari 2014) or learning driven plasticity (LDP; Menary 2015a). LDP refers to the idea that the acquisition of a certain cognitive capacity is associated with changes to the structural, functional, and effective connectivity of cortical areas.[3] LDP is not an open-ended process of resource allocation in the brain. Rather, it is constrained by anatomical properties and functional biases that define the possibility space of plastic changes in the brain (Anderson and Finlay 2014; Anderson 2015; Anderson 2016). A functional bias is defined as "a set of dispositional tendencies that capture the set of inputs to which the circuit will respond and govern the form of the resulting output" (Anderson 2015, p. 15). According to Michael Anderson's (Anderson 2015) interactive differentiation and search (IDS) framework, the functional and structural potentials of specific brain areas and local neural circuitry are exploited by a 'neural search' mechanism. The idea is that "[a] set of […] neural structures with different functional biases (different input-output mappings) would be enough to allow an ongoing process of neural search to identify and consolidate the sets of partnerships that reliably supported skills being acquired during development" (Anderson and Finlay 2014, p. 12; see also Anderson 2015). From the perspective of IDS, LDP would be about the establishment of neuronal coalitions as a result of 'neural search', which is constrained by the structural and functional properties and the connectivity potentials of cerebral units. Below, I will look at the neuronal changes associated with reading acquisition as being a paradigm example of the relation between IDS and LDP.

The IDS framework is of particular importance for our understanding of the acquisition of cognitive practices. Cognitive practices are a specific, socio-culturally shaped type of cognitive capacities. The acquisition of cognitive practices, such as reading, writing, and arithmetic augment and support the completion of certain cognitive tasks (Menary 2012; Menary 2015a). Cognitive practices are evolutionarily recent, which excludes the possibility that dedicated, exclusive, and modular neural circuitry may have emerged on a phylogenetic timescale (Anderson 2010; Anderson 2015; Dehaene 2010; Heyes 2012; Van Atteveldt and Ansari 2014). The question is how the human brain has adapted to the

3 For a thorough discussion of effective connectivity and its relation to PP, see Clark (Clark 2016, pp. 146-150; Clark 2013b).

processing needs afforded by recent socio-culturally shaped ways of cognizing. This includes the interaction with epistemic resources such as writing systems or number systems. A reasonable answer to this question is that global brain organization is governed by the principle of *neural reuse (*Anderson et al. 2012; Anderson 2010; Anderson 2015; Anderson 2016).

Neural reuse is defined as "the use of local regions of the brain for multiple tasks across multiple domains" (Anderson 2015, p. 4). On this view, "[…] individual pieces of the brain, from cells to regions to networks, are used and reused in a variety of circumstances, as determined by social, environmental, neurochemical, and genetic contexts" (Anderson 2015, p. 36). Dehaene's (Dehaene 2005; Dehaene 2010) *neuronal recycling* hypothesis is a specification of this principle, which is supposed to account for the cerebral realization of ontogenetically acquired cognitive practices. He defines neuronal recycling as a neuro-functionally realized "[…] form of reorientation or retraining: it transforms an ancient function, one that evolved for a specific domain in our evolutionary past, into a novel function that is more useful in the present cultural context" (Dehaene 2010, p. 146).

The combination of neuronal recycling, neural search, and the consideration of functional biases offers an account of the possibilities and constraints that determine the developmental potential of LDP for the acquisition of a certain cognitive practice. A reasonable conjecture at this point is that LDP and its guiding principles are evolutionary adaptations that allow for sufficient, yet not open-ended, flexibility in response to and in close interaction with the cognitive niche (Clark 2008; Dehaene 2010; Menary 2013; Sterelny 2003).

The PP scheme has the conceptual resources to integrate these considerations regarding LDP and the importance of neural reuse and neural search. It "[…] combines functional differentiation with multiple (pervasive and flexible) forms of informational integration" (Clark 2016, p. 150). On this construal, the principles governing LDP are realized by the ongoing minimization of prediction error in the course of PACC. Prediction error minimization is thus a means of exploiting and connecting cerebral units that are apt to contribute to the realization of a certain neuronal function.

To give an example, there is converging evidence suggesting that the left ventral occipito-temporal (vOT) area is reliably and consistently recruited in the course of reading acquisition (Dehaene 2005; Dehaene 2010; McCandliss et al. 2003; Price and Devlin 2003; Price and Devlin 2004; Vogel et al. 2013). This area has a functional bias that makes it most suitable to be reused or recycled so as to contribute to a large-scale neural circuit that is associated with visual word recognition (Dehaene 2010; Vogel et al. 2014). According to Price and Devlin's (Price and Devlin 2011, p. 248) PP account of visual word recognition, "[…] learning involves experience-dependent synaptic plasticity, which changes connection strengths and the efficiency of perceptual inference." Price and Devlin (Price and Devlin 2011) propose a three-stage model of reading acquisition in terms of prediction error minimization, with a focus on the contribution of the left vOT area to PACC. This model is the first of its kind that is able to provide an explanation for the observation that the left vOT area appears to be a hub for the development of new feed-forward *and* feed-back connections. Additionally, this model is unprecedented because it offers an account of the experimental data on the activation levels in the left vOT area across the whole trajectory of reading acquisition. The three-stage model depicts the developmental trajectory as an "inverted U-shape of activation levels" in the left vOT area and the relevant cortical areas that become increasingly functionally connected to it (Price and Devlin 2011).

At the first stage prior to the proper acquisition of reading, predictions and prediction errors associated with this cognitive practice are hardly detectable. This is because the predictive system is not yet able to identify specific task-relevant sensory signals as salient and potentially precise. At the second stage, which Price and Devlin (Price and Devlin 2011) call "early learning", prediction errors realized in the left vOT area are significantly high. This is because the top-down predictions that influence this area are inaccurate and thus are not sufficiently adjusted to the wealth of salient sensory information. The prediction error signals have a potent causal influence on supra-ordinate hierarchical levels. This normally leads to the refinement and modification of future predictions. It is at this stage that the

'neural search' mechanism is successful in the exploitation and recruitment of a cortical area that has the right kind of functional profile so as to contribute to a new, neuronal function. At the third stage, which is associated with a sufficient degree of proficiency, the relevant predictions influencing the left vOT area become increasingly accurate. Thus, they have a more pronounced causal influence on the entire processing hierarchy. Accordingly, the overall causal influence of prediction error associated with significant neuronal activation patterns in the left vOT area decreases over time.

This model of PACC is supported by converging empirical evidence on the neuronal changes that occur in the course of reading acquisition. Several neuroimaging studies suggest that the activation level of the left vOT area peaks in beginning readers and significantly decreases over time with the development of reading proficiency (Ben-Shachar et al. 2011; Brem et al. 2010; Maurer et al. 2006). In addition, it has been shown that there is a significant increase in functional connectivity between the left vOT area and higher-level left-hemispheric frontal and temporal areas that are reliably associated with visual word recognition and language processing more generally (Dehaene et al. 2010; Gaillard et al. 2003; Turkeltaub et al. 2003).

This example offers a proposal for the application of PP to concrete cases of LDP and the associated search for a cortical area that has the functional bias to contribute to a particular neuronal function. Furthermore, this example also illustrates how neuronal recycling might be realized, under the condition that prediction error minimization is a pervasive mechanistic principle that guides and refines the acquisition of new neuro-functional profiles. In this sense, PACC provides a proposal about the temporal unfolding of neural search and neuronal recycling. This is in line with Clark's (Clark 2016, p. 150) view that prediction error minimization captures the emergence of cognitive capacities in a way that is fully consistent with neural reuse and neural search: "Distinctive, objectively-identifiable, local processing organizations now emerge and operate within a larger, more integrative, framework in which those functionally differentiated populations and sub-populations are engaged and nuanced in different ways so as to serve different tasks". Importantly, this process of constrained neuronal resource allocation is modulated by the optimization of precision estimations. In particular, precision estimations determine the causal influence (or 'weight') of prediction error signals on adjacent levels of the hierarchical generative model. Clark (Clark 2016, p. 277) describes this process in the following way:

> Highly-weighted errors, if the system is unable to explain them away by recruiting some model that it already commands, result in increased plasticity and (if all goes well) the acquisition of new knowledge about the shape and nature of the distal causes responsible for the surprising inputs.

Considering our example, prediction errors realized in the left vOT area are estimated as being precise. Therefore, they have a pronounced influence on the revision and modification of supra-ordinate levels of the cortical hierarchy. This leads to the establishment of new structural, functional, and effective connections between the left vOT area and higher-level cortical areas that are equipped to generate accurate predictions. These predictions are increasingly able to meet prediction error signals driven by linguaform visual input. Given that the optimization of precision estimations is associated with attention, attention is a powerful means of altering the causal flow of predictions and prediction error signals in the course of PACC. On this construal, "[a]ttention […] is simply one means by which certain error-unit responses are given increased weight, hence becoming more apt to drive learning and plasticity, and to engage compensatory action" (Clark 2013a, p. 190). The last point is important, because it emphasizes that both perceptual inference and active inference guide PACC.

PACC is realized by the close interaction of perceptual inference and active inference, which is constrained by the assignment of precisions to certain prediction error signals. This gives rise to the idea that LDP is complemented by a genuinely embodied component of PACC, which I call learning dependent bodily adaptability (LDBA). LDBA guides the developmental trajectory of skilled motor action in close interaction with plastic changes in the brain. The ensuing development of new motor

patterns and action routines is constrained by the overall morphology of human bodies and their constitutive parts, in addition to the functional biases of cerebral regions that realize the initiation of embodied active inference.

In human vision, for example, the acuity limitations of the retina beyond the fovea constrain the emerging oculo-motor patterns displayed by proficient readers (Rayner et al. 2007; Rayner 1998; Rayner 2009). In the case of reading acquisition, eye tracking studies have revealed that these constraints lead to a characteristic profile of the alternation between saccades and fixations in the course of skill acquisition (Huestegge et al. 2009; Joseph and Liversedge 2013; Rayner et al. 2001; Seassau et al. 2013). This profile is best characterized by an efficiency gain in terms of a significant decrease of saccades, refixations, and fixation durations and a significant increase of saccade amplitudes. This efficiency gain represents the developmental trajectory of active inference in the course of PACC. Another example for the importance of bodily and motor constraints on LDBA is the anatomical and physiological organization of human hands and arms. This determines the degrees of freedom of joints and muscles, which constrain the development of movement patterns associated with writing. The physiological and mechanical properties of the involved effectors thus determine the potential characteristics of motor patterns (Alsmith 2012; Dounskaia et al. 2000; Phillips et al. 2009).

In these cases, the specific ways of embodied interaction with epistemic resources in terms of active inference are both enabled and constrained by the functional biases displayed by the anatomical and mechanic configuration of the human body. This potential for embodied active inference complements the biases of cerebral areas to contribute to the acquisition of new cognitive functions. This complementarity is orchestrated by the optimization of precision estimations. Both LDP and LDBA are thus integral components of PACC.

According to Clark's (Clark 2016) and Hohwy's (Hohwy 2013) versions of the PP framework, PACC can be seen as a specific variant of supervised learning. Hohwy (Hohwy 2013, p. 49) argues that "[t]he predictions of the internal models that determine perception are supervised by the world itself." On Clark's (Clark 2016, p. 18) view of the relationship between prediction error minimization and supervised learning, PP style accounts of learning and cognitive development (both PACC and PICC) "offer a form of self-supervised learning, in which the 'correct' response is repeatedly provided, in a kind of ongoing rolling fashion, by the environment itself." This suggests that it is the local environment that supervises the continuous process of prediction error minimization. In this way, the features and particular organization of the local environment shape and constrain the developmental trajectory of prediction error minimization and the optimization of precision.

Taken on its own, PP is primarily an account of the cerebral processes that are associated with specific instances of perception, action, cognition, and attention (traditionally conceived). Embodied active inference in the service of prediction error minimization may (Clark 2016) or may not (Hohwy 2013) serve as a reason to extend the units of analysis to include embodied interaction with the local environment. If it is the environment (or the world) which supervises learning and development, we are well advised to include the 'supervisor' and its impact on PACC into our considerations. This is where enculturation enters the picture.

## 4 The Enculturated Predictive Acquisition of Cognitive Capacities

Enculturation is the temporally extended transformative acquisition of embodied cognitive practices in the cognitive niche. According to proponents of enculturation, our cognitive capacities are "augmented and transformed by the acquisition of cognitive practices" (Menary 2012, p. 148). The idea of enculturated cognition commits us to the view that our understanding of the genuinely neuronal and bodily underpinnings of cognitive practices needs to be complemented by considerations of the embodied interaction of cognitive systems with their cognitive niche. The cognitive niche can be defined as the incrementally, trans-generationally structured socio-cultural environment that provides human

organisms with epistemic resources for the completion of cognitive tasks. Examples of resources in the cognitive niche include tools, artefacts, and representational systems. In addition, the cognitive niche is also characterized by socio-cultural institutions like kindergartens, schools, and universities. Call this the *niche aspect* of enculturation. In addition, cognitive practices are a specific kind of "patterned practice" (Roepstorff et al. 2010), because they are shared by a large number of individuals in the socio-culturally structured cognitive niche. Therefore, the skillful performance of cognitive practices is constrained by sets of *cognitive norms* (Menary 2007; Menary 2010; Menary 2013; Menary 2015a). These norms regulate interactions with epistemic resources. They need to be learned and automatized in the course of acquisition. This is the *normative aspect* of enculturation. Since cognitive practices are socio-cultural phenomena, their acquisition is in itself a socio-culturally structured process. This process is characterized by *scaffolded learning*. In its most general form, scaffolding "denotes a broad class of physical, cognitive, and social augmentations — augmentations that allow us to achieve some goal that would otherwise be beyond us" (Clark 1997, pp. 194-195). In our context, scaffolding refers to the idea that the acquisition of a cognitive practice is a systematic process of novice-expert interaction in the cognitive niche (Menary 2013; Menary 2015a; Fabry 2015). This interaction is structured by the current developmental stage of the novice and a specific set of skills and knowledge that needs to be acquired in the long run (Estany and Martínez 2014). Call this the *scaffolding aspect* of enculturation. I will consider all three aspects of enculturation in turn.

Cognitive practices are the result of enculturation. PACC is supposed to account for the acquisition of cognitive practices. Hence, PACC should take the niche aspect, the normative aspect, and the scaffolding aspect of enculturation into account. This requires the investigation of the enculturated predictive acquisition of cognitive capacities (EPACC).

Recall from the previous section that Clark (Clark 2016) and Hohwy (Hohwy 2013) classify prediction error minimization as a specific case of supervised learning. The cognitive niche fulfills the role of a 'supervisor', enabling and facilitating the acquisition of cognitive practices. In order to be able to specify the conditions of the predictive acquisition of a certain cognitive practice, we need to take the diachronic and synchronic features and properties of the cognitive niche into account. Work on *cognitive niche construction* suggests that the trans-generational modification of the local environment has facilitated and amplified the development of increasingly fine-grained and sophisticated problem solving routines (Clark 2006; Clark 2008; Kendal 2011; Laland and O'Brien 2011; Odling-Smee and Laland 2011; Sterelny 2003; Sterelny 2012; Stotz 2010; Stotz 2014). Cognitive niche construction is further characterized by *epistemic engineering* (Menary 2014; Menary 2015a), i.e., by the principle of "organizing our physical environment in ways that enhance our information-processing capacities" (Sterelny 2012, p. xii). An important consequence of cognitive niche construction is that the phylogenetic and ontogenetic development of human cognitive capacities cannot be understood independently from considerations about the structured environment in which they are situated; we need to take the interaction and the mutual dependence of human cognitive systems and their cognitive niche into account.

In various ways, the cognitive niche assembles an abundance of *epistemic resources*, which have been shaped and reshaped in the course of cultural evolution. These epistemic resources are the product of epistemic engineering. They include but are not limited to writing systems, number systems, artefacts, and tools. Embodied interaction with these materials enables innovative and efficient solutions to cognitive challenges. These solutions are not only constrained by the overall possibilities and limitations of neuronal and bodily processing routines, but also by the properties of epistemic resources and by the principles that govern the interaction with these resources (Menary 2015a). For example, the alphabetic principle, i.e., the correspondence of specific graphemes and phonemes, constrains the ways in which cognizers can employ tokens of an alphabetic writing system in the completion of a certain cognitive task (Frith 1985; Rayner et al. 2001; Snowling 2000; Ziegler and Goswami 2006). In this sense, the exploitation of epistemic resources is governed by cognitive norms.

Cognitive norms specify the relationship between epistemic resources and the cognitive systems interacting with them. As Menary (Menary 2010, p. 229) puts it, cognitive systems are "embedded in a physical and social environment, and that environment contains norms which determine the content of environmental vehicles and how we manipulate them." Cognitive norms constrain the manipulation and interpretation of epistemic resources (Menary 2007). For example, the cognitive norm that alphabetic writing systems are spatially arranged from left to right and from top to bottom permits specific ways of interacting with a printed text and prohibits others.

Cognitive norms are acquired through scaffolded learning (Menary 2010). In the course of the acquisition of a cognitive practice, cognitive systems learn to manipulate environmental resources automatically and fluently in accordance with these cognitive norms (Menary 2013). The specific components of cognitive norms are adapted to the learner's current stage of competence. Accordingly, the "properties of cognitive norms are altered from being entirely explicit and context free to being entirely implicit and embodied" (Menary and Kirchhoff 2013, p. 9). This account of scaffolded learning is consistent with Vygotsky's (Vygotsky 1978) approach to ontogenetic cognitive development. On his view, the acquisition of a certain cognitive capacity is structured by the learner's *zone of proximal development* (Clark 1997; Menary 2007). It is defined as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky 1978, p. 86; italics removed). Someone's current state of scaffolded learning is thus determined by the developmental stage of the novice and the scope and limits of her cognitive potential as defined by the zone of proximal development. This idea can also be found in the consideration of scaffolded learning by Wood, Bruner, and Ross (Wood et al. 1976). In particular, they argue that "scaffolding consists essentially of the adult 'controlling' those elements of the task that are initially beyond the learner's capacity, thus permitting him to concentrate upon and complete only those elements that are within his range of competence" (Wood et al. 1976, p. 90).

An example of the temporal unfolding of scaffolded learning is Uta Frith's (Frith 1985) influential three-stage model of reading acquisition. According to this model, the process of becoming a proficient reader is comprised of the progressive acquisition of logographic, alphabetic, and orthographic skills. At the first stage, specific words are unsystematically identified as a whole on the basis of their visual properties. At the second stage, novices are systematically trained in the application of the alphabetic principle. The emerging ability to apply grapheme-phoneme correspondence rules "is an analytic skill involving a systematic approach, namely decoding grapheme by grapheme" (Frith 1985, p. 306). Finally, orthographic skills are developed on the basis of the abilities that have been acquired at previous stages. Once alphabetic skills have reached a certain stage of automaticity and fluency, words can be recognized without explicitly associating them with their relata in spoken language.

This example emphasizes the idea that the temporal structuring of the interaction with epistemic resources is an important condition for successful enculturation. It also suggests that scaffolded learning is an inherently social process that systematically shapes the relationship between a novice and the epistemic resources located in her cognitive niche. In this sense, *path-dependent learning* is a guiding principle of scaffolded learning. Path-dependent learning induces a specific kind of "cognitive path-dependence", according to which "you can't get everywhere from anywhere, and where you are now strongly constrains your potential future intellectual trajectories" (Clark 1997, p. 205). Path-dependent learning is a process that selects and shapes the particular features of the scaffolding procedure. Clark's (Clark 1997) considerations on path-dependence develop a new twist when applied to the guiding principles of PACC:

> The hierarchical nature of the prediction-based approaches […] makes them especially well-suited as inner mechanisms capable of supporting complex patterns of path-dependent learning in which

> later achievements build on earlier ones. At the same time, however, prior learning makes certain other regularities harder (at times impossible) to spot. (Clark 2016, p. 288)

However, the broad perspective on the acquisition of cognitive capacities I have offered so far suggests that prediction error minimization is more than just an "inner mechanism" that is apt to learn in a specific, path-dependent manner. Rather, it is in virtue of the embodied interaction of enculturated predictive systems with their cognitive niche that path-dependent learning is a potent means of EPACC. The temporally ordered and systematic exposure of scaffolded novices to structured epistemic resources, orchestrated by the optimization of precision estimations, facilitates and constrains efficient perceptual inferences and active inferences. Scaffolded learning, enacting the principle of path-dependence, is thus a good candidate for an in-depth description of the particular features of PP style supervised learning.

Taken together, all three aspects of enculturation are indispensable components of the acquisition of cognitive practices (and most likely of other types of cognitive capacities). This suggests that PACC, taken on its own, does not have the explanatory power to account for the full range of factors that contribute to the realization of cognitive practices in an important and indispensable way. The interpretation of PP advocated for here offers a plausible and epistemically rewarding "mechanism sketch" of the perceptual, active, and cognitive processes that give rise to the successful performance of cognitive practices (Piccinini and Craver 2011). This sketch is intended to be enriched by new empirical evidence for the particulars of the structural and functional realization of PP along the way (Harkness 2015). However, it does not and need not account for all components that have an indispensable impact on the acquisition of cognitive practices. We need EPACC in order to consider and describe the entire range of neuronal, bodily, and environmental components that shape and re-shape the acquisition of cognitive capacities. In this sense, the enculturation perspective and PP are complementary.

This complementarity is characterized by the "division of labour" between PP and the account of enculturation (Menary 2015b, p. 7). This becomes obvious as soon as we distinguish at least three temporal units of analysis and their association with specific levels of explanation. This is important given that enculturation is a process that unfolds in and across time (Menary 2015b). On a physiological timescale, we can attempt to provide an account of the neuronal and bodily changes that are associated with EPACC at a sub-personal level of explanation. It is here that prediction error minimization avails itself as a mechanistic proposal for the realization of LDP and LDBA.[4] On an organismic timescale, we can investigate the specific changes that characterize EPACC on a personal level of explanation. From this perspective, we can investigate the embodied interaction of human organisms with epistemic resources, the set of cognitive norms that guides this interaction, and the scaffolding procedures that facilitate the acquisition of cognitive capacities. Finally, on an evolutionary timescale, we may approach the co-ordination of the cognitive niche and the phylogenetic trajectory of human organisms at a supra-personal level of explanation.

These distinctions emphasize that the ontogenetic acquisition (and improvement) of cognitive capacities is a complex and multi-level phenomenon. But they also highlight why we should prefer EPACC over PACC: EPACC has the conceptual resources to account for the entire range of components that give rise to the acquisition of cognitive capacities. In contrast, PACC restricts itself to the specification of the neuronal and bodily underpinnings of the acquisition of a certain cognitive capacity on a physiological timescale and at a sub-personal level of explanation. From this perspective, EPACC offers a proposal for how prediction error minimization "fits within a larger account of the brain-body-niche nexus" (Menary 2015b, p. 7).

4 This account is markedly different from recent attempts to explain the phenomenal experience of time and temporal order in terms of PP (Friston 2016; Hohwy et al. 2016; Kiebel et al. 2008). These attempts are not interested in the temporal resolution of explanations in terms of PP per se, but rather in a PP-style explanation of the properties of temporality that characterizes our phenomenology.

## 5 Inside, Outside, and Beyond

EPACC speaks directly to the emerging discussion between Jakob Hohwy (Hohwy 2013) and Andy Clark (Clark 2016) about the wider theoretical ramifications of PP. In this section, I will consider how their views on PP sit with the EPACC account.

According to Hohwy (Hohwy 2013), PP commits us to thinking of cognition (traditionally conceived) in internalistic, neurocentric terms. This interpretation of PP emphasizes that "the sensory boundary between the brain and the world is not arbitrary, but rather a principled, indispensable, and epistemically critical element for identifying what is inferring what in perceptual and active inference" (Hohwy 2013, p. 240). On this construal, the "sensory boundary" separates the brain from the world in absolute terms, such that active inference and the optimization of precision estimations are nothing but functional attempts to improve the model parameters in the brain and the brain's certainty in the sensory signals it receives. The theoretical consequence is that "the mind remains secluded from the hidden causes of the world, even though we are ingenious in using culture and technology to allow us to bring these causes into sharper focus and thus facilitate how we infer to them" (Hohwy 2013, p. 239). This view, which I dub *seclusionism*, implies that the functional relevance of the embodied interaction with epistemic resources in the cognitive niche can be neglected if we try to understand the realization of cognitive processes in terms of prediction error minimization. This also applies to the developmental processes captured by the EPACC account. If "learning is just longer-term revision of hypotheses in the light of prediction error" (Hohwy 2013, p. 162), we must not consider the particular socio-cultural and bodily conditions that specify and constrain the acquisition and subsequent improvement of cognitive capacities. This is because learning would be a matter of prediction error minimization that is exclusively realized in the secluded brain. As a consequence, seclusionism is opposed to EPAAC, because EPACC emphasizes the indispensable functional role of the body and the socio-culturally shaped cognitive niche for the emergence of cognitive capacities.

Hohwy argues that "many of the ways we interact with the world in technical and cultural aspects can be characterized by attempts to make the link between the sensory input and the causes more precise (or less uncertain)" (Hohwy 2013, p. 238). On this view, the primary function of epistemic resources in the cognitive niche would be to increase the estimated precision given socio-culturally derived types of sensory input. It seems to me that Hohwy depicts the contribution of the cognitive niche to the optimization of precision estimations in relational terms. On this construal, epistemic resources would provide more precise sensory signals to the secluded brain when compared to those parts of the world that are detached from socio-cultural and technological epistemic resources.[5] This distinction between more and less precise sensory signals — based on the classification of their cultural or non-cultural causes — is hard to conceive, because human learning and cognizing are pervasively and inevitably situated in the socio-culturally structured cognitive niche. There simply does not seem to be an ecologically valid scenario under which the conceptual distinction between precise signals in the niche on the one hand and rather imprecise signals beyond the niche on the other hand would be meaningful and epistemically fruitful. Therefore, it would be highly speculative to argue that predictive systems would receive less precise sensory signals if they were detached from epistemic resources. If this is correct, learning in terms of prediction error minimization can only be understood if we take the entire "brain-body-niche nexus" into account (Menary 2015b). This is just the purpose of the EPACC approach.

Hohwy's (Hohwy 2013) idea that epistemic resources tend to increase the precision of sensory signals is questionable for another reason. The estimated precision of a prediction error signal is not fully determined by its worldly origin, e.g. by its connection to a salient aspect of a certain epistemic resource. Rather, it is determined by the *interaction* of the predictive system with that resource *across time*. For example, mathematical cognition appears to be dependent in a non-trivial sense on the tem-

5 Ubiquitous examples of technological resources are computers and smartphones that facilitate the completion of cognitive tasks.

porally extended embodied manipulation of mathematical symbols (Menary 2015a). Based on empirical evidence, Dutilh Novaes argues that "in typical cases manipulating portions of writing (broadly construed) is a fundamental aspect of mathematical practice" (Dutilh Novaes 2013, p. 60). In this sense, the importance of mathematical symbols would be neglected if we simply assumed that these symbols happen to provide potentially precise sensory signals. Rather, it is the on-going interaction with mathematical symbols — understood as tokens of a large-scale representational system — that brings about the completion of cognitive tasks. For this reason, I will argue that we can only understand the impact of the cognitive niche on prediction error minimization if we take the brain, the acting body, the world, and their intricate relationships into account. Vice versa, we can only understand the ubiquitous modification of the cognitive niche if we investigate the influence of embodied predictive systems on their local environment. According to the PP version of embodied action, the modification of the cognitive niche is primarily achieved by active inference, where active inference is a means of minimizing prediction error and of optimizing precision estimations. On Hohwy's (Hohwy 2013, p. 238) construal, "[t]his can generally be done by removing sources of noise in the environment and by magnifying signal strength." In these cases, active inference can only be understood if we consider its function in the context of brain-body-niche interactions (Fabry 2017). This becomes obvious when we consider the different temporal scales at which we can investigate the details of the acquisition of cognitive capacities in terms of EPACC.

On a physiological timescale, active inference is always relative to the efficiency of perceptual inference, current precision estimations, and states in the world at any given time. But the generation and execution of active inference is not only determined by the brain's current states (including the concentration of neurotransmitters in the postsynaptic gap), but also by that part of the cognitive niche that would change if an active inference were performed.

On an organismic timescale, embodied action employs and sculpts components of epistemic resources. That is, the sensory input derived from the cognitive niche is the result of previous interactions with these epistemic resources. This may explain why epistemic resources may provide precise sensory signals. Whether this kind of signal amplification applies is dependent on the particular organismic history of the predictive system's embodied interaction with specific epistemic resources. This is of particular relevance for EPACC.

Finally, on an evolutionary timescale the relationship between prediction error minimization, the evolution of the brain, the rest of the body, and the cognitive niche becomes even more intriguing. We have seen that human brains exhibit a high degree of LDP relative to functional biases and other constraints on brain functioning. We have also seen that LDP is continuous with LDBA. If the evolved primary function of the brain — in perceptual inference, active inference, and the optimization of precision estimations — were to render the sensory signals reliable and precise, it would be hard to assess how the abundance of neural reuse and the multiplicity of motor programs, cognitive resources, and institutionalized forms of scaffolded learning would have emerged during phylogeny. This potential shortcoming of seclusionism is important, because the temporal unfolding of cognitive capacities on both the organismic and the physiological scale is the result of the long phylogenetic history of humans' cognitive potential. The upshot is that we need the multi-level EPACC perspective, in addition to a non-seclusionistic interpretation of PP, if we wish to do conceptual justice to the complexity and interactive nature of our acquired and to-be-acquired cognitive capacities: "Our cognitive abilities are the result of a synergistic combination of internal neural mechanisms, bodily capacities and constraints, and environmental and social context; if we are ever to understand our brains, we must thoroughly absorb this lesson" (Anderson 2015, p. 249).

Andy Clark's (Clark 2016) interpretation of PP seems to be more compatible with this non- internalistic view of the phylogenetic and ontogenetic development of cognitive capacities. Like Hohwy (Hohwy 2013), Clark emphasizes the potentially precision-increasing role of components of the cognitive niche in prediction error minimization. On his view, "[o]rganismically salient (high precision)

prediction error may thus be the all-purpose adhesive that, via its expressions in action, binds elements from brain, body, and world into temporary problem-solving wholes" (Clark 2016, p. 262). In contrast to Hohwy (Hohwy 2013), Clark (Clark 2016) emphasizes the continuity of the prediction error minimizing processes realized in the human brain with the embodied interactions with epistemic resources necessary to complete cognitive tasks. Epistemic resources are supposed to deliver precise sensory signals. According to Clark, however, this insight can only be understood when we consider the predictive brain as part of a larger unit of analysis that does not stop at the skull, but extends into the rest of the body and the cognitive niche with its unique, and often salient, properties.

In this sense, supervised learning, as it is supposed to be realized by prediction error minimization (see Section 3), needs to be understood in terms of the predictive system's enculturating interaction with the epistemic resources in its cognitive niche. As Clark (Clark 2016, pp. 276-77) puts it, "[p]rediction-driven learning routines make human minds permeable, at multiple spatial and temporal scales, to the statistical structure of the action-ready, organism-salient world, as reflected in the training signals." EPAAC provides initial considerations on the trans-temporal and multi-level features of this tight relationship between the neuronal and bodily potential of predictive cognitive systems and the cognitive niche. The 'training signals' that induce the minimization of prediction error are more than specifically precise sensory inputs to the secluded brain. On Clark's (Clark 2016, p. 171) view, "the probabilistic inference engine in the brain does not constitute a barrier between agent and world. Rather, it provides a unique tool for encountering a world of significance, populated by human affordances" (italics in original). This suggests that the continuous process of prediction error minimization does not prevent, but rather enables the skillful embodied interaction with the cognitive niche. The upshot is that we can only understand the genuinely cerebral contribution to our cognitive endeavors if we relate this contribution to the possibility landscape for embodied action delivered by the cognitive niche. In this sense, EPACC is entirely consistent with Clark's suggestion that "prediction-driven learning delivers a grip upon affordances: the possibilities for action and intervention that the environment makes available to a given agent" (Clark 2016, p. 171; italics in original). Importantly, this consideration of predictive ways of acquiring certain cognitive capacities and affordances implies that these affordances are determined by the temporally extended trajectory of the co-evolution of human organisms and their cognitive niche (Estany and Martínez 2014). This suggests a profound connection between ontogenetic and phylogenetic development of the human brain and the rest of the body on the one hand, and the unfolding of ever new epistemic resources in the cognitive niche on the other. Thus, the developmental trajectory of the genuinely cerebral processing units cannot be isolated from the on-going modification of the cognitive niche realized by embodied interactions on both organismic and evolutionary timescales.

For this reason, it is epistemically justified to include the properties of the cognitive niche into our considerations of brain functioning and the brain-body interface. These broader ramifications of PP appear to be in line with the following suggestion:

> Action and perception then work together to reduce prediction error only against the more slowly evolving backdrop of a culturally distributed process that spawns a succession of practices and designer environments whose impact on the development […] and unfolding of human thought and reason can hardly be overestimated. (Clark 2016, p. 280)

The EPACC account can be understood as a modest proposal of how we could begin to appreciate this intricate relationship. It thus offers a contribution to the emerging debate over the most suitable interpretation of the PP framework. EPACC suggests that Hohwy's (Hohwy 2013) view of the secluded predictive brain neglects the influence of the cognitive niche and bodily interaction repertoires on the phylogenetic and ontogenetic unfolding of predictive ways of acquiring and improving cognitive capacities. If seclusionism were a position that had epistemic advantages over EPACC style inter-

pretations of PP, it would offer an argument for why we can reasonably neglect the socio-cultural conditions of genuinely human ways of completing cognitive tasks on physiological, organismic, and evolutionary timescales. Clark (Clark 2016), in contrast to Hohwy (Hohwy 2013), paves the way to interpretations of PP that take the relationship of the brain, the rest of the body, and the cognitive niche at physiological, organismic, and evolutionary timescales into account. On his view, "[t]he full potential of the prediction error minimization model of how cortical processing most fundamentally operates may thus emerge only when that story is paired with an appreciation of what immersion in a huge variety of sociocultural designer environments can do" (Clark 2016, p. 280). EPACC offers a first sketch of this relationship between prediction error minimization, embodied interaction, and the cognitive niche in the case of one of the most complex achievements of human cognition: the acquisition of cognitive practices.

## 6 Concluding Remarks

I have investigated how the emerging PP perspective can contribute to a better understanding of the conditions that lead to the acquisition of cognitive practices (and possibly many other types of cognitive capacities). I have suggested that PP provides us with mechanistic considerations that complement and enrich core principles of the phylogenetic and ontogenetic functional development of the brain and the rest of the body. If this line of reasoning is on the right track, it appears to be a tenable position to understand perceptual inference, active inference, and the optimization of precision estimations as a potent means to realize LDP, neural reuse, neural search, and LDBA. This idea is at the core of PACC. However, I have argued that we need to enrich PACC with a careful examination of the delicate relationship between the brain, the rest of the body, and the socio-culturally structured cognitive niche, and we need to do so across several timescales and at several levels of explanation. Therefore, consideration of enculturation and its components — especially the niche, the normative, and the scaffolding aspects — appears to be an epistemically justified complement to the descriptions offered by PACC. The core insight of EPACC is therefore that we can only understand the possibilities and limitations of human cognitive achievements if we take the entire human organism including its ongoing embodied interactions with epistemic resources in its cognitive niche into account. This insight offers a timely contribution to the emerging discussion on the ramifications of PP for our attempts to understand human cognition. In particular, EPACC reveals that we should be skeptical that the seclusionistic interpretation suggested by Hohwy (Hohwy 2013) provides us with the conceptual tools and hypotheses to allow us to account for the entire scope of the acquisition of cognitive capacities. Clark's (Clark 2016) view of predictive systems, however, seems to be more compatible with EPACC. This is because it advocates for the feasibility of considerations of cognitive growth that transcend the brain and include the rest of the body and the cognitive niche. At least as far as cognitive development and the acquisition of cognitive capacities are concerned, PP on its own may not have and need not to have the potential to cover all indispensable components that contribute to cognitive success (and failure). However, if PP is complemented by a thorough examination of enculturation and its enablers, we can begin to appreciate the possibilities and limitations of human cognition.

## References

Alsmith, A. (2012). The concept of structural affordance. *AVANT, III* (2), 94–107.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences, 33* (04), 245–266. https://dx.doi.org/10.1017/S0140525X10000853.

———(2015). *After phrenology: Neural reuse and the interactive brain.* Cambridge, MA: MIT Press.

——— (2016). Neural reuse in the organization and development of the brain. *Developmental Medicine and Child Neurology, 58* (Suppl 4), 3–6. https://dx.doi.org/10.1111/dmcn.13039.

Anderson, M. L. & Finlay, B. L. (2014). Allocating structure to function: The strong links between neuroplasticity and natural selection. *Frontiers in Human Neuroscience, 7*. https://dx.doi.org/10.3389/fnhum.2013.00918.

Anderson, M. L., Richardson, M. J. & Chemero, A. (2012). Eroding the boundaries of cognition: Implications of embodiment. *Topics in Cognitive Science, 4* (4), 717–730. https://dx.doi.org/10.1111/j.1756-8765.2012.01211.x.

Ansari, D. (2012). Culture and education: New frontiers in brain plasticity. *Trends in Cognitive Sciences, 16* (2), 93–95. https://dx.doi.org/10.1016/j.tics.2011.11.016.

——— (2015). Mind, brain, and education: A discussion of practical, conceptual, and ethical issues. In J. Clausen & N. Levy (Eds.) *Handbook of neuroethics* (pp. 1703–1719). Dordrecht, Springer Netherlands.

Ben-Shachar, M., Dougherty, R. F., Deutsch, G. K. & Wandell, B. A. (2011). The development of cortical sensitivity to visual word forms. *Journal of Cognitive Neuroscience, 23* (9), 2387–2399. https://dx.doi.org/10.1162/jocn.2011.21615.

Brem, S., Bach, S., Kucian, K., Guttorm, T. K., Martin, E., Lyytinen, H., Brandeis, D. & Richardson, U. (2010). Brain sensitivity to print emerges when children learn letter-speech sound correspondences. *Proceedings of the National Academy of Sciences, 107* (17), 7939–7944. https://dx.doi.org/10.1073/pnas.0904402107.

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

——— (2006). Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences, 10* (8), 370–374. https://dx.doi.org/10.1016/j.tics.2006.06.012.

——— (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. New York: Oxford University Press.

——— (2013a). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36* (03), 181–204. https://dx.doi.org/10.1017/S0140525X12000477.

——— (2013b). The many faces of precision (Replies to commentaries on "Whatever next? Neural prediction, situated agents, and the future of cognitive science"). *Frontiers in Psychology, 4*. https://dx.doi.org/10.3389/fpsyg.2013.00270.

——— (2014). *Mindware: An introduction to the philosophy of cognitive science*. Oxford, New York: Oxford University Press.

——— (2015). Embodied prediction. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1–21). Frankfurt am Main, MIND Group.

——— (2016). *Surfing Uncertainty: Prediction, action, and the embodied mind*. New York: Oxford University Press.

Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. In S. Dehaene, J.-R. Duhamel, M. D. Hauser & G. Rizzolatti (Eds.) *From monkey brain to human Brain: A Fyssen Foundation symposium* (pp. 133–157). Cambridge, MA, MIT Press.

——— (2010). *Reading in the brain: The new science of how we read*. New York: Penguin Books.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Filho, G. N., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J. & Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science, 330* (6009), 1359–1364. https://dx.doi.org/10.1126/science.1194140.

Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review, 3* (4), 357–370.

Dounskaia, N., van Gemmert, A. W. A. & Stelmach, G. E. (2000). Interjoint coordination during handwriting-like movements. *Experimental Brain Research, 135* (1), 127–140. https://dx.doi.org/10.1007/s002210000495.

Dutilh Novaes, C. (2013). Mathematical reasoning and external symbolic systems. *Logique & Analyse, 221*, 45–65.

Estany, A. & Martínez, S. (2014). "scaffolding" and "affordance" as integrative concepts in the cognitive sciences. *Philosophical Psychology, 27* (August), 98–111. https://dx.doi.org/10.1080/09515089.2013.828569.

Fabry, R. E. (2015). Enriching the notion of enculturation: Cognitive integration, predictive processing, and the case of reading acquisition - A commentary on Richard Menary. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1–23). Frankfurt am Main, MIND Group.

———(2017). Transcending the evidentiary boundary: Prediction error minimization, embodied interaction, and explanatory pluralism. *Philosophical Psychology*, 1–20. https://dx.doi.org/10.1080/09515089.2016.1272674.

Feldman, H. & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*. https://dx.doi.org/10.3389/fnhum.2010.00215.

Fletcher, P. C. & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symp-

toms of schizophrenia. *Nature Reviews Neuroscience, 10* (1), 48–58. https://dx.doi.org/10.1038/nrn2536.

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

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

Friston, K. & Buzsáki, G. (2016). The functional anatomy of time: What and when in the brain. *Trends in Cognitive Sciences, 20 (7)*, 500–511. https://dx.doi.org/10.1016/j.tics.2016.05.001.

Friston, K. J. & Stephan, K. E. (2007). Free-energy and the brain. *Synthese, 159* (3), 417–458. https://dx.doi.org/10.1007/s11229-007-9237-y.

Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. E. Patterson, J. C. Marshall & M. Coltheart (Eds.) *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading* (pp. 301–330). Hillsdale, N.J., Erlbaum.

Gaillard, W. D., Balsamo, L. M., Ibrahim, Z., Sachs, B. C. & Xu, B. (2003). FMRI identifies regional specialization of neural networks for reading in young children. *Neurology, 60* (1), 94–100. https://dx.doi.org/10.1212/WNL.60.1.94.

Harkness, D. L. (2015). From explanatory ambition to explanatory power - A commentary on Jakob Hohwy. *Open MIND*. https://dx.doi.org/10.15502/9783958570153.

Heyes, C. (2012). Grist and mills: On the cultural origins of cultural learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367* (1599), 2181–2191. https://dx.doi.org/10.1098/rstb.2012.0120.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3*. https://dx.doi.org/10.3389/fpsyg.2012.00096.

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

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

——— (2015b). Prediction error minimization, mental and developmental disorder, and statistical theories of consciousness. In R. J. Gennaro (Ed.) *Philosophical psychopathology: Disturbed consciousness: New essays on psychopathology and theories of consciousness* (pp. 293–324). Cambridge, MA, MIT Press.

Hohwy, J., Paton, B. & Palmer, C. (2016). Distrusting the present. *Phenomenology and the Cognitive Sciences, 15* (3), 315–335. https://dx.doi.org/10.1007/s11097-015-9439-6.

Huestegge, L., Radach, R., Corbic, D. & Huestegge, S. M. (2009). Oculomotor and linguistic determinants of reading development: A longitudinal study. *Vision Research, 49* (24), 2948–2959. https://dx.doi.org/10.1016/j.visres.2009.09.012.

Joseph, H. S. S. L. & Liversedge, S. P. (2013). Children's and adults' on-line processing of syntactically ambiguous sentences during reading. *PLoS ONE, 8* (1). https://dx.doi.org/10.1371/journal.pone.0054141.

Kendal, J. R. (2011). Cultural niche construction and human learning environments: Investigating sociocultural perspectives. *Biological Theory, 6* (3), 241–250.

Kiebel, S. J., Daunizeau, J. & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology, 4* (11). https://dx.doi.org/10.1371/journal.pcbi.1000209.

Laland, K. N. & O'Brien, M. J. (2011). Cultural niche construction: An introduction. *Biological Theory, 6* (3), 191–202.

Maurer, U., Brem, S., Kranz, F., Bucher, K., Benz, R., Halder, P., Steinhausen, H.-C. & Brandeis, D. (2006). Coarse neural tuning for print peaks when children learn to read. *NeuroImage, 33* (2), 749–758. https://dx.doi.org/10.1016/j.neuroimage.2006.06.025.

McCandliss, B. D., Cohen, L. & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences, 7* (7), 293–299. https://dx.doi.org/10.1016/S1364-6613(03)00134-7.

Menary, R. (2007). *Cognitive integration: Mind and cognition unbounded.* Basingstoke, New York: Palgrave Macmillan.

——— (2010a). Cognitive integration and the extended mind. In R. Menary (Ed.) *The extended mind* (pp. 227–243). Cambridge, MA: MIT Press.

——— (2010b). Dimensions of mind. *Phenomenology and the Cognitive Sciences, 9* (4), 561–578. https://dx.doi.org/10.1007/s11097-010-9186-7.

——— (2012). Cognitive practices and cognitive character. *Philosophical Explorations, 15* (2), 147–164. https://dx.doi.org/10.1080/13869795.2012.677851.

——— (2013). The enculturated hand. In Z. Radman (Ed.) *The hand, an organ of the mind: What the manual tells the mental* (pp. 349–367). Cambridge, MA: MIT Press.

——— (2014). Neural plasticity, neuronal recycling and niche construction. *Mind & Language, 29* (3), 286–303. https://dx.doi.org/10.1111/mila.12051.

——(2015a). Mathematical cognition: A case of enculturation. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1–20). Frankfurt am Main, MIND Group.

——(2015b). What? Now: Predictive coding and enculturation. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*.

——(2016). Pragmatism and the pragmatic turn in cognitive science. In A. K. Engel, K. Friston & D. Kragic (Eds.) *Where is the action? The pragmatic turn in cognitive science* (pp. 219–237). Cambridge, MA: MIT Press.

Menary, R. & Kirchhoff, M. (2013). Cognitive transformations and extended expertise. *Educational Philosophy and Theory, 46* (6), 610–623. https://dx.doi.org/10.1080/00131857.2013.779209.

Odling-Smee, J. & Laland, K. N. (2011). Ecological inheritance and cultural inheritance: What are they and how do they differ? *Biological Theory, 6* (3), 220–230.

Phillips, J. G., Ogeil, R. P. & Best, C. (2009). Motor constancy and the upsizing of handwriting. *Human Movement Science, 28* (5), 578–587. https://dx.doi.org/10.1016/j.humov.2009.07.004.

Piccinini, G. & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese, 183* (3), 283–311. https://dx.doi.org/10.1007/s11229-011-9898-4.

Price, C. J. & Devlin, J. T. (2003). The myth of the visual word form area. *NeuroImage, 19* (3), 473–481. https://dx.doi.org/10.1016/S1053-8119(03)00084-3.

——(2004). The pro and cons of labelling a left occipitotemporal region: "The visual word form area". *NeuroImage, 22* (1), 477–479.

——(2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences, 15* (6), 246–253. https://dx.doi.org/10.1016/j.tics.2011.04.001.

Quattrocki, E. & Friston, K. J. (2014). Autism, oxytocin and interoception. *Neuroscience & Biobehavioral Reviews, 47*, 410–430. https://dx.doi.org/10.1016/j.neubiorev.2014.09.012.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124* (3), 372–422. https://dx.doi.org/10.1037/0033-2909.124.3.372.

—— (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62* (8), 1457–1506. https://dx.doi.org/10.1080/17470210902816461.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D. & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2* (2), 31–74. https://dx.doi.org/10.1111/1529-1006.00004.

Rayner, K., Juhasz, B. J. & Pollatsek, A. (2007). Eye movements during reading. In M. J. Snowling & C. Hulme (Eds.) *The science of reading: A handbook* (pp. 79–97). Malden, MA, Blackwell Pub.

Roepstorff, A., Niewöhner, J. & Beck, S. (2010). Enculturing brains through patterned practices. *Neural Networks, 23* (8-9), 1051–1059. https://dx.doi.org/10.1016/j.neunet.2010.08.002.

Seassau, M., Bucci, M.-P. & Paterson, K. (2013). Reading and visual search: A developmental study in normal children. *PLoS ONE, 8* (7). https://dx.doi.org/10.1371/journal.pone.0070261.

Seth, A. K. (2015). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 1–24). Frankfurt am Main, MIND Group.

Snowling, M. J. (2000). *Dyslexia.* Malden, MA: Blackwell Publishers.

Sterelny, K. (2003). *Thought in a hostile world: The evolution of human cognition.* Malden, MA: Blackwell.

——(2012). *The evolved apprentice: How evolution made humans unique.* Cambridge, MA: MIT Press.

Stotz, K. (2010). Human nature and cognitive–developmental niche construction. *Phenomenology and the Cognitive Sciences, 9* (4), 483–501. https://dx.doi.org/10.1007/s11097-010-9178-7.

——(2014). Extended evolutionary psychology: The importance of transgenerational developmental plasticity. *Frontiers in Psychology, 5*. https://dx.doi.org/10.3389/fpsyg.2014.00908.

Turkeltaub, P. E., Gareau, L., Flowers, D. L., Zeffiro, T. A. & Eden, G. F. (2003). Development of neural mechanisms for reading. *Nature Neuroscience, 6* (7), 767–773. https://dx.doi.org/10.1038/nn1065.

Van Atteveldt, N. & Ansari, D. (2014). How symbols transform brain function: A review in memory of Leo Blomert. *Trends in Neuroscience and Education, 3* (2), 44–49.

Vogel, A. C., Church, J. A., Power, J. D., Miezin, F. M., Petersen, S. E. & Schlaggar, B. L. (2013). Functional network architecture of reading-related regions across development. *Brain and Language, 125* (2), 231–243. https://dx.doi.org/10.1016/j.bandl.2012.12.016.

Vogel, A. C., Petersen, S. E. & Schlaggar, B. L. (2014). The VWFA: It's not just for words anymore. *Frontiers in*

*Human Neuroscience, 8*. https://dx.doi.org/10.3389/fnhum.2014.00088.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Wood, D., Bruner, J. S. & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17* (2), 89–100.

Ziegler, J. C. & Goswami, U. (2006). Becoming literate in different languages: Similar problems, different solutions. *Developmental Science, 9* (5), 429–436. https://dx.doi.org/10.1111/j.1467-7687.2006.00509.x.

# Meeting in the Dark Room: Bayesian Rational Analysis and Hierarchical Predictive Coding

Sascha Benjamin Fink & Carlos Zednik

At least two distinct modeling frameworks contribute to the view that mind and brain are Bayesian: Bayesian Rational Analysis (BRA) and Hierarchical Predictive Coding (HPC). What is the relative contribution of each, and how exactly do they relate? In order to answer this question, we compare the way in which these two modeling frameworks address different levels of analysis within Marr's tripartite conception of explanation in cognitive science. Whereas BRA answers questions at the computational level only, many HPC-theorists answer questions at the computational, algorithmic, and implementational levels simultaneously. Given that all three levels of analysis need to be addressed in order to explain a behavioral or cognitive phenomenon, HPC seems to deliver more complete explanations. Nevertheless, BRA is well-suited for providing a solution to the *dark room problem,* a major theoretical obstacle for HPC. A combination of the two approaches also combines the benefits of an embodied-externalistic approach to resolving the dark room problem with the idea of a persisting evidentiary border beyond which matters are out of cognitive reach. For this reason, the development of explanations spanning all three Marrian levels within the general Bayesian approach may require combining the BRA and HPC modeling frameworks.

**Keywords**

Bayesian rational analysis | Dark room problem | Embodiment | Hierarchical predictive coding | Levels of analysis | Modeling frameworks

## 1 Introduction

Two methodologically distinct modeling frameworks contribute to the rising prominence of the view that the mind and brain are Bayesian. On the one hand, Bayesian Rational Analysis (BRA) is used in cognitive psychology to characterize behavioral and cognitive phenomena as forms of optimal probabilistic inference (Anderson 1991; Griffiths et al. 2008; Oaksford 2001; Oaksford and Chater 2007). On the other hand, Hierarchical Predictive Coding (HPC) is used in theoretical neuroscience to model information processing in the brain (Clark 2013; Friston 2010; Hohwy 2013; Rao and Ballard 1999). Although both of these modeling frameworks are grounded in the formal tools and concepts of Bayesian statistics, they differ with respect to their explanatory scope.[1] In particular, they address different *levels of analysis* in David Marr's tripartite conception of explanation in cognitive science (Marr 1982).

As is well-known, Marr argued that in order to "completely understand" the visual system, it must be analyzed at three distinct levels. Marr's levels can be distinguished according to the different types of questions investigators are likely to ask about a particular cognitive system (see also McClamrock 1991). The *computational level* is characterized by questions about what the system is doing, and why it is doing it. Questions of this kind can be answered by specifying mathematical functions that describe the system's behavior, and by determining the extent to which these functions reflect relevant structures in the environment (Shagrir 2010). In contrast, the *algorithmic level* of analysis concerns questions about how the system does what it does—questions that can be answered by specifying

1 Our focus is on BRA and HPC as *frameworks,* rather than on specific models developed within either one of these frameworks. Frameworks, in our understanding, provide tools to apply to certain phenomena, specify the particular kinds of parameters that are available for explanations, and provide methods for model-building. In Marrian terms, frameworks are akin to the languages in which answers to what-, why-, how- and/or where-questions are expressed, as well as the methods that are deployed to answer questions of each type. Models, in contrast, are akin to specific answers: they connect and set the available parameters in order to explain specific phenomena.

Fink, S. B. & Zednik, C. (2017). Meeting in the Dark Room: Bayesian Rational Analysis and Hierarchical Predictive Coding.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 14. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573154

the individual steps of an algorithm for computing or approximating the mathematical function that describes the system's behavior. Finally, at the *implementational level* of analysis, questions are asked about where in the brain the relevant algorithms are actually realized, by identifying individual steps of the relevant algorithm with the activity of particular physical structures in the brain *(*Zednik 2017).

Ever since Marr applied this three-level scheme to the phenomenon of visual perception, it has served as a backdrop for comparing and evaluating the explanatory scope of modeling frameworks in cognitive science quite generally. Therefore, the first aim of the present discussion is to highlight the differences between the BRA and HPC modeling frameworks by illuminating them against this backdrop. Specifically, it will be argued that whereas the BRA framework answers what- and why-questions and therefore speaks directly to Marr's computational level, it is neutral concerning the algorithmic and implementational levels of analysis. In contrast, proponents of HPC are keen to address all three levels of analysis simultaneously (see also Harkness and Keshava 2017).

The second aim is to explore the relationship between the BRA and HPC modeling frameworks, and to suggest that even though HPC is broader in scope and might therefore be thought to supplant BRA, they in fact complement each other in mutually beneficial ways. On the one hand, HPC puts paid to the allegation that the general Bayesian approach "eschews mechanism altogether" (Jones and Love 2011, p.173), because it answers questions at the algorithmic and implementational levels of analysis in addition to the computational level. On the other hand, as we will show, BRA helps to address concerns related to the *dark room problem* (Clark 2013; Mumford 1992; Sims 2017), which has been thought to undermine the explanatory credentials of HPC. Additionally, a combination of the two approaches marries the benefits of an embodied-externalistic approach to resolving the dark room problem with the idea of a persisting evidentiary border. Because they complement one another in these ways, a combination of the Bayesian Rational Analysis and Hierarchical Predictive Coding modeling frameworks offers a promising avenue to full-fledged Bayesian explanations in cognitive science.

## 2 Bayesian Rational Analysis and the Computational Level

Bayesian approaches in cognitive science are motivated by the insight, often attributed to Hermann von Helmholtz, that many kinds of behavior and cognition can be viewed as solutions to problems of inference under uncertainty (von Helmholtz 1867). For example, perception can be viewed as a solution to the problem of inferring the cause of a particular sensation ("Was it a bird?"), and motor action might be viewed as a solution to the problem of selecting an adequate course of action ("Should I try to catch it?"). In line with Helmholtz' insight, the aim of BRA is to formally characterize a cognitive system's behavior as an optimal solution to a particular probabilistic inference task in the environment (Anderson 1991; Griffiths et al. 2008; Oaksford 2001; Oaksford and Chater 2007). To this end, proponents of BRA specify the probabilistic inference tasks in which cognitive systems appear to be engaged, formally characterize the environment in which these tasks are solved, and derive optimal solutions to those tasks according to the rules of probability theory, most notably among them Bayes' *rule*:

$$P(H|\mathrm{E}) = \frac{P(E|\mathrm{H}) \times P(H)}{P(E)}$$

We can consider the left side of this equation to be a formalization of Helmholtz' insight: *P(H|E)* is the *posterior probability* of some hypothesis *H* (e.g. "It was a bird"), given evidence *E* that may speak either for or against it (e.g. "I saw a yellow beak"). The right side prescribes that the posterior probability should depend on the *prior probability P*(*H*) that the hypothesis is true independent of the evidence (e.g. the probability of encountering birds in a given environment), as well as on the *likelihood P(E|H)* that evidence *E* will be available if *H* is in fact true (e.g. the salience of beaks) and the probability of

encountering evidence *E* independent of the truth of *H*. Notably, because natural environments are unpredictable and complex, real-world inference typically involves considering not just the probability of a single hypothesis, but rather considering a distribution of probabilities over a space of competing hypotheses. Indeed, many different things could have caused the relevant sensation (e.g. a bird, a plane, Superman), and many different behavioral actions could be performed (e.g. catching, but also running, screaming, laughing maniacally).

The optimal solutions derived using Bayes' rule closely approximate the behavioral data in a wide variety of behavioral and cognitive domains. Phenomena as varied as perceptual cue-combination (Ernst and Banks 2002), memory and categorization (Anderson 1991), judgment and decision making (Griffiths and Tenenbaum 2006), sensorimotor learning (Körding and Wolpert 2004), reasoning (Oaksford 2001), and language learning (Xu and Tenenbaum 2007) can all be viewed as forms of optimal probabilistic inference in an uncertain environment.[2] But what exactly is the explanatory import of this finding? The characterization of behavior and cognition as a form of optimal probabilistic inference allows investigators to answer questions at Marr's computational level of analysis (Jones and Love 2011; Oaksford 2001; Oaksford and Chater 2007; Zednik and Jäkel 2016; Harkness and Keshava 2017; cf. Bowers and Davis 2012). Specifically, it allows them to answer questions about *what* a cognitive system is doing, and *why*. In general, whereas an answer to a what-question is delivered by describing a system's behavior, an answer to a why-question is delivered by demonstrating this behavior's "appropriateness" with respect to the "task at hand" (Marr 1982: 24; see also: Shagrir 2010). BRA delivers on both counts. Regarding questions about *what* a system is doing, whenever an optimal solution is closely approximated by the behavioral data, the former provides an empirically adequate description of the latter. As for questions about *why* a system does what it does, there is a clear sense in which the system can be thought to behave as it does *because* that way of behaving is optimal in the sense prescribed by probability theory.[3]

The fact that BRA is purpose-built for answering what- and why-questions at the computational level distinguishes it from many other modeling frameworks in cognitive science. Traditional frameworks such as classical computationalism and connectionism are designed to answer questions at the algorithmic level of analysis about *how* the relevant system does what it does, and to a lesser extent, questions at the implementational level about *where* in the brain the relevant structures and processes are located. Put differently, whereas BRA is mostly concerned with describing behavioral and cognitive phenomena as well as with assessing their appropriateness with respect to some particular task environment, most other modeling frameworks in cognitive science are designed to describe the component parts, operations, and organization of the *mechanisms* responsible for these phenomena (Bechtel and Richardson 1993; Piccinini and Craver 2011). Because BRA remains neutral with respect to the algorithmic and implementational levels, however, it has been accused of "eschew[ing] mechanism altogether" (Jones and Love 2011, p.173).

To what extent is BRA's focus on the computational level and simultaneous neglect of mechanisms a virtue rather than a vice? Insofar as the computational level of analysis—and in particular, the answering of why-questions—remains somewhat underappreciated (Marr 1982; Shagrir 2010), BRA is poised to make an important explanatory contribution: By providing formal answers to questions

2 Investigators may also often find deviations from optimality, of course. Some theorists argue that such deviations show that real cognizers are not ideally rational in the Bayesian sense (Kwisthout and van Rooij 2013), and that the models developed in BRA should be considered normative models that set a benchmark on performance, rather than descriptive models thereof (Colombo and Series 2012). However, proponents of this modeling framework also regularly tweak their assumptions about the statistical structure of the environment until the model does in fact accommodate the data (Anderson 1991; Bowers and Davis 2012). In this way, they are often able to preserve the assumption that cognitive systems behave optimally in the sense prescribed by probability theory (for discussion see Zednik and Jäkel 2016).

3 Many proponents of BRA take themselves to be answering why-questions in this way (e.g. Griffiths et al. 2012; Oaksford and Chater 2007). Notably, in line with Marr's own understanding of what it takes to answer why-questions, no appeal is made to ontogenetic or phylogenetic considerations. Although some commentators have argued that this way of answering why-questions is explanatorily deficient (Danks 2008), others have defended it (Shagrir 2010; Zednik and Jäkel 2016). Whether or not direct reference is made to a particular behavior's ontogenetic or phylogenetic history, characterizing it as an optimal solution might suggest "why natural selection might favor one mechanism rather than another" (Griffiths et al. 2012).

about what a cognitive system is doing, proponents of this modeling framework can attain a heightened understanding of the nature of cognition and behavior itself, including its mathematical structure. As for why-questions, BRA may be poised to contribute to our understanding of a particular behavior's teleology and role in a containing environment (Griffiths et al. 2012; Oaksford and Chater 2007)—despite the fact that questions remain about how teleological considerations factor into explanation in cognitive science (Zednik 2017; cf. Danks 2008).

That said, in line with Marr's three-level account, it would be a mistake to think that answering what- and why-questions is sufficient for the purposes of explaining a behavioral or cognitive phenomenon. To wit, Jones and Love have recently argued that:

> [I]t would be a serious overreaction simply to discard everything below the computational level. As in nearly every other science, understanding how the subject of study (i.e., the brain) operates is critical to explaining and predicting its behavior [… M]echanistic explanations tend to be better suited for prediction of new phenomena, as opposed to post hoc explanation. [...] Much can be learned from consideration of how the brain handles the computational challenge of guiding behavior efficiently. (Jones and Love 2011, p. 177)

In other words, the explanatory success of the general Bayesian approach arguably depends on the extent to which the computational-level insights delivered by BRA can be supplemented with insights into behavioral and cognitive mechanisms at the algorithmic and implementation levels of analysis.[4]

Unfortunately, there is considerable disagreement about how best to supplement the BRA modeling framework so as to address questions at levels below the computational. Some investigators—most notably proponents of the so-called *Bayesian coding hypothesis (*Knill and Pouget 2004; Ma et al. 2006)—have sought to identify probability distributions and Bayes' rule with specific physical structures and processes in the brain. However, it would be a mistake to think that the answers BRA provides at the computational level impose significant constraints on the answers that may be given to questions at the algorithmic and implementational levels. As Marr himself has previously argued, "there is a wide choice available at each level, and the explication of each level involves issues that are rather independent of the other two" (Marr 1982, p. 25). Indeed, although the Bayesian coding hypothesis may yet be confirmed, the ability to describe behavior and cognition as a form of optimal probabilistic inference at the computational level does not require or even imply that the brain actually invokes Bayes' rule to compute over probability distributions (Colombo and Series 2012; Maloney and Mamassian 2009). Perhaps for this reason, an increasing number of investigators instead co-opt techniques from machine learning and artificial intelligence to develop biologically plausible algorithms that approximate optimal probabilistic inference without directly implementing either Bayes' rule or probabilistic representations (e.g. Griffiths et al. 2015; Sanborn et al. 2010, cf. Kwisthout and van Rooij 2013). However, there exist a great number of options, and few principled guidelines for how to choose between them (see Zednik and Jäkel 2016 for discussion). In general, therefore, despite the fact that the BRA modeling framework is useful for answering what- and why-questions at the computational level, it remains unclear how to proceed so as to develop full-fledged scientific explanations that span all three of Marr's levels.

## 3 Hierarchical Predictive Coding: Complement or Alternative?

Although there may be many different ways in which to supplement the computational-level insights provided by BRA, it is worth considering one particularly prominent candidate: Hierarchical Predictive Coding (HPC). HPC-theorists have developed a wide range of algorithms that exhibit a common

4 Without such supplementation, it remains unclear how some form of abstract optimality on the computational level can be interpreted as teleologically apt at all. Teleology itself explains too little if it is not grounded in specific mechanism exposed to evolutionary pressure. Vice versa, specific mechanism may explain too little if not considered in the broader context of their place in a system within its ecological niche.

computational architecture: a hierarchy of processing stages, where each higher stage is tasked with predicting the state of the preceding stage, and where each lower stage forwards an error signal—a measure of a prediction's accuracy—to the higher stage.[5] At every stage in this hierarchy, Bayes' rule is used (or approximated) to combine past predictions with error signals so as to result in the construction of increasingly veridical representations of the world. Although these representations are typically used to infer the causes of perceptual stimuli (Rao and Ballard 1999), proponents of the HPC modeling framework have also argued that cognitive systems often "bring the world in line" by "seeking or generating the sensory consequences that they (or rather, their brains) expect" (Clark 2013, p. 186; see also: Friston 2005).

Whereas the BRA modeling framework can be used to formally characterize perception and action as forms of optimal probabilistic inference, HPC is used to develop algorithms that actually perform this kind of inference. That is, the algorithms developed within the HPC modeling framework can be viewed as potential answers to questions about *how* a particular cognitive system does what it does, i.e. as descriptions of the functional processes that contribute to that system's behavior. Although much work has yet to be done to determine which (if any) of these algorithms actually constitute a *correct answer*—i.e. a true description of functional processes in our brains—the fact that answers to how-questions are being developed is often considered the central explanatory contribution of HPC (e.g., Spratling 2013).

Although the focus may be on the algorithmic level of analysis, many HPC-theorists also make it a point to address questions at the implementational level that ask *where* in the brain the relevant algorithms might be realized. To this end, they identify the particular steps of an HPC-algorithm, or elements of the general HPC-architecture, with particular neuronal structures or processes (e.g. Bastos et al. 2012). For example, the claim that perception and action depend on the propagation of predictions and error signals has motivated the search for specific neural pathways along which this two-way propagation could take place. In particular, Friston (Friston 2005, p. 829) proposes to identify such pathways in "functionally distinct subpopulations [of neurons]". He suggests the deep pyramidal cells as the locus of error propagation, and the superficial pyramidal cells as pathways for transmitting expectations (see also: Friston 2009). In this way, in addition to answering how-questions at the algorithmic level of analysis, proponents of HPC also often seek to answer where-questions at the implementational level.

Insofar as HPC promotes the formulation of testable claims about the algorithms that are used to perform optimal probabilistic inference, and about the neural structures in which these algorithms are implemented, HPC and BRA might be thought to complement one another. Clark appears to suggest as much when he argues that:

> [T]he hierarchical and bidirectional predictive processing story, if correct, would rather directly underwrite the claim that the nervous system approximates, using tractable computational strategies, a genuine version of Bayesian inference. The computational framework of hierarchical predictive processing realizes, using the signature mix of top-down and bottom-up processing, a robustly Bayesian inferential strategy, and there is mounting neural and behavioral evidence [...] that such a mechanism is somehow implemented in the brain. (Clark 2013, p. 189)

That said, although Clark espouses the idea that human and animal cognizers may in fact perform optimal probabilistic inference—the central *claim* of BRA—it is worth noting that he does not explicitly endorse the *methods* of BRA. Indeed, it is fair to question whether these methods provide explanatory insights that go beyond the ones delivered by HPC. Although answers to what- and why-questions at the computational level may not impose significant constraints on the algorithmic and implementa-

5 Friston (Friston 2010, p. 10) presents an overview of such algorithms. See also (Sims 2017) for a review and comparison of different interpretations of the HPC framework.

tional levels, the opposite may still be true. Algorithms always produce a particular output that can be measured or described. Therefore, an understanding of *how* a cognitive system does what it does should allow investigators to understand *what* that system is actually doing. Indeed, the algorithms used in the HPC modeling framework are known to compute or approximate optimal solutions to problems of probabilistic inference under uncertainty (Rao and Ballard 1999; Friston 2010). Thus, HPC-theorists agree with proponents of BRA that cognitive systems optimally solve probabilistic inference tasks in their environments, but they arrive at this conclusion indirectly, via the algorithmic level, rather than directly, by considering the computational level itself. In this sense, the generic answer to what-questions given by proponents of BRA is implicit in the answers given to how-questions by proponents of HPC.

In addition to answering what-questions, the HPC modeling framework also answers questions about *why*. Many HPC-theorists argue that cognitive systems behave as they do *because* that way of behaving leads to the minimization of prediction error (e.g. Clark 2013; Friston 2009; Friston 2010). At first glance, this may seem to differ from the generic way of answering why-questions in the BRA modeling framework, which appeals to the claim that cognitive systems behave as they do because that way of behaving is optimal with respect to the relevant task environment. Indeed, whereas HPC answers why-questions by looking at features *internal* to a particular system—the algorithms being deployed—BRA answers these questions by considering *external* features, namely the statistical structure of environment in which that system is situated. Still, HPC's answers to why-questions entail the answers developed in BRA: prediction error will be minimized whenever Bayes' rule is applied to update representations of the external world, and whenever cognitive systems act so as to "bring the world in line". By minimizing prediction error in either one of these two ways, the system's behavior inevitably approaches optimality in the sense prescribed by probability theory. Therefore, like the answers to what-questions, the answers to why-questions developed within the BRA modeling framework are in fact entailed by the answers developed in HPC.

In summary, although the HPC modeling framework is most clearly directed at the algorithmic and implementational levels of analysis, it is also well-suited for answering questions at the computational level. Perhaps for this reason, while BRA is plagued by the accusation that it "eschews mechanism" (Jones and Love 2011), HPC-theorists regularly present their approach as a unifying framework that is capable of simultaneously addressing all three levels of analysis (e.g. Clark 2013; Hohwy 2013). In this sense, the explanatory scope of HPC exceeds the scope of BRA. Not only that, the scope of HPC appears to fully subsume the scope of BRA. Because the answers given to computational-level questions by proponents of HPC entail the answers that would also be given by advocates of BRA, it is unclear what BRA's own unique contribution actually is. Does HPC render BRA superfluous?

## 4 Meeting in the Dark Room

BRA can contribute to a unified Bayesian conception of the mind in several ways. We will mainly focus on how the methods and practices of BRA are poised to solve one of HPC's most pernicious puzzles, the *dark room problem* (Mumford 1992; Sims 2017; see also the commentary on Clark 2013). But, in passing, we will address how BRA provides us with additional interpretational tools to understand behavior, and how BRA complements HPC-explanations such that we may distinguish how-possibly from how-actually explanations. If BRA contributes in these ways, then although HPC exceeds BRA in explanatory scope and subsumes its answers to what- and why-questions, there are reasons to believe that the development of satisfying three-level explanations involves a combination of resources from both modeling frameworks.

Recall that, from the perspective of HPC, a cognizer behaves as it does in order to minimize prediction error. It can do so either by adjusting its represented predictions about what happens in the environment so that they better fit the incoming sensory data, or by acting in order to make the

world match its predictions. Minimization of prediction error therefore can come in both directions of mind-world-fit. In the fleeting, dynamic world which we inhabit, however, one particularly easy way to ensure this kind of fit is to seek an evenly heated, silent, dark place which deprives the system of any sensory stimulation whatsoever, and to predict that nothing about this place will change. In such a "dark room", the error for predicting that everything will stay the same is minimized—because *ex hypothesi,* nothing ever changes. Therefore, seeking such a dark room would appear to be a cognizer's best strategy for prediction-error minimization. Evidently, however, there is a mismatch between this strategy and the ways in which biological cognizers actually behave: In the real world, we avoid such dark rooms for much of our lives. HPC seems unable to explain why cognizers are found playing bridge in living rooms, chasing mates in noisy clubs, listening in lecture halls, navigating the woods and busy market streets, and so on. The erratic nature of such dynamic, complex, and chaotic environments increases the chance that predictions will fail. Still, we prefer them over dark rooms. The fact that real-world cognizers are regularly found in dynamic, unpredictable environments challenges the adequacy of the generic HPC-answer to why-questions: cognitive systems do what they do because it minimizes prediction error. The best strategy for doing just that—going into the dark room—appears to be widely ignored.[6]

Can HPC-theorists explain why cognitive systems avoid dark rooms, and instead behave in far more interesting ways? Several HPC-theorists have tried to explain why certain features of our cognitive architecture lead us to avoid dark rooms. For example, Andy Clark argues:

> [C]hange, motion, exploration, and search are themselves valuable for creatures living in worlds where resources are unevenly spread and new threats and opportunities continuously arise. This means that change, motion, exploration, and search themselves become predicted. (Clark 2013, p. 193)

Hohwy similarly argues that predictions about surprisal rates of an internalized model—so called "hyper-priors"—are what keep us out of dark rooms:

> [W]e don't end up in dark rooms. We end up in just the range of situations we are expected to end up in on average. It is true we minimize prediction error and in this sense get rid of surprise. But this happens against the background of models of the world that do not predict high surprisal states, such as the prediction that we chronically inhabit a dark room. (Hohwy 2013, p. 87)

Finally, Schwartenbeck et al. 2013 strike a similar chord when they analyze exploration as a comparison between two different models the agent has of itself: the agent predicts that it will perform diverse actions in the future, compares these predictions to the actions it is currently performing, and if there is a mismatch, acts in order to minimize the prediction error.

On each one of these responses, dark rooms are avoided because the sensory stimuli encountered in such rooms diverge from the predicted stimuli. Notably, saying that change, motion, exploration, search, future acts and surprisal states themselves become predicted is tantamount to saying that they are internally represented at some level of the hierarchy. Thus, responses of this type can be thought to be *internalistic*: It is an internal feature of the system that contributes to the avoidance of dark rooms. As such, this response is well in line with HPC's focus on the algorithmic level: it seems easy to encode such a prediction as, for example, a prior probability at a particular level of the processing hierarchy. But there is reason to be unsatisfied with any internalistic response insofar as the likelihood of these

6 According to Schwartenbeck et al. 2013, the dark room problem brings together two questions. First, why does the imperative to minimize prediction error not lead us to seek dark rooms? Second, how does HPC motivate the active exploration of new states? Both are, however, entangled: If there is a good answer to the second question and if that answer can be generalized, an answer to the first question is in reach; and any answer to the first question must, if it is adequately detailed, suggest an answer to the second question. *Pace* Schwartenbeck et al. 2013, we therefore treat these questions together.

internalistic predictions themselves can be adjusted. It is accurate that going into a dark room would increase prediction error on the level where change, motion, exploration, etc. are predicted, but it would seem to simultaneously *decrease* prediction error on lower levels of the hierarchy which are closer to the sensorimotor periphery: whenever a cognizer is situated in the dark room, predicting that everything will remain dark produces no error at these lower levels. Thus, a question remains: at which hierarchical level should prediction error be minimized?

As long as both predictions about sensory input and surprisal rates etc. are malleable, no answer follows by necessity. Should the higher level predictions change so as to fit the incoming sensory data in a dark room? Or should the higher level predictions remain fixed, so that the system moves away from the dark room so as to bring the incoming sensory information in line with the higher level prediction? The internalist's answer would be: act in such a way that the error for higher-level predictions is minimized. But an equally adequate strategy would seem to be: lower the precision estimates associated with these predictions while staying in the dark room. Why shouldn't an agent decrease the strength of its predictions that the world will change, that it will perform diverse acts in the future, or that it will inhabit high surprisal states? Insofar as any one of these strategies would lead to the avoidance of dark rooms, it is unclear why the internalist response should be preferred.[7] At the same time, it is unclear why exactly this choice is to be preferred over the alternative of staying in the dark room. In other words, any of the proposed models, understood in such a way that the predictions are all malleable, might explain why an organism left the dark room (if it did); but they can equally well explain why an organism stayed in the dark room (if it does).[8]

There is an extended (or embodied) counterpart to the internalistic response that may be better suited to avoiding dark rooms. According to this extended view, some predictions or priors are kept stable by tying them to fixed features of the organism or its environment. For example, Karl Friston expresses it as follows:

> [E]very organism (from viruses to vegans) can be regarded as a model of its econiche, which has been optimized to predict and sample from that econiche. [...] This means that a dark room will afford low levels of surprise if, and only if, the agent has been optimized by evolution (or neurodevelopment) to predict and inhabit it. Agents that predict rich stimulating environments will find the "dark room" surprising and will leave at the earliest opportunity. This would be a bit like arriving at the football match and finding the ground empty. (Friston et al. 2012, p. 3)

Crucially, on this response, "model" encompasses the system's "interpretive disposition, morphology, and neural architecture, and as implying a highly tuned 'fit' between the active, embodied organism and the embedding environment" (Friston et al. 2012, p. 6). This answer is not fully internalistic, but mixes internal and external aspects. Specifically, it assumes a matching of external factors, such as the configuration of the ecological niche an embodied cognitive system inhabits, with the internal model the organism has of that niche. On this extended response, we do not dwell in dark rooms because we are embodied agents that need to sustain homeostasis in a world where means and resources are unevenly spread in a changing environment (Klein in press), and we represent the world as such.[9]

This extended response is better suited to answering the dark room problem than the internalist alternative, because some of the model's predictions are in fact ineligible for Bayesian updating: an animal's morphology or fit to its econiche cannot be altered in the same quick way as its internal rep-

7 This kind of uncertainty affects any solution that relies on a comparison between two represented probability distributions where both can be altered. In order to minimize the Kullback-Leibler-Distance between them (and thereby minimize prediction error or free energy), either one of the compared distributions can be altered. Defenders of an internalistic response to the dark room problem choose to alter distributions at lower hierarchical levels over distributions at higher levels, such that the organism has to act in order to match the input to these novel predictions.

8 See Klein in press for further critique of current solutions to the dark room problem.

9 Here, we focus on a reading where body and environment not merely influence a cognitive model but are actual parts of this model. We believe that such views can be found in Friston et al. 2012 and arguably in Bruineberg et al. forthcoming.

resentations. For this reason, the argument against the purely internalistic answer outlined above does not apply: if some of the predictions of change, motion, exploration, and search are embodied rather than represented, they necessarily remain fixed. Thus, there is only one way for such agents to act, as they cannot adjust these embodied priors.[10] For embodied systems, moving into a dark room will not decrease the likelihood of predictions of change, exploration, motion, nor change or alter their influence on behavior. Rather, these predictions are anchored in the body and its adaptation to the specific econiche of the organism (see also Bruineberg et al. forthcoming). Therefore, going into a dark room is not an option for any systems not adapted to caves; anything but sessile cave dwellers will inherently prefer dynamic environments.

By adopting an extended or embodied solution to the dark room problem, HPC departs from the brain-centric focus advocated by some of its proponents (e.g. Hohwy 2013). Unfortunately, this has the disadvantage of blurring a cognitive system's boundaries. By viewing a system's morphology and econiche as being part-and-parcel of its "model" of the environment, we lose the clear demarcation between the evidence that is available to the system and what this is evidence for (see Hohwy 2016 as well as Hohwy 2017, and Clark 2017). Intuitively, there should be a difference between predictions made and the evidence that is used to evaluate them. Hohwy 2016 expresses this view when he argues that there should be a tightly woven evidentiary blanket which makes part of the world, as well as our own bodies, cognitively unavailable to us—their properties are to be inferred but are not directly available or immediately known. If morphology and ecological niche are themselves part of the model, however, this evidentiary blanket is lost. Therefore, this solution is unlikely to be attractive to those proponents of HPC who hope to retain a clear distinction between predictive mind and predicted world (see also Hohwy 2013).

Is there a way to retain the sharp boundary between internal and external while also getting the benefits of the extended response? Here, the tools and methods of BRA might complement the ones of HPC. BRA is purpose-built for specifying the task environments inhabited by particular cognitive systems. Indeed, BRA-theorists have developed specialized techniques with which to formalize assumptions, including about the nature of the hypothesis space being considered; the prior knowledge possessed; the likelihood of experiencing particular stimuli given that certain hypotheses are true for an environment; and perhaps most importantly, the relative costs or benefits of particular actions in a particular environment (Anderson 1991; Oaksford and Chater 2007). Recall that proponents of BRA invoke Bayes' rule to compute posterior probability distributions over a space of, for example, possible causes of a particular visual stimulus. Thus, they assume that real-world behavior depends not only on an estimation of which causes are the most probable, but also on a calculation of which course of action is the most prudent. Indeed, behavioral actions typically have consequences that should influence whether or not they are actually performed: to any villainous inhabitant of Metropolis, erroneously classifying Superman as a bird (leading to a false feeling of security, detection, and swift justice) will be more costly than erroneously classifying a bird as Superman (which merely incurs ridicule). For this reason, even if the posterior probability of "It's a bird!" is high, the villain's best course of action might be to declare "It's Superman!" in order to avoid swift justice. Notably, *Bayesian Decision Theory* may be used to specify how posterior probability distributions should be combined with *cost functions* that formalize such consequences, so as to minimize the expected costs to the organism. Proponents of the BRA modeling framework regularly incorporate such cost functions into their computational-level characterizations of human and animal behavior (for discussion see e.g. Gershman and Daw 2012).

Some HPC-theorists are averse to invoking formal constructs such as cost functions that go beyond the calculation of probably distributions—some actively avoid them or think that they can do without (see Schwartenbeck et al. 2013 and Friston et al. 2012). But there are reasons to think this

10 That is, given the clash between sensory input and the predictions of search and change, only one particular probability distribution can be altered in order to reduce the relevant Kullback-Leibler-Distance, as the other is not subject to adjustment.

aversion is ill-advised. First, cost functions are just the kind of formal construct that may be needed to add precision to Clark's rather intuitive appeal to "change, motion, exploration, and search" and related proposals by other HPC-theorists. Whereas dark-room-seeking behavior could be associated with a high cost to the organism, explorative behavior might be rewarded. In this way, BRA provides formal interpretational tools that help to describe why going into a dark room is in fact unreasonable given a dynamic environment where resources are unevenly spread. HPC might have formal tools to model under which circumstances an organism does not go into the dark room. But what is lacking is a tool for evaluating the extent to which this behavior can be seen as *reasonable* or *prudent*. As well as adding further formal tools to the Bayesian toolkit, BRA and Bayesian Decision Theory together provide an interpretational tool for understanding behavior. These interpretational tools can be used to supplement the suspicions of HPC-theorists that dark-room-dwelling would be detrimental to an organism, by precisely modeling the system's behavior *in relation to its environment* in a way that the customary tools and concepts of HPC cannot. At the same time, these tools do not require that a cognitive system's body and econiche themselves be viewed as a "model". Rather, bodily features and environmental constraints are encoded as costs and benefits that are poised to influence the relevant system's behavior.

Thus supplemented by the tools of BRA and Bayesian Decision Theory, the HPC modeling framework has a way of explaining why cognizers generally avoid dark rooms, while retaining a clear distinction between cognizers and their environment. However, some HPC-theorists reject this story as they maintain that as long as we minimize surprisal in our world, exploration and high-utility-gaining states come naturally. For example, Schwartenbeck et al. 2013 write that:

> [M]inimizing surprise leads naturally to concepts such as exploration and novelty bonuses. In this approach, agents infer a policy that minimizes surprise by minimizing the difference (or relative entropy) between likely and desired outcomes, which involves both pursuing the goal-state that has the highest expected utility (often termed "exploitation") and visiting a number of different goal-states ("exploration"). (Schwartenbeck et al. 2013, p. 1)

Views like these bring us to another proposal of how BRA contributes to HPC, because they raise the question: why does such behavior lead to the highest expected utility? One can only say that it maximizes expected utility if a model of the environment is implicitly presumed where such behavior does maximize expected utility. The tools of BRA are ideally suited to make this implicit model of the organism-environment-coupling explicit.

We also believe that including the formal tools of BRA circumvents another problem of HPC-accounts: only HPC-accounts with fixed and stable (or at least specific) priors can account for observed behavior which is stable across organisms of a species. But how do such fixed priors come to be? HPC-theorists ought to give some explanation as to why we have the priors we have which explain the behavior we show. Giving such explanations often involves references to the environment in which an organism evolved and developed. But Hohwy (Hohwy 2013; see also Bowers and Davis 2012) warns us of Bayesian *just-so stories*, where we merely hypothesize about how a prior might arise without actually checking for evidence for such stories:

> The challenge […] is to avoid just-so stories. That requires avoiding priors and likelihoods that are posited only in order to make them fit an observed phenomenon. To avoid just-so stories any particular ordering of priors and likelihoods should be supported by independent evidence, which would suggest that this ordering holds across domains. (Hohwy 2013, p. 94)

One source of such independent evidence for specific priors might come from BRA, where we model the environment with its specific features, distribution of resources, and the costs and bene-

fits for certain actions. This then might serve as a tool for distinguishing Bayesian just-so stories or how-possibly-explanations from how-actually-explanations. Supplementing a successful HPC-story with some of BRA's tools then gives us a broader Bayesian explanation of the behavior we exhibit and the mind that brings it about.[11]

Therefore, we believe that BRA complements HPC. First, because only HPC-solutions where some priors are fixed or ineligible for Bayesian updating can explain why *all* nontroglobionitic animals avoid dark rooms. A prominent solution for this is an extended account where certain stable features of both organism and environment are part of the organism's model, thereby fixing certain priors. However, this endangers the strong evidentiary blanket, which some HPC-theorists like to maintain. In order to preserve this blanket, the tools of BRA can be used to explain why certain behaviors come with certain costs and benefits and are therefore performed. Second, because BRA provides an interpretational tool, telling us why certain behaviors are prudent. Third, if we avoid cost-functions and argue that minimizing surprisal comes with high expected utility, our best evidence for this is an implicit model of the environment, which can best be made explicit by using BRA. Fourth, if HPC-explanations rely on specific priors, BRA may help us to distinguish Bayesian how-possibly-explanations (just-so stories) from how-actually-explanations. We take these reasons as sufficient for advocating a combination of BRA and HPC.

## 5 Conclusion

We have sought to clarify how the Bayesian Rational Analysis and Hierarchical Predictive Coding modeling frameworks relate, and did so by comparing them vis-à-vis Marr's influential three-level conception of explanation in cognitive science. Whereas BRA-theorists answer questions at the computational level of analysis only, HPC-theorists focus primarily on the algorithmic level, while also addressing the computational and implementational levels. Given that answering questions at the computational level is insufficient for full-fledged explanation, the methods and practices of HPC can appear to offer a far more likely avenue to explanatory success. Nevertheless, because BRA (i) appears well-suited for supporting a solution to the dark room problem due to its specific tools and concepts for modeling task environments, (ii) provides an interpretational tool (iii) allows us to make implicit assumptions about the structure of the environment explicit, and (iv) helps to distinguish how-possibly- from how-actually-explanations, it appears that three-level explanations of behavior and cognition are most likely to be forthcoming if the Bayesian Rational Analysis and Hierarchical Predictive Coding frameworks are combined.

11 We are grateful to Wanja Wiese for discussions which inspired this section about Bayesian just-so stories.

## References


Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14* (3), 471–485.

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*, 695–711.

Bechtel, W. & Richardson, R. C. (1993). *Discovering complexity. Decomposition and localization as strategies in scientific research.* Cambridge, MA: MIT Press.

Bowers, J. S. & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138* (3), 389–414. https://dx.doi.org/10.1037/a0026450.

Bruineberg, J., Kiverstein, J. & Rietveld, E. (forthcoming). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*. https://dx.doi.org/10.1007/s11229-016-1239-1.

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

——— (2017). How to knit your own Markov blanket: Resisting the second law with metamorphic minds. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Colombo, M. & Series, P. (2012). Bayes in the brain–On Bayesian modelling in neuroscience. *British Journal for the Philosophy of Science, 63* (3), 697–723.

Danks, D. (2008). Rational analyses, instrumentalism, and implementations. In N. Chater & M. Oaksford (Eds.) *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 59–75). Oxford: Oxford University Press.

Ernst, M. O. & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415* (6870), 429–433. http://dx.doi.org/10.1038/415429a.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 360* (1456), 815–836. https://dx.doi.org/10.1098/rstb.2005.1622.

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

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

Friston, K., Thornton, C. & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology, 3*.

Gershman, S. J. & Daw, N. D. (2012). Perception, action and utility: The tangled skein. In M. I. Rabinovich, K. J. Friston & P. Varona (Eds.) *Principles of brain dynamics: Global state interactions.* MIT Press.

Griffiths, T. L. & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17* (9), 767-773. https://dx.doi.org/10.1111/j.1467-9280.2006.01780.x. http://pss.sagepub.com/content/17/9/767.abstract.

Griffiths, T. L., Kemp, C. & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.) *The Cambridge handbook of computational cognitive modeling.* Cambridge, UK: Cambridge University Press.

Griffiths, T. L., Chater, N., Norris, D. & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on bowers and davis (2012). *Psychological Bulletin, 138* (3), 415–422. https://dx.doi.org/10.1037/a0026884.

Griffiths, T. L., Lieder, F. & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science, 7* (2), 217–229.

Harkness, D. L. & Keshava, A. (2017). Moving from the what to the how and where – Bayesian models and predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

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

———(2016). The self-evidencing brain. *Nous, 50* (2), 259–285.

——— (2017). How to entrain your evil demon. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34* (4), 169–188.

Klein, C. (in press). What do predictive coders want? *Synthese*. https://dx.doi.org/10.1007/s11229-016-1250-6.

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. https://dx.doi.org/10.1016/j.tins.2004.10.007.

Kwisthout, J. & van Rooij, I. (2013). Bridging the gap between theory and practice of approximate Bayesian inference. *Cognitive Systems Research,* (24), 2–8.

Körding, K. P. & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427* (6971), 244–247. http://dx.doi.org/10.1038/nature02169.

Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9* (11), 1432–1438. http://dx.doi.org/10.1038/nn1790.

Maloney, L.T. & Mamassian, P. (2009). Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience, 26*, 147–155.

Marr, D. (1982). *Vision: A computational approach.* San Francisco: Freeman.

McClamrock, R. (1991). Marr's three levels: A re-evaluation. *Minds and Machines, 1* (2), 185–196. https://dx.doi.org/10.1007/BF00361036.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics, 66* (3), 241–251.

Oaksford, M. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Sciences, 5* (8), 349–357.

Oaksford, M. & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* Oxford: Oxford University Press.

Piccinini, G. & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese, 183* (3), 283–311.

Rao, R. P. N. & 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. http://dx.doi.org/10.1038/4580.

Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117* (4), 1144–1167.

Schwartenbeck, P., FitzGerald, T., Dolan, R. J. & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology, 4* (710), 1–5.

Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science, 77* (4), 477–500.

Sims, A. (2017). The problems with prediction. The dark room problem and the scope dispute. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Spratling, M. W. (2013). Distinguishing theory from implementation in predictive coding accounts of brain function. *Behavioral and Brain Sciences*, 36 (3), 231–232.

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik*. Leipzig: Leopold Voss.

Xu, F. & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114* (2), 245–272.

Zednik, C. (2017). Mechanisms in cognitive science. In S. Glennan & P. Illari (Eds.) *The Routledge handbook of mechanisms and mechanical philosophy*. London: Routledge.

Zednik, C. & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese 193:* 3951. https://dx.doi.org/10.1007/s11229-016-1180-3.

# The Evidence of the Senses

## A Predictive Processing-Based Take on the Sellarsian Dilemma

Paweł Gładziejewski

Traditional foundationalist empiricist projects in epistemology postulated that sensory states of the subject are epistemically basic, in that they are capable of conferring justification on mental representations of the world without themselves needing to be (inferentially) justified by any antecedent representational states. This sort of view faces a seemingly hopeless dilemma, whose recognition is usually attributed to Wilfrid Sellars. If we treat sensory states as brute stimulations devoid of intentional content, then it is hard to see how the senses could provide subjects with anything that could possibly feature in justification-conferring relations with representational states. If we treat them as contentful, then in order to justify contentful states, sensory states themselves would presumably need to be justified by other representational states; but if this is so, they are not able to play a properly foundational epistemic role.

In the article, I use the Predictive Processing (PP) view of perception in order to sketch a possible resolution of the Sellarsian dilemma. I draw on PP in order to show how sensory states could actually serve a normative role that is recognizably similar to the one envisioned by traditional empiricists. To do this, I first distinguish representational from non-representational posits of PP and subsequently focus on the role that PP ascribes to sensory or "driving" signal. In particular, I argue that (1) the driving signal plays a role of a non-representational, contentless detector; at the same time, (2) it serves as an "impartial" or "theory-neutral" tribunal against which contentful internal models are actively tested and updated. Drawing on Anil Gupta's work, I discuss the epistemic involvement of the sensory signal in perceptual inference and show how the signal provides conditional justification (i.e. justification that is conditional on the justification or rationality of prior knowledge) to perceptual hypotheses. Then I discuss the role the sensory signal plays in perceptual learning. I employ the notion of "epistemic convergence" to sketch out how the sensory signal could provide perceivers with unconditional justification (i.e. one that is not relativized to the justification of prior knowledge). If this approach is right, the Sellarsian dilemma seems to be averted. We can see how the senses can be at the same time silent (i.e. contentless) and capable of playing a sort of foundational epistemic role.

**Keywords**

Anil Gupta | Epistemic justification | Immanuel Kant | Kantian receptive sensibility | Perceptual inference | Perceptual justification | Predictive processing | Sellarsian dilemma

**Acknowledgments**

Work on this paper was supported by the Polish National Science Centre FUGA 3 grant (UMO-2014/12/S/HS1/00343). I thank Michael Anderson, Jakob Hohwy and other participants of the Philosophy of Predictive Processing workshop at the Frankfurt Institute for Advanced Studies for their comments and questions. I am especially indebted to Jona Vance, Krys Dołęga, Thomas Metzinger and Wanja Wiese for their critical assessments of the earlier versions of the paper. When working on the paper, I also benefited from conversations with Marcin Miłkowski, Paweł Grabarczyk and Przemysław Nowakowski.

### 1 Introduction

One of the philosophically important aspects of the predictive processing framework (PP) is its postulate that our perceptual contact with the world is active and interpretative in nature. According to PP, perception relies on an internal generative model which estimates the states of the environment and on this basis predicts, in a top-down manner, the sensory states of the perceiver. The emphasis on the active role played in perception by the perceiver's cognitive apparatus and prior beliefs is by no means historically new. But at the very least it constitutes a notable divergence from a view that has, in recent decades, been quite popular in naturalistically oriented philosophy of mind — a view on which perception (and perhaps representation more generally) is largely a matter of *detecting* states of affairs by being causally attuned to them (see Ramsey 2007, Chapter 4 for a discussion of the "receptor" notion of representation and its prevalence in philosophy and cognitive science).

Gładziejewski, P. (2017). The Evidence of the Senses - A Predictive Processing-Based Take on the Sellarsian Dilemma.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 15. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573161

It is not surprising, then, that the view of perception as active in this sense caught the attention of philosophers (Clark 2013a; Clark 2013b; Clark 2016; Hohwy 2013). But perception is not *just* active according to PP. In addition to postulating the internal probabilistic inference that generates top-down predictions, the theory also attributes a crucial role to the sensory signal which results from the world's interaction with the perceiver's sensorium. It is through confronting the predicted with *actual* sensory states that the prediction error signal is generated and propagated bottom-up, which in turn enables the model to adjust its estimates. So despite having a clearly constructive flavor, PP rather views perception as resulting from the interplay between — to use Immanuel Kant's terminology — mind's spontaneity (i.e. the endogenous, prior-belief-based activity of the generative model) and receptive sensibility (i.e. sensory signal caused by the environment). In the present paper, I want to investigate more closely this latter, "receptive" aspect of PP, with emphasis on epistemological considerations.

Notably, traditional empiricist projects in philosophy used to attribute to the sensory states a foundational role. The sensory states were to act as epistemological "unmoved movers", capable of conferring justification or warrant on the subject's representational states, without needing a separate source of justification or warrant for themselves (BonJour 1985; Chisholm 1977). This empiricist-foundationalist project received deep and influential criticisms in the latter half of 20th century, to the point where it became often regarded as conclusively refuted (Davidson 1973; Quine 1969; Sellars 1956). One way to see what is wrong with this project is to consider a seemingly hopeless dilemma whose recognition is sometimes attributed to Wilfried Sellars (Sellars 1956). If we treat sensory states as devoid of intentional content, then it is hard to see how they could possibly feature in inferential, justification-conferring relations with representational states. But if we treat sensory states as contentful, then in order to play a justificational role, they would presumably require to be justified by other, antecedent representational states. However, if this is so, then sensory states are not able to play a properly *foundational* epistemic role.

The aim of this paper is to revisit the Sellarsian dilemma with the PP account of perception in view. I will argue that although the sensory signal that PP appeals to is best construed as nonrepresentational, it nonetheless plays a crucial epistemic role. In particular, it serves as an "impartial" tribunal against which contentful internal models are actively tested and updated. This testing-and-updating is a probabilistically rational process, usefully interpreted as a form of abductive inference. If this is right, then we can see how the senses can be at the same time "silent" (i.e. contentless) and capable of playing a crucial role in providing justification[1] for representational states. The resulting view is distinct from classic empiricist foundationalism in many important respects, but it preserves the basic empiricist notion of "the evidence of the senses".

The discussion is structured as follows. In section 2, I discuss the Sellarsian dilemma in more detail. Next, in section 3, I provide a philosophical interpretation of PP which distinguishes representational from nonrepresentational posits of the theory. There, I also argue that the sensory signal belongs to the latter category. In section 4, I explain how, despite being nonrepresentational, the sensory signal still plays a major epistemic role according to PP, a role which is not completely unlike the role played by sensory "given" in traditional empiricism. In particular, I examine the epistemic involvement of the sensory signal in perceptual inference and perceptual learning.

1 Throughout this paper, I use the notions of "justification" and "warrant" to signify positive epistemic status very generally understood. For a representational state to be epistemically justified in this sense, it need not be produced by a subject that has a conscious or reflective grasp of the justification in question, or by way of conscious, intentionally controlled inference (for a similar approach, see e.g. the notion of "epistemic entitlement" as introduced in Burge 2003). On this encompassing approach, states that are formed through unconscious inferential processes — such as those postulated in PP — can count as epistemically evaluable.

## 2 Senses, their (Purported) Normative Role, and the Sellarsian Dilemma

It is easy to see how the senses could causally shape our representations of the world. On the one hand, the senses establish a world-mind causal link, as they are reliably affected by external stimuli. In fact, Kant is thought to have generalized this point into a claim that the very nature of sensibility lies in its constituting a passive capacity to be affected by things (Langton 1998; I will turn to this notion in section 4). On the other hand, the senses also establish a mind-mind causal link in that the (externally generated) sensory states are causally involved in producing subject's intentional states, such as perceptual beliefs.

It is philosophically crucial to distinguish this causal role of sensory states from the *normative* role that they play, at least according to classical, foundationalist versions of empiricism. This latter role is connected to a specific place that the senses supposedly occupy in the structure of justification or warrant of contentful states (BonJour 1985). Broadly speaking, the point is that the sensory contact with the world provides subjects with *reasons* for holding a belief or even a specific conception of the world (be it folk, scientific, or both). And these reasons are supposed to be epistemologically basic. Although many of our beliefs may acquire their justification by virtue of their inferential relations to other justified beliefs, this inferential chain ends with the deliverances of the senses. Through the sensory contact with the world, the subjects are supposed to be "immediately apprehending" states of affairs, or be "directly aware" of them, or be "acquainted" with them, or have those states of affairs "given". Whatever the terminology, these sensory acquaintances can transmit justification or warrant to intentional states, but themselves do not require to be justified or warranted by being inferentially related to other contentful states.

Since serving as a reason for an intentional state is not the same as serving as a cause of that state, it seems impossible to account for the normative, justification-generating role of the senses by simply pointing to their causal role. In fact, as already mentioned, the very claim that we can even *make sense* of the normative role of sensory states faces the "Sellarsian" dilemma (Sellars 1956; see also BonJour 1985; Lyons 2008). Sensory states can be either taken as devoid of intentional content, or they can be treated as states that are contentful themselves. Both options seem to render them unfit to play their postulated normative role.

Accepting the first option amounts to accepting the claim that sensory states are contentless. They do not represent anything as being a certain way; rather, they constitute brute sensory stimulations. But brute sensory stimulations do not seem like things that could justify or warrant intentional states (Davidson 1986). In particular, they cannot enter inferential relations with intentional states. Because of this, sensory states, nonintentionally construed, are epistemically inert.

If we rather go with the second option and treat sensory states as contentful, then another problem arises. For any sensory state which asserts some content (i.e. represents the world as being in a particular way), we may legitimately ask about what justifies this state having (asserting) this particular content and not some other. After all, we presumably do not want this content to be arbitrarily asserted. Thus, if sensory states are contentful, then perhaps they can act as reasons, but they will also *require* reasons for themselves.[2] However, if this is the case, then the chain of inferential justification cannot end at the senses, hence, the latter cannot play a *foundational* epistemic role. Perhaps sensory representations do obtain their justification inferentially, from other, antecedently held representational states. But if so, then they are not "theory-neutral". Their justification rests on conceptions and beliefs

2 As one of the reviewers of this paper points out, an epistemologist might accept that sensory states are contentful (i.e. assessable for accuracy) and yet deny that they are rationally evaluable (i.e. assessable for their epistemic status). But note that although this is *some* way out of the Sellarsian dilemma, it does not seem to take us far. At the heart of the dilemma is the worry about how sensory states could provide reasons for beliefs. If a state is contentful but not rationally evaluable, then — as I understand it — we can establish what its conditions of accuracy are but not whether this state is justified or not (or whether the subject is justified or not in being in the state with that content). If so, it is not easy to see how the sensory state in question could generate a positive epistemic status for other representational states, i.e. play a justificational role. What we get is a contentful, but epistemically intert state.

which are merely presupposed (if we want to treat those background presuppositions as *empirically* justified, then the problem reappears, and so a regress looms). What the senses deliver is not bare, inferentially uncontaminated "Given".

One possible reaction to the Sellarsian dilemma might be to claim that it rests on an overly restrictive, internalist and doxastic conception of epistemic justification, one captured in the dictum that "nothing can count as a reason for holding a belief except another belief" (Davidson 1986, p. 310). To pose the dilemma, one needs to assume that justification or warrant is a matter of inferential relations between intentional states of the subject. But perhaps the whole ordeal can be easily averted if we extend our notion of justification so that it counts, as possible justifiers, extra-mental and non-intentional factors, e.g. causal processes that reliably produce true or accurate representations (see Lyons 2008). With such a view in hand, it might be argued that the basic sensory representations do not require *inferential* justification; them having an appropriate, reliable *causal* history will suffice.

It is not my intention here to take stance on the internalism-externalism debate about epistemic justification. However, I do think that the Sellarsian dilemma *is* of relevance at least in the context of PP with its basic theoretical commitments. First, according to PP, perceivers estimate the states of the environment using only the resources that do not go beyond the statistical patterns that arise at their sensorium (Clark 2013b; Clark 2016; Hohwy 2013). In other words, perceivers update their model of the environment based solely on evidence that is internal to their mental lives. Second, PP construes this model updating in terms of unconscious probabilistic *inference* that employs prior ("presupposed") representations of the causal structure of the environment. So the epistemology of perception inherent in PP seems distinctly internalist[3] and inferentialist. And these ideas are precisely the philosophical incubators of the Sellarsian dilemma. Indeed, if the present paper is on the right track, then it turns out that PP enables us to tackle the Sellarsian dilemma largely *on its own epistemological terms*, without relying on the externalist view of justification.

## 3 Representational and Nonrepresentational Posits of Predictive Processing

One of the philosophical discussions sparked by PP centers around the question of whether PP is best seen as representationalist, and if so, then what sort of internal representations it is committed to, exactly (Hohwy 2013; Hohwy forthcoming; Clark 2016; Gładziejewski 2016; Orlandi forthcoming). However, we need to be cautious to notice that it might be too simplistic to construe the debate as a competition between thoroughly representationalist and thoroughly antirepresentationalist readings of PP. To explain perception, PP postulates a complex processing scheme engaged in minimizing the prediction error. It is entirely possible that this scheme includes both representational and nonrepresentational aspects or parts. If so, then in order to get a complete philosophical understanding of PP, we need to carefully examine which of its theoretical posits belong to which category.

Following William Ramsey (2007), I take it that any genuinely representational theoretical posit of a cognitive theory owes its status to its meeting an appropriate "job description", i.e. to the function that it plays in a cognitive or computational economy of a larger system. In other words, a theoretical posit counts as truly representational if it can be shown to *serve* as a representation in some nontrivial way or sense.

Now, we may take PP to be committed to at least following four posits: (1) the sensory signal which results from the world affecting the sensory apparatus of an organism, (2) the (hierarchical) generative model that sends top-down sensory predictions or "mock" sensory signals, (3) the prediction error signal which is propagated bottom-up and signifies the divergence between the predicted and actual sensory signal, (4) precision estimators which regulate the gain on prediction error signal. For each

3 Because these epistemically relevant mental factors may be not consciously or reflexively accessible for the subject, the brand of epistemic internalism at play here is so-called "mentalism" rather that access internalism (see Conee and Feldman 2001).

of those posits, we may ask whether its functioning merits a representational reading. Here, I want to raise this question for (1) and (2) in particular, as they are directly relevant to the present discussion.

Let me start with posit (2), i.e. the generative model. There are strong reasons to regard it as playing a nontrivially representational role. First, it generates, in perceptual inference, estimates of the environment which guide the cognitive system's practical engagements with the environment (in active inference). Second, the model's ability to play this action-guiding function is dependent on how well its structure captures or resembles the causal structure — hierarchically nested at different timescales — of the environment. Third, the model does not simply passively register external states of affairs, but rather performs a sort of, largely endogenously-controlled, predictive simulation. It displays at least some degree of detachment or independence from current states of the world; in fact, a case could be even made that it can be used purely off-line, i.e. outside of any direct engagements with the environment. Fourth, insofar as the model undergoes correction in light of prediction error signals, it can be said to be capable of detecting cases when its estimates or hypotheses are inaccurate. The upshot, then, is that the generative model constitutes an action-guiding, detachable structural representation, capable of detecting representational error. The way this representation functions is not unlike the way that familiar and noncontroversial examples of external representational artifacts, like cartographic maps, function (for a much more in-depth exposition of this view, see Gładziejewski 2016).

If the generative model is in the business of representing the causal structure of the environment, then, according to PP, this representation is formed on the basis of the trace that the causal structure in question leaves on the outer boundary of the central nervous system. The job of the internal model is to recover the causal furnishing of the world using its sensory effects, i.e. the sensory signal. The question, now, is whether the sensory signal itself can be reasonably construed as representational.

Some might feel inclined to defend a representational reading of the sensory signal by pointing to the role of the senses as *detectors*. A detector is a structure that causally co-varies with certain states of affairs, and its function is to do so. In particular, detector's function is to reliably react to some conditions and in turn initiate certain reactions or downstream effects. Unsurprisingly, the sensory signal can be regarded as detecting the presence of environmental conditions. Sensory signals result from a reliable causal covariance between the sensory apparatus and the presence of proximal physical stimuli. And given the larger processing scheme that they participate in (according to PP), it is the function of the senses to be reliably affected by the stimuli. In particular, they give rise to prediction error signal, as unpredicted or "unexplained" aspects of the sensory signal get propagated bottom-up, eventually prompting the generative model to revise its estimates of the environment, if needs be.

However, it is far from clear how functioning as a detector could be by itself *sufficient* to regard something as representational (Orlandi forthcoming; Ramsey 2007). This is because any causal mediator meets the conditions of serving as a detector. On such a view, a light switch, a firing pin in a pistol, or the reactions of an immune system to bodily injury would have to count as representational. Claiming that the senses "represent" anything in this extremely liberal sense would amount to making an uninteresting claim that they causally mediate between what happens in the world and what happens in the perceiver (i.e. that they are transducers). In other words, attributing the sensory signal with a role of representing something simply because it detects proximal conditions would mean subscribing to a hopelessly trivial notion of representation.

The case for a representational reading of the driving signal weakens even further when we notice that it lacks other functional properties that we often attribute to representations (see also Orlandi forthcoming). First, the signal is not directly involved in guiding the actions (active inferences) of the cognitive system. What guides active inference is not the sensory input itself, but rather the interpretation of the input provided by the generative model in the form of perceptual hypotheses. Second, signals generated in the sensorium necessarily track current proximal stimuli; they only react to what is actually present. They do not entertain any sort of independence from their physical causes. As such, the sensory signal is incapable of acting in a way that is even minimally detached or off-line.

Given how there is no representation without the possibility of representational error, it is also natural to consider the question of whether we could make sense of the claim that the incoming sensory signals are capable of *misrepresentation*. Admittedly, it makes perfect sense within the PP framework to say that the senses can sometimes mislead or misguide the perceiver, say when she tries to find her way in a deep fog. These are the cases where the sensory signal is noisy — in the sense of having low precision or inverse variance — to the point where it tends to lead perceptual inference astray. But it does not seem quite right to count such cases as examples of the *senses* somehow failing at performing their duties, let alone misrepresenting anything. The crucial point to note here is illuminatingly expressed by Anil Gupta:

> If, during a walk in a forest, I bump my head on a low branch of a tree, it is better that I assume responsibility (and change my ways) than that I pin the blame on the tree. The tree is passive. It is bound to be the way it is, given the circumstances, and it is useless to blame it for my sore head. Similarly, if, having suffered an experience, I acquire a false perceptual belief, it is better that I assume responsibility (and change my manner of "reading" experience) than that I pin the blame on the experience. The experience is bound to be the way it is, given the circumstances, and it is useless to blame it for my false belief. (Gupta 2006, pp. 28–29)

This idea can be in the following way translated into more subpersonal notions that PP trades in. When we qualify noisy sensory signal as misleading, unreliable or ambiguous, it is not due to the fact that the sensory apparatus somehow fails to represent the world correctly. Photoreceptors in the retina do not fail to fulfill their function, representational or otherwise, when we are taking a stroll on a foggy day. They are "bound" to generate a noisy reaction, given the circumstances, and it is useless to "blame" them, i.e. treat as malfunctioning[4]. Rather, the qualification of the sensory signal as misleading is parasitic on the use that the inferential machinery, i.e. the generative model, makes of the signal.[5] It is the model's job to provide hypotheses about determinate causal etiology of the incoming signal. The noisy sensory signal itself is misleading only in the sense that it can make the generative model more prone to come up with inaccurate hypotheses about the signal's causal origins, and hence less effective at minimizing the prediction error. It is not the signal itself that misrepresent the world, but the model that "misreads" the signal and misrepresents its distal causal etiology.

These considerations present a case to think that the explanatory repository of PP involves both representational and nonrepresentational posits. In particular, while the generative model can be treated as playing a genuinely representational role, the same cannot be said about the sensory signal. What the *senses* deliver, then, is *not* representations. This echoes Immanuel Kant's claim that "although it is correct to say that the senses do not err, this is not so because they always judge correctly but because they do not judge at all" (Kant 1781/1996, p. 128).

## 4 The Sellarsian Dilemma in Light of Predictive Processing

### 4.1 Epistemic Role of the Senses According to Predictive Processing: an Outline

The foregoing discussion suggests that if we are to tackle the Sellarsian dilemma using PP's conceptual resources, we will have to proceed by accepting the claim that the sensory signal can somehow play a sort of foundational epistemic role despite the fact that it is not contentful. However, if we assume,

4 This is *not* to say that the sensory apparatus cannot malfunction in cases where the physiology of the sensory organ itself is damaged or otherwise changed. However, this sort of malfunctioning is *not* an instance of *misrepresentation* (assuming that after the damage or modification, the sensory organ still functions as a causal mediator, and so does not have representational content).

5 A remark is in order here. When an inaccurate representation is formed based on a noisy signal, there is usually not one, but two culprits: the model which misestimates the state of the environment and the precision estimators which fail to lower (to a sufficient degree) the gain on prediction error signals.

rather uncontroversially, that inferences are transitions between *representational* states which accord to some normative epistemic rule, then contentless sensory signals turn out incapable of acting as contentful *premises* from which the latter could be inferentially derived. The question, then, is how to characterize the epistemic or rational, and recognizably "foundational" bearing that the sensory signal has — assuming that it has — on perceiver's internal representations?

To avoid confusion, we need to respect the distinction between sensation and perception. In PP, sensation corresponds to the contentless sensory signal.[6] Perception, construed in terms of perceptual inference, corresponds to the forming of a representation of the sensory signal's distal causal origins. One might wonder whether we should rather go with the second option of the dilemma and treat the perceptual representations as epistemically basic. But it seems that perceptual representations or hypotheses, as PP construes them, do not even purport to be epistemically basic. According to PP, cognitive systems form perceptual hypotheses by engaging in approximate Bayesian inference which relies on preexisting probabilistic representations of (or prior beliefs about) the causal structure of the environment (see section 4.2). Thus, perception rests on presuppositions and is inherently theory-laden in this sense. What cognitive systems perceive is inferentially shaped by what they already represent the world to be like.[7]

But this is not the whole story about how perceptual representations are formed in PP. Insofar as perception is prediction error minimization, updating perceptual representations relies not *only* on preexisting beliefs but also depends on the incoming sensory signal. After all, the prediction error is estimated and minimized relative to actual sensory input. Of course, we need to be careful to distinguish between sensory signal's causal involvement in forming representations and its epistemically normative involvement. The present point is that the sensory signal can be attributed with not only a causal but also a normative, epistemic role.

Remember the Kantian lesson (which also resonates in the passage from Gupta cited in the previous section) that our sensory contact with the world is *receptive* in the sense that it constitutes a passive capacity to be affected by things; the raw sensory manifold is simply received rather than endogenously constructed through the spontaneous activity of the perceiver (Langton 1998). The sensory input, as construed by PP, *is* receptive in this sense. The states of the sensory apparatus are treated as a function of their hidden worldly causes (Friston 2010; Friston and Kiebel 2009). What happens at the perceiver's sensory boundary is outside her jurisdiction, in the sense of being contingent solely on the causal structure of the world (which includes the perceiver's own body), and independent from the perceiver's prior beliefs or inferential activity. Metaphorically, the senses constitute a point at which inference and top-down prediction bottoms out, and the system becomes purely responsive to external factors.

6 There is a way in which the present PP-based construal of sensory states diverges from epistemological tradition. Traditionally, the purportedly foundational sensory states were assumed to have a qualitative or phenomenal character. They were often seen as raw conscious "feels". Here, they are treated as a bottom-up signal that the results from the world causally affecting the sensory apparatus of the perceiver. I do not assume that sensory states thus understood have to determine or be manifest in the phenomenal perceptual experience of the subject, even if they shape it in major ways. In the context of PP, it seems more reasonable to claim that contents of consciousness correspond to generative-model-derived perceptual hypotheses which already populate the world with familiar objects, properties and relations (see the discussion of sensation/perception distinction in main text). The analogy between "sensation" in the present sense and the way this notion was used in more traditional approaches *only* pertains to the epistemic role (see the following discussion in main text).

7 As has been already noted in the literature (Hohwy 2013; Lupyan 2015), the issue of cognitive penetrability of perception naturally crops up in the context of PP. Cognitive penetrability is sometimes construed in terms of, roughly, high-level cognitive states (which may correspond to personal-level beliefs or desires) directly affecting the content of perceptual representations. Cognitive penetrability thus understood is a local phenomenon, as many percepts seem *not* to be penetrable in this sense. For example, in the hollow mask illusion, the mask persists to be perceptually represented as convex even in light of knowledge that it is actually concave. But notice that PP commits us to the view that perception is inferential all the way down. Perceiving may be sometimes independent of representations stored at *higher* levels of the generative model, but it still rests on *some* assumptions, presumably stored at lower levels of the processing hierarchy. For example, on PP view of things, the supposedly convex shape of the hollow mask is inferred from preexisting premises that are low-level and beyond the conscious control of the subject. When I talk about perceptual representations being theory-laden in PP, I mean that they are strongly or globally cognitively penetrable in this sense. Notice that this strong penetrability of perception is presumably what we should care about in the context of foundationalist empiricism. When epistemologists seek properly epistemically basic representations, they do not mean representations whose justification does not depend on a particular subclass of intentional states (say, personal-level beliefs about the truth of particular scientific theories; see Fodor 1984), but ones whose justification does not depend on (inferential relations to) any antecedent intentional states.

By being passive in this sense, the sensory signal is anything but theory-laden. Rather, it is "pure" or "impartial", as it only depends on what the worldly causes are, not on what they are represented to be.[8]

The claim about the receptivity of the sensory signal requires qualifications. There is a sense in which the sensory signal *does* depend on the activity of the perceiver and her representations of the environment. For one thing, perceptual inference is intertwined with active inference, whereby the perceiver actively samples the environment in order to *make* the sensory input fit the perceptual hypotheses. Perceivers intervene in the world with their bodies to cause it to conform with their sensory predictions. Also, through estimating precision of the sensory input, the perceiver has the ability to regulate the weight of the prediction error signal, and hence the degree to which it affects perceptual inference. For example, in a deep fog, the perceiver may "decide" to largely ignore the noisy visual input. However, on closer inspection, neither of those points is inconsistent with the claim about the essentially receptive nature of the sensory signal. *Once* the perceiver decided to actively sample the environment in a particular way (say, by performing a particular series of saccades) and *once* the precision of the input is estimated, it is no longer up to the perceiver what will actually happen in her sensorium; it depends on the external causes. Whether active inference actually succeeds at making incoming signal conform to the signal predicted requires "cooperation" on the part of the external environment. Similarly, although the perceiver may lower the degree to which the sensory input affects hypothesis revision, it does not thereby change the sensory signal itself; again, it depends on the world.

The claim about the receptivity of the senses relates not only to how they are causally situated, but it has a significant epistemological import as well. To get a general gist of how the epistemic role of the sensory signal is to be understood, imagine a cognitive system which lacks anything that fulfills this role. That is, imagine a system which gets sensorily disconnected from the world — say, we disable its sensory system — but still uses an internal generative model to get around. The system samples from its model (construed here as a Bayesian network) and thereby performs an internal predictive simulation of the causal processes in the world, on the basis of which it decides how to act. The simulation in question relies *only* on prior and likelihood probability distributions already encoded in the model. No sensory input is received, so no prediction error is computed. Our system freely "dreams" reality (see Hobson and Friston 2012; Bucci and Grasso 2017),[9] rather than properly perceives it.

Even if we assume that the hypothesis (or a set of hypotheses) which initiates internal sampling from the model actually corresponds to some initial state of the environment, given the complexity and unpredictability of the latter, it is natural to expect that the internal simulation will at some point diverge from actual states of the world, eventually leading to potentially catastrophic results. Without the sensory input with which the internal predictions are confronted, the process of revising representations lacks some sort of *external constraint* on the space of currently relevant hypotheses, a constraint which makes the process in question responsive to what is actually going on in the environment. When the sensory constraint or guide of this sort is not present, the internal simulation is like, to use John McDowell's (1994, p. 11) poetic phrase, "frictionless spinning in the void". And this shows what the epistemic role of the senses consists in. As active and constructive as perception is according to PP, if it is to put us in touch with the world, there needs to be a point where it meets resistance or friction and becomes answerable to some external authority. This is precisely the job of the sensory signal with its passive or receptive nature. Perceivers minimize the prediction error which reliably depends on both internally-generated predictions and actual sensory input; but the latter ultimately

8 This claim about "purity" of the senses needs to be qualified. The physical/physiological makeup of sensory transducers themselves is, of course, a product of evolutionary forces. The senses are selective with respect to the physical energies they react to, in a way that expresses general "assumptions" what the organism's Umwelt is (roughly, about what in the environment is biologically useful or salient). It is not my intention here to treat the senses a providing some kind of "view from nowhere". The senses count as passive receivers, but what they receive (and how they receive it) is determined by the evolutionary history. I am indebted to Thomas Metzinger for pushing me on this.

9 As one reviewer points out, this analogy with dreaming is limited in two important ways. First, certain kinds of sensory signals (e.g. proprioceptive and interoceptive) are actually present during sleep. Second, the content of perceptual-like states in dreaming is less detailed than the content of actual perceptual states.

depends on worldly causes. This way, through in the involvement of the sensory input, perception becomes supervised by the world itself (Hohwy 2013).[10]

Crucially, for the senses to play this role, they do not have to supply perceivers with epistemically basic *representations*. The sensory apparatus does not provide prediction-error-minimizing systems with a set of ultimate, non-inferentially justified premises, but rather with raw, uninterpreted data against which sensory predictions are tested. The system can then use the result of this testing — the prediction error signal — to correct its internal representation. The causal structure that generates the sensory input is *indirectly* recovered by adjusting the model to optimize prediction error minimization. It is not simply inferred from a set of ready-made, pristine premises, "direct acquaintances", or anything of that sort.

This provides us with a general understating of what the "evidence of the senses" amounts to, according to PP. Now I want to fill this general story with some details, by discussing the epistemic role that the sensory signal plays in perceptual inference (section 4.2) and perceptual learning (section 4.3).

## 4.2 Sensory Signal and the Epistemology of Perceptual Inference

In a way, epistemological considerations lie at the very heart of PP. Despite being naturalistic through and through, the theory views perception as a process that conforms to *normative* principles. This means that, on PP view of things, organisms in fact (tend to) form perceptual representations that they *ought* to form *were* they to follow a rational norm (Hohwy 2013; Hohwy et al. 2008; see also a related discussion in Rescorla 2016). One way to understand the norm in question is in terms of causal abduction (inference to the best explanation), whereby percepts are hypotheses that "best explain" the sensory input by citing its worldly cause(s) as an explanans (Hohwy 2014; Seth 2015). As already mentioned, in PP, this general idea is cashed out in Bayesian terms. The assumption is that perceivers update (infer) perceptual hypotheses in a way that aims at maximizing posterior probability. This process is only approximately Bayesian, as PP is not committed to the claim that brain literally works by implementing Bayes' rule (Hohwy 2013). Rather, the point is that the process of prediction error minimization with the use of generative models approximates Bayesian inference. A system updating its generative model to minimize prediction error is a system that updates its internal estimates of the environment in a way that conforms with Bayes' rule. The upshot, then, is that PP construes perception as a form of probabilistically rational abduction.[11]

Given that perceptual inference is rational in this sense, how are we to understand the epistemic involvement of sensory signal in this process? The preceding section already characterized this role in broad strokes. To a first approximation, sensory input constrains "from the outside" the updating of hypotheses in a way that enables them to actually track environmental states of affairs. Now I want to

10 The claim about the world acting as a supervisor does not imply that the notion of justification at use here is externalist after all. The process of internal model formation (and perceptual hypothesis formation) is truth-conductive in the sense that if the sensory signal one receives is actually causally generated by the familiar physical world, then one will likely form — on the basis of the input — a more or less accurate representation of the world. This does not have to imply that an actual reliable causal connection to the world acts as a justifier in this story. To understand this point, consider a PP-inspired variation on the so-called "new evil demon" scenario (Cohen 1984). Imagine your epistemic copy that, through life, receives a series of sensory signals which are exactly like the signals that you receive. Imagine that this copy performs a series of (subpersonal) probabilistic inferences to construct a model of the worldly causes of its sensory states — again, the copy proceeds exactly like you do (or your brain does). The copy ends up with a model of the external world which is exactly like yours. Given the same sensory signals, it will form the same perceptual hypotheses. Now imagine that this copy is actually fed its sensory signals by an evil, misleading demon, such that the copy has no reliable causal connection to a real world. The model of the world harbored by your copy is systematically false. But it is constructed by way of the same Bayesian inferences operating on the same data as is the case with you. So it seems that your copy is, given all available sensory evidence, as rationally entitled to/justified in forming the model of the environment as you are.

11 I take it that this inferential aspect is inherent in Bayesian accounts of perception, including PP. It is hard to see how perception can be truly Bayesian without constituting Bayesian inference (see Hohwy forthcoming; Kiefer 2017; Rescorla 2015). But it needs to be acknowledged that some authors make attempts to reconstrue PP in a way that aims to avoid this inferential commitment (see Orlandi forthcoming). It is beyond the scope of the present paper to discuss how this anti-interential reconstrual could affect the issues about the epistemology of perception in the context of PP.

propose how we can enrich this story by linking it to some ideas found in Anil Gupta's "Empiricism and Experience" (Gupta 2006).

Gupta's aim is to understand the "rational contribution" that conscious experience of the subject makes to knowledge. Right now I undergo a certain visual experience which, intuitively speaking, makes *reasonable* my perceptual belief that there is a computer screen in front of me. Gupta coins this rational contribution of my experience to my belief as "the given". On his account, the given in any experience is not, so to speak, epistemically autonomous. In his words, every experience is "multiply factorizable" in the sense that "no experience carries with it its own genealogy" (Gupta 2006, p. 7). For example, an experience that one has when looking at a bright green wall so large that it occupies one's whole visual field could have been obtained by looking at a white wall through bright green glasses, or at a blue wall through yellow glasses. Experience *alone* cannot decide between these options. This gives rise to the claim that the given in experience is "hypothetical" in nature. That which judgment is rational to hold on the basis of a particular experience is conditional on the background "view" that one brings to bear on that experience. The view consists of "concepts, conceptions and beliefs" that perceiver is already in possession of. For example, if I antecedently believe myself to be placed in a large building, and I am in the possession of the concept wall, and my view does not include the belief that I am wearing yellow glasses, etc., then, given a particular visual experience, I can reasonably form the perceptual belief that there is a bright green wall in front of me. A person with a different view could well be entitled to a different judgment, given the same experience. This leads Gupta to conclude that the "logical category" of the given in a particular experience $e$ is that of a *function* $\Gamma e$, which takes the view $v$ as the input and yields (classes of) perceptual judgments $\Gamma e(v)$ as outputs. The resulting position can be summarized with the following schema:

View $v \Rightarrow$ (Experience $e \Rightarrow$ Perceptual judgments $\Gamma e(v)$)

On this view, then, experience provides the perceiver not with an absolute, but only with conditional entitlement to her perceptual judgment(s); one could say that perceptual judgments are "rationally underdetermined" by experience. The rationality (justification or warrant) of the perceptual judgment is conditional on the rationality (justification or warrant) of the view that one brings with herself. Importantly, the *whole* positive nature of the given in experience consists in how it provides rational guidance in forming perceptual judgments. In particular, although the given in experience outputs judgments, it is not itself contentful or "propositional".[12] Taken apart from the view, the experience as such is "silent", rather than composed of basic propositions regarding sense-data, subjective mental states or medium-sized physical objects.

Gupta construes his account as concerned with strictly logical or normative matters, and distances it from naturalism, claiming that "the project of constructing a naturalistic account of rationality and of perceptual judgment is, at the present stage of inquiry, nothing but quixotic" (Gupta 2006, p. 54). Nonetheless, against the author's intentions, I want to suggest that his treatment of "the given" proves strikingly useful as a framework for understanding the epistemic role that the sensory signal plays in perceptual inference according to the naturalistic theory that is PP.

For starters, we may observe that basic categories present in Gupta's account have analogues in PP's story. Namely, we may suppose that (1) the raw sensory signal corresponds to Gupta's "experience"[13]; (2) the generative model that encodes the causal structure of the environment corresponds to "the

12 When Gupta 2006 characterizes the given as nonpropositional, the claim is not (only) that it lacks "conceptual" or "propositional" type of content, but that it lacks content *altogether*.

13 Importantly, the analogy here is supposed to pertain only to the similarity at the level of the type of normative role that the analogues play. In Gupta's view, the given is manifest in conscious experience (Gupta 2006, p. 30). But if my proposal is right, this particular point of Gupta's view is not preserved in PP. To repeat, it is not my suggestion that conscious perceptual *experience* is determined by or somehow corresponds to what happens at the level the sensory input.

view"; (3) perceptual hypotheses (estimates) generated by the model to minimize the prediction error relative to the sensory input correspond to perceptual judgments entailed or yielded by the experience when combined with the view. Abstractly, we might treat the perceptual representation or hypothesis as an output of a function $\Gamma_s$ (where *s* stands for sensory signal) that takes the generative model (i.e. a particular prior and likelihood distribution) *m* as input. We may then reformulate Gupta's schema as follows:

Generative model $m \Rightarrow$ (Sensory signal $s$ Perceptual hypothesis(es) $\Gamma s(m)$)

Paralleling Gupta's account, the transition from the combination of the incoming signal and the preexisting model of the environment to the perceptual hypothesis is rational. Namely, it is rational, as it (approximately) conforms to Bayes' rule; and Bayesian inference is by its nature truth-conductive: if you follow the rule, you produce hypotheses with the largest posterior probability. Also, on this picture, the normative contribution of the sensory signal consists entirely in the guidance that it provides for the generative model. Apart from the generative model that interprets it, the sensory input is silent (contentless). This is PP's incarnation of Gupta's claim that the given in experience is "nonpropositional". Furthermore, the guidance that the senses provide for the model is essentially corrective in nature. Its epistemic value lies in the fact that in enables the cognitive system to be responsive to cases in which what actually happens in the environment diverges from what is already "contained" (encoded) in, and hence expected on the basis of, the model alone. This, after all, is the lesson to be drawn from the dreamer-perceiver example given earlier.

Notice that on such a view the epistemic role the sensory signal does *not* consist in the signal *transferring* its epistemic status to the perceptual hypothesis. The signal as such is not rationally evaluable. Its rational involvement consists solely in how it is essential (in combination with the internal model) in enabling perceptual inference to be truth-conductive in virtue of approximately embodying Bayesian rationality. Without the signal, the rational link between priors and the perceptual hypothesis is broken. That is, were perceptual inference not guided by the senses, it would fail to be Bayes-rational.[14]

Crucially, this picture parallels Gupta's proposal in that, on PP view of things, perceivers obtain only conditional epistemic entitlement (justification or warrant) to their perceptual hypotheses. Imagine two perceivers facing the same ambiguous stimulus, like the duck-rabbit. If one of them ascribes a significantly larger prior probability to ducks than to rabbits, and the other vice versa, they will (all else things being equal) end up forming different perceptual representations. But both representations can be considered equally probabilistically rational in a conditional sense. In each case, the hypothesis is formed which has the largest posterior probability in light of the probabilistic generative model used to interpret the sensory input. Both perceivers can be said to have conditional entitlement to their perceptual hypotheses — they are entitled to their perceptual representations *if* they are right about their respective priors.

## 4.3 Sensory Signal and the Epistemology of Perceptual Learning

Can perception, as PP construes it, give rise to justification that is unconditional or not relativized to a particular set of background "beliefs" encoded in the generative model? This is a subtle issue, and it

14 Here, it is important not to overplay the causality/normativity distinction. Although the distinction needs to be drawn to avoid basic philosophical confusions, the point is *not* that matters of causality are completely disconnected from normative matters. The transition from the signal-model combo to a perceptual hypothesis is rational in virtue of its conforming to a rational, truth-conductive principle. But in PP, the move form the signal-model combo to the hypothesis is, of course, *causally realized*. Furthermore, one could argue that the rationality of token perceptual inferences lies precisely in the fact that their causal profiles track or accord to a rule of inference (here, Bayes' rule; see also Kiefer 2017). The sensory signal's normative involvement is constituted by how it shapes the *causal* transitions between contentful states in such a way that they accord to a *normative* rule. Thanks to Thomas Metzinger and Jona Vance for pointing this issue out to me.

is not my aim here to settle it definitely. Nonetheless, let me tentatively sketch out how this problem could be handled in a way that suggests a positive answer to the original question.

A crucial observation to make is that although perceptual inference makes use of a pre-existing model of the environment, the model as such is subject to change through perceptual learning (Clark 2013a; Clark 2013b; Clark 2016). If we treat the model as a hypothesis space, then this space is anything but fixed once and for all. Even if perceptual inference is only as good as the range of available hypotheses, the latter can change and improve over time through learning. One of the factors that make learning possible is, again, the receptivity of the senses. Because the world is basically independent of how it is represented in one's generative model, the patterns that arise in the sensorium can, and do leave average residual prediction error. This way, the sensory input enables the system to "recognize" that the world escapes even the best predictions that its current model can afford. In learning, this fact is exploited to optimize the effectiveness of the model at minimizing the prediction error over longer periods of time. At heart, learning consists in modifying the model parameters, presumably with the use of some form of gradient descent algorithm (Clark 2013b; Friston 2003; Friston and Kiebel 2009). Starting from an initial "guess", the hierarchical structure of the model is iteratively adjusted so as to improve its overall ability to minimize prediction error. In other words, not only perceptual inference (i.e. forming hypotheses to deal with current sensory input) but also perceptual learning (i.e. changing the structure of the model itself) is basically prediction error minimization, performed at different time scales (Clark 2013b).

Importantly, in learning thus construed, the causal structure of the environment is gradually recovered from the sensory input itself. This process does not require supervision by an external observer providing the cognitive system with data that are already pre-classified (Hohwy 2013; see also Eliasmith 2005). Because what is available at the outset is basically raw data (perhaps accompanied with relatively minimal learning biases), learning on this view can be seen as a form of bootstrapping (Clark 2013a; Clark 2013b; Clark 2016). Furthermore, a body of computational work suggests that systems which learn using multi-level Bayesian strategies are able to recover the values of hyperpriors or "overhypotheses" (see the work on learning with Hierarchical Bayesian Models: Kemp et al. 2007; Kemp and Tenenbaum 2008; Tenenbaum et al. 2011; a summary can be found in Clark 2016). These are high-level priors that structure the range of hypotheses available at lower levels of the processing hierarchy. This way the learner can recover "deep" or abstract overarching principles of how the data are to be categorized, like the shape bias in categorizing physical objects (see Clark 2016, pp. 171–175). Hyperpriors are themselves selected by maximizing posterior probability. This means that a hypothesis space available at a lower level is selected (out of a range of possible such spaces) based on whether it is most likely to capture the causal structure that generated the data (Kemp et al. 2007).

But what consequences do these considerations have for the question of unconditional justification or warrant, assuming that they do? Here is a way of thinking about this, albeit admittedly sketchy and speculative. Consider an approach that, in a way, puts classical empiricist foundationalism on its head. On this view, representations that are unconditionally justified are not the starting points of the inquiry but, so to speak, its *endpoints*. The crux of this proposal has been formulated by Charles Sanders Peirce:

> Different minds may set out with the most antagonistic views, but the progress of investigation carries them by a force outside of themselves to one and the same conclusion. This activity of thought by which we are carried, not where we wish, but to a fore-ordained goal, is like the operation of destiny. No modification of the point of view taken, no selection of other facts for study, no natural bent of mind even, can enable a man to escape the predestinate opinion. This great law is embodied in the conception of truth and reality. (Peirce 1878/2011, p. 63)

On this general approach, the notion of unconditional justification (or rational entitlement) is cashed out in terms of what we might call "epistemic convergence". Roughly, the idea is that there are beliefs that an agent will eventually form, regardless of what she accepts at the starting point, if she proceeds by rationally updating her beliefs in light of evidence. These are the beliefs on which rational inquiry converges, and they are justified in an unconditional sense. Gupta, who himself adopts this approach (Gupta 2006), expresses this in terms of a convergence that results from a revision sequence, in which successive experiences $< e_0, e_1, e_2, \dots, e_n, \dots >$ result in a sequence of views $< v_1, v_2, v_3, \dots, v_n, \dots >$, where each previous experience revises subsequent view. Formal and philosophical technicalities aside, the idea is that revision sequences may converge on particular views or propositions, regardless of their starting points. For example, a proposition $p$ constitutes a point of convergence at stage $n$ of the revision sequence generated by a particular series of experiences $E$ if, and only if, regardless of the starting point $v0$, all views that survive revision in light of $E$ include $p$ at $m \geq n$. A proposition on which the revision process eventually converges regardless of where it starts is a proposition which the subject is rationally entitled to accept in an unconditional sense.

It seems that the general notion of epistemic convergence can be accommodated by PP. The hypothesis is that systems which learn by iteratively revising their generative models to optimize long-term prediction error minimization can, and perhaps do reach points of convergence. And we may treat the representations on which such systems converge, regardless particular starting points and learning trajectories, as ones which are justified in an unconditional sense. For example, we may speculate that cognitively healthy people (as well as, perhaps, members of many other species) inhabit a perceptual world that is roughly the same in terms of its basic ontological furnishing. It is a world filled with medium-sized physical objects, which persist through time, occupy spatial locations, can retain their identities despite changing their properties, and which undergo causal interactions with each other. From the perspective of PP, this unity of how the world is perceptually represented may stem precisely from a sort of epistemic convergence. Prediction error minimizing systems (with relevantly similar bodies and sensory apparati) will converge to acquire (learn) generative models that parse the world into ordinary physical objects. This convergence point may be actually reached quite early in ontogeny. But it still presumably results from a bootstrapping process in which the values of hyperpriors are adjusted to constrain the range of hypotheses available at lower levels. The common-sense perceptual ontology could then be treated as unconditionally justified in virtue of it constituting such a convergence point of iterative model revision. Crucially for the present discussion, the "driving" influence of the sensory signal has major significance in enabling such convergence to occur. The convergence is possible because the sensory signals that organisms receive embody similar statistical regularities (which is in turn due to the signals actually being generated by the same physical world), so that the learning processes in different agents tend to proceed, at least at some level, in uniform direction. In other words, it is the sensory signal that constitutes the "outside force" which inevitably carries the learning process to "one and the same conclusion".

## 5 Conclusion

The basic tenet of empiricism is that we owe (a large chunk of) our knowledge to the epistemic authority of "the evidence of the senses". When we perceive, our thinking becomes answerable to the empirical world. Despite intuitive plausibility, this claim proves notoriously difficult to defend or even formulate in a way that is not philosophically problematic. One problem that empiricism needs to face is the Sellarsian dilemma, which purports to show that sensory states are (if we consider them to be nonintentional) incapable of conferring justification on intentional states, or (if we construe them as intentional) incapable of terminating the chain of inferential justifications.

In the present article, I attempted to reconsider the Sellarsian dilemma from the perspective of the Predictive Processing (PP) view of perception. According to PP, perception centrally involves an active

or constructive aspect, whereby an internal generative model attempts to predict the inflow of sensory signals. But it also involves an aspect that is "receptive" in roughly Kantian sense, as the model adjusts its predictions to match the signal that results from the world impinging on the perceiver's sensory apparatus. On such a view, the epistemic authority of the senses lies in how the sensory signal constrains the activity of the model, making it "answerable" (through error-correction) to external states of affairs. The justification or rational entitlement that perceivers thus obtain to their perceptual hypotheses is conditional, as it depends on the epistemic standing of the generative model that interprets the incoming signal. However, because of how, in perceptual learning, the model itself is corrected using the guidance of sensory input, we can at least begin to make sense of the idea that our sensory contact with the world can provide us with unconditional justification or entitlement as well. The resulting view significantly diverges from classical empiricist foundationalisms which saw the evidence of the senses as consisting in basic, non-inferentially justified representational states. Contrary to this tradition, PP rather leads us to view the sensory signal as a purely contentless "foundation", one that passively causally registers, rather than represents, what happens in the empirical world.

## References


BonJour, L. (1985). *The structure of empirical knowledge.* Harvard: Harvard University Press.

Bucci, A. & Grasso, M. (2017). Sleep and dreaming in the predictive processing framework. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND-Group.

Burge, T. (2003). Perceptual entitlement. *Philosophy and Phenomenological Research, LXVII*, 503-548.

Chisholm, R. (1977). *Theory of knowledge.* Englewood Cliffs, NJ: Prentice-Hall.

Clark, A. (2013a). Expecting the world: Perception, prediction and the origins of human knowledge. *The Journal of Philosophy, CX*, 469-496.

——— (2013b). Whatever next? Predictive brains, situated agents and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181-204.

———(2016). *Surfing uncertainty. Prediction, action, and the embodied mind.* Oxford: Oxford University Press.

Cohen, S. (1984). Justification and truth. *Philosophical Studies, 46*, 279-295.

Conee, E. & Feldman, R. (2001). Internalism defended. *American Philosophical Quarterly, 38*, 1-18.

Davidson, D. (1973). On the very idea of a conceptual scheme. *Proceedings and Addresses of the American Philosophical Association, 47*, 5-20.

———(1986). A coherence theory of truth and knowledge. In E. Lepore (Ed.) *Truth and interpretation: Perspectives on the philosophy of Donald Davidson* (pp. 307-319). Oxford: Blackwell.

Eliasmith, C. (2005). A new perspective on representational problems. *Journal of Cognitive Science, 6*, 97-123.

Fodor, J. (1984). Observation reconsidered. *Philosophy of Science, 51*, 23-43.

Friston, K. J. (2003). Learning and inference in the brain. *Neural Networks, 16*, 1325-1352.

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

Friston, K. J. & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of Royal Society B, 364*, 1211-1221.

Gupta, A. (2006). *Empiricism and experience.* Oxford: Oxford University Press.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese, 193*, 559-582.

Hobson, J. A. & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology, 98*, 82-98.

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

——— (2014). *The self-evidencing brain.* Noûs. https://dx.doi.org/10.1111/nous.12062.

———(forthcoming). The predictive processing theory and 4e cognition. In A. B. Newen & S. Gallagher (Eds.) *The Oxford handbook of cognition: Embodied, embedded, enactive and extended.* Oxford University Press.

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

Kant, I. (1781/1996). *Critique of pure reason, abridged.* Indianapolis/Cambridge: Hackett Publishing Company.

Kemp, C. & Tenenbaum, J. B. (2008). *The discovery of structural form* (pp. 10687-10692). 105.

Kemp, C., Perfors, A. & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*, 307-321.

Kiefer, A. (2017). Literal perceptual inference. In T. Metzinger & W. Wiese (Eds.) *Philosophy of predictive processing*. Frankfurt am Main: MIND-Group.

Langton, R. (1998). *Kantian humility. Our ignorance of things in themselves.* Oxford: Oxford University Press.

Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology, 6*, 547-569.

Lyons, J. (2008). Evidence, experience, and externalism. *Australasian Journal of Philosophy, 86*, 461-479.

Orlandi, N. (forthcoming). *Bayesian perception is ecological perception.* Philosophical Topics.

Peirce, C. S. (1878/2011). How to make our ideas clear. In R. B. Talisse & S. F. Aikin (Eds.) *The pragmatism reader: From peirce through the present* (pp. 50-65). Princeton, NJ: Princeton University Press.

Quine, W. V. O. (1969). Epistemology naturalized. *Ontological relativity and other essays* (pp. 69-90). New York: Columbia University Press.

Ramsey, W. (2007). *Representation reconsidered.* Cambridge, MA: The MIT Press.

Rescorla, M. (2015). Bayesian perceptual psychology. In M. Matthen (Ed.) *The oxford handbook of philosophy of perception* (pp. 694-717). Oxford: Oxford University Press.

——— (2016). Bayesian sensorimotor psychology. *Mind & Language, 31*, 3-36.

Sellars, W. (1956). Empiricism and the philosophy of mind. In H. Feigl & N. Scrived (Eds.) *Minnesota studies in the philosophy of science* (pp. 253-329). Minneapolis, MN: University of Minnesota Press.

Seth, A. (2015). Inference to the best prediction. A reply to Wanja Wiese. In T. Metzinger & J. M. Windt (Eds.) *Open MIND* (pp. 4-20). Frankfurt am Main: MIND Group. http://open-mind.net/papers/inference-to-the-best-prediction.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. (2011). How to grow a mind: Statistics, structure and abstraction. *Science, 331*, 1279-1285.

# Moving from the What to the How and Where – Bayesian Models and Predictive Processing

Dominic L. Harkness & Ashima Keshava

The general question of our paper is concerned with the relationship between Bayesian models of cognition and predictive processing, and whether predictive processing can provide explanatory insight over and above Bayesian models. Bayesian models have been gaining influence in neuroscience and the cognitive sciences since they are able to predict human behavior with high accuracy. Models based on a Bayesian optimal observer are fitted on behavioral data. A good fit is hence interpreted as human subjects "behaving" in a Bayes' optimal fashion. However, these models are performance-oriented and do not specify which *processes* could give rise to the observed behavior.
Here, David Marr's (Marr 1982) levels of analysis can help understand the relationship between performance- and process-oriented models or explanations. Bayesian models are situated at the computational level since they specify *what* the system (in this case the brain) does and *why* it does it in this manner. Although Bayesian models can constrain the search space for hypotheses at the algorithmic level, they do not provide a precise solution about *how* a system realizes the observed behavior. Here predictive processing can shed more light on the underlying principles. Predictive processing provides a unifying functional theory of cognition and can thus i) provide an answer at the algorithmic level by answering *how* the brain realizes cognition, ii) can aid in the interpretation of neurophysiological findings at the implementational level.

**Keywords**

Bayesian models | Explanation | Marr | Predictive processing | Unification

## 1 Introduction

Recent findings indicate that the human brain can be seen as a Bayesian inference machine (Knill and Pouget 2004). This is motivated by the fact that the brain faces the so-called inverse problem: the brain cannot gain access to the world outside of the skull by itself, and must rely on the sensory organs to make sense of the world. This sensory input is oftentimes noisy, and furthermore, a many-to-many relationship holds between external causes and perceived sensory effects. As a consequence, the brain must *infer* the causes of sensory input from these effects.

Bayesian models of cognition create ideal observer models of cognitive phenomena and use them as a backdrop to which human performance is compared (Colombo and Seriés 2012). What these models have shown is that humans often behave in a Bayes' optimal fashion and therefore the behavior itself can be accurately predicted (Ernst and Banks 2002). Bayesian models can thus be seen as performance-oriented, since it is solely the behavior of the ideal observer that is compared to the behavior of human subjects. However, this paper will argue that this does not suffice if one wants to reach an explanation of human cognition.

What is additionally needed are process-oriented models that can describe which functional or physical components are involved in giving rise to human behavior as well as the causal relations that hold between these components. We will argue in this paper that Bayesian models of cognition alone do not provide such insight. Instead, we propose that predictive processing (Clark 2013; Hohwy 2013) can provide a process-oriented theory that can give an account of *how* Bayesian inference is realized by the brain. Furthermore, unlike Bayesian models, predictive processing can serve as a so-called mechanism sketch, thus being able to guide researchers in finding mechanistic explanations of a target

Dominic L. Harkness & Ashima Keshava (2017). Moving from the What to the How and Where – Bayesian Models and Predictive Processing.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 16. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573178

cognitive phenomenon. To exemplify how Bayesian models relate to predictive processing we turn to David Marr's (Marr 1982) three level account of explanations. The paper is structured as follows: first we will describe Bayesian models and predictive processing and follow with an introduction to David Marr's levels. Finally we will compare Bayesian models and predictive processing in light of Marr's levels of analysis.

## 2 Bayesian Models

One of the central problems in the cognitive sciences is how the brain builds rich and generalized models of the world given sparse and noisy sensory data. In this regard, Bayesian models have become an increasingly popular tool in the understanding of cognition. They have been used to model, amongst others, visual perception (Yuille and Kersten 2006), inductive learning and human reasoning (Tenenbaum et al. 2006), motor planning (Körding and Wolpert 2006), or multisensory integration (Ernst and Banks 2002). In this section, we will explore the basis of Bayesian modeling, taking the problem of multisensory cue integration as an example.

One of the attractions of the Bayesian approach is its simplicity of formulation. The three core tenets of the modeling scheme are: the task of the organism, prior knowledge of the environment and the knowledge of the way the environment is sensed by the organism (Kersten et al. 2004). These three basic components can then be used to model and further predict an organism's behavior.

This approach can be realized with Bayes' rule. For instance, if we have a set of hypotheses $H$ and want to test the probability of a given hypothesis h within this set. Before obtaining evidence, we assume that this hypothesis h has a probability, $P(h)$, which is the prior probability. We can then observe certain data d given this hypothesis which brings us to $P(d|h)$ which is the likelihood of observing this data. Using Bayes' rule we can then update the probability of the hypothesis given the data, i.e. $P(h|d)$, which is the posterior probability of h given the data d. The remaining term $P(d)$ is the marginal probability of d, i.e. the sum of the joint probabilities over the hypothesis space $H$. Simply put:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

The problem of how the brain integrates different sensory signals to form a coherent picture of the world is one that is of great importance, and has been approached by Bayesian modelling techniques extensively. Bayesian models of multisensory integration have thus far shown that humans are close to Bayes' optimal when integrating sensory signals of different modalities, i.e. human performance closely fits the predictions made by these models (Ernst and Banks 2002; Körding and Wolpert 2004; Triesch et al. 2002). In a similar vein, Shams et al. (Shams et al. 2005) showed that visual and auditory signals are integrated or segregated depending on the stimulus condition in a Bayesian fashion and thus gave a computational account of the phenomenon of sound-induced flash illusion (Shams et al. 2000; Shams et al. 2002). They investigated this by presenting subjects with a varying number of light flashes on the screen with a simultaneously varying number of sound beeps in each trial and having them report the perceived number of flashes and beeps. When one flash was presented with one beep, the signals 'appear' to originate from the same source. Whereas, when one light flash is coincident with four sound beeps, the two signals are perceived to originate from different sources and are, hence, considered separate events. If a single flash is accompanied by two beeps, the single flash is often perceived as two flashes and the flashes and beeps are perceived to originate from the same source. Based on this paradigm, Shams et al. (Shams et al. 2005) developed an ideal observer model that would account for bimodal sensory signal segregation, partial integration and complete integration. Human behavioral data was then compared with the ideal observer model.

The Bayesian model (ideal observer model) was developed with the assumption that auditory signal ($A$) and visual signal ($V$) are statistically independent as noise processes that corrupt these signals

are independent. It is also assumed that $A$ and $V$ are caused by separate sources $Z_A$ and $Z_V$ respectively. Thus, information about the likelihood of signal $A$ occurring given a source $Z_A$ is given by the probability distribution $P(A|Z_A)$ Similarly, $P(V|Z_V)$ represents the likelihood of sensory signal $V$ given a source $Z_V$. The distribution $P(Z_A|Z_V)$ denotes the observer's prior knowledge about auditory and visual events in the environment. As Shams et al. mention, "the priors may reflect hard-wired biases imposed by the physiology and anatomy of the brain, [...] as well as biases imposed by the task, the observer's state, etc." (Shams et al. 2005, p. 1924). Given the auditory and visual signals, an ideal observer would try to best estimate the sources $Z_A$ and $Z_V$, which can be represented as a posterior probability distribution $P(Z_A, Z_V|A, V)$. By applying Bayes' rule, the following results:

$$P(Z_A, Z_V|A, V) = \frac{P(A|Z_A)P(V|Z_V)P(Z_A, Z_V)}{P(A, V)}$$

Shams et al. (Shams et al. 2005) approximated the priors from the observed data by marginalizing the joint probability across all combinations of $A$ and $V$. To do this, data from half of the participants was used to estimate the priors and the rest of the data was used to test the model's predictive accuracy:

$$P(Z_A, Z_v) = \sum_{A,V} P(Z_A, Z_V|\, A,V)P(A,V)$$

Results suggested that human observer's performance were consistent with the ideal observer's. The observed human behavioral data fitted well with the model predictions for conditions in which signals from the two modalities were integrated and segregated. Hence, it was shown that "the brain uses a mechanism similar to Bayesian inference [...] as a statistically optimal computational strategy" (Shams et al. 2005, p. 1927) to integrate signals from different modalities.

It must be mentioned here that the example taken above deviates from traditional models of Bayesian cue integration (Beierholm et al. 2007), that assume a single signal source location that gives rise to latent variables belonging to different modalities e.g. a visual and an auditory. In these cases the source is determined using the maximum likelihood estimation method. This is done by taking the inverse of the variances of the visual and the auditory signal and representing them as precisions of the visual and auditory signal.

As seen above, Bayesian models specify when a behavioral response to a given stimulus can be regarded as rational or optimal according to Bayes' rule. However, since Bayesian models are performance-oriented they do not give an account of *how* this Bayesian optimality is achieved in the brain, either functionally or physically. They do not inform us about the underlying causal structure that leads to the human behavior predicted by Bayesian models. One hypothesis that does aim at giving a process-oriented theory in the realm of the Bayesian brain hypothesis (Knill and Pouget 2004) is called predictive processing (PP) (Clark 2013; Clark 2016) or prediction error minimization (PEM) (Hohwy 2013).

## 3 Predictive Processing

Picture the brain as a black box that cannot directly access the world outside of the skull. The only means it has to gain such access consists of utilizing the sensory information that is provided to the brain via the sensory organs. For example, the eyes provide the brain with visual sensory information. As a consequence, the brain does not have access to the external causes that lead to the sensory input. Thus the brain is in the business of having to infer the causes of sensory input from its effects. This has been coined the inverse problem. The question then is how is the sensory input used to gain any knowledge about the hidden external causes that lead to that sensory input?

The general assumption is that the information that is provided by the sensory organs will always contain a certain amount of uncertainty due to noise or other factors that may distort the sensory signal. Also, one cause may have many different effects and the opposite is also possible, that many causes can have the same effect. Taking an example from Hohwy (Hohwy 2013), an actual bicycle, a bicycle poster, or even a bee swarm that coincidentally flew in a formation that resembles a bicycle might all give rise to the same visual percept. How then could the brain discern between these different possibilities? For one it could assess the probability of each possibility, e.g. how probable is it that the bicycle-like visual image being perceived is actually a swarm of bees? Then, depending on the context, different hypotheses are more likely. For example, if a person perceives a bicycle-like visual input in front of a train station, the priors of the different hypotheses will not be equal from the start. The bee hypothesis seems unlikely in any case, since bees do not usually fly in bicycle-like formations. The poster hypothesis seems unlikely as well, since people generally do not park bicycle posters in front of a train station. Yet, bicycle posters do exist and are thus still more likely than the bee hypothesis. The most likely hypothesis seems to be that the visual information represents an actual bicycle.

How then can the brain assess how likely any of these hypotheses are? Why will humans, even if the sensory information could be the same in all three cases, i.e. different causes leading to identical effects, nonetheless perceive an actual bicycle in most cases?

Here, we must appreciate that sensory inputs contain statistical regularities. These regularities are captured in so-called generative models — models that aim at representing the causal structure of the world in a probabilistic format (every time I clap my hands a sound occurs) and are continuously updated in light of the evidence. One of predictive processing's core tenets is that these models are updated and new information is integrated according to Bayes' rule. With the help of these generative models, the brain can then attempt at predicting what the next sensory input might be given the last sensory input. For example, if a face that is lit from the top turns to the left, the brain can persistently predict how the lighting on the face will change since these lighting effects are also subject to statistical regularities that have been perceived in the past.

We now have two values that can be continuously compared with one another: the actual sensory input and the predicted sensory input which is dependent on previous observations about some facet of the world. To solve the inverse problem and thereby infer external causes from their effects, predictive processing argues that the brain is a system that constantly compares predicted with actual sensory input and tries to minimize the discrepancy between these two values up to expected levels of noise, i.e. to minimize prediction errors. The reasoning behind this is that the less prediction error occurs, the better the model fits the sensory input, and a model that has a good fit accurately represents the causal structure of the world. However, prediction errors will occur in every case due to the noise in sensory input. In the case of high prediction error, the brain can reduce these prediction errors in two ways. Either it can change its generative models according to the sensory input. This can be seen as learning (or perceptual inference), since the models are updated in light of new sensory information to further refine the models and ultimately achieve a more accurate representation of the external causes that lead to that sensory input. Alternatively, it can change the sensory input to match the predictions deriving from the models. This would then be action, or more formally active inference (Friston et al. 2009). Referring back to the bicycle example from above, by moving closer to the seen bicycle-like visual image, the agent could more accurately discern different hypotheses. Yet, what determines which path is chosen to reduce the amount of prediction error?

As mentioned before, sensory input always contains a varying amount of noise. To deal with this, the brain must "take [this] variability [...] into consideration — it needs to assess the precision of the prediction error." (Hohwy 2012, p. 4). In other words, since sensory information can contain different amounts of noise, the brain must be able to determine whether a given sensory input can be regarded as 'trustworthy', i.e. does this sensory information provide *precise* information about the external causes or should the brain rather rely on its models? It is then the precision of prediction errors (Hes-

selmann et al. 2010) that determines whether models are updated in light of new evidence or the agent engages in active inference to match the sensory input to its expected states. If noise in sensory input is low, prediction errors are amplified, since they are precise and model revisions can ensue. If noise is high, the resulting prediction errors are deemed imprecise and lead to an amplification of predictions (ibid., p. 1). Just as there are generative models about external causes that lead to the effects observed by the sensory organs, there are also models about precisions. Since noise in sensory input is context-sensitive and may vary between sensory modalities and the environmental setting (for example that precision declines on foggy days in the visual domain), it is necessary for the brain to have these precision estimates (Hohwy 2012; Friston and Stephan 2007; Feldman and Friston 2010) in order to better minimize prediction errors.

The optimization of these precision estimates have been identified with attention (Feldman and Friston 2010; Hohwy 2012), arguing that it may serve as "a gating or gain mechanism that somehow optimizes sensory processing." (Hohwy 2012, p. 6). Thus, by attending to a certain feature of the sensory input the precision can be increased and consequently the amount of prediction error minimized in a more efficient fashion.

Another crucial feature of predictive processing is that prediction errors are minimized across a cortical hierarchy spanning over numerous levels (Friston 2005; Mumford 1991). Each of these levels is concerned with different properties of the sensory input and contains generative models about these properties. Thus, any level attempts to predict the next state of the level below. Furthermore, as one goes down the hierarchy, the temporal grain at which predictions are made increases and sensory properties are more variant. For example, predictions in V1 must be able to process fast-changing properties of the visual input. As one goes up the hierarchy, the processed properties become more invariant and the temporal scale increases, as for example in temporal cortex. This allows "the brain to build up representations of environmental causes from basic stimulus attributes to more and more abstract and invariant properties." (Hohwy 2010, p. 136). As a consequence, "each cortical area is an expert for inferring certain aspects of the visual scene" (Lee and Mumford 2003, p. 1436) and this can actually be seen if one looks at the functional segregation between cortical areas. For example, V1 processes basic sensory properties, such as orientation or location, of the visual input via simple and complex cells (Hubel and Wiesel 1968), V4 has been associated with processing form and structure (Desimone and Schein 1987) and area MT with motion perception (Maunsell and Essen 1983).

The hierarchical structure of predictive processing implies that the major portion of processing proceeds top-down, since predictions are passed down from higher to lower levels. Bottom-up messages on the other hand then only carry the prediction errors to the levels above for further processing.

As a short summary, predictive processing argues that the brain only has access to its own states due to its 'black-box' status. It uses sensory information provided by the sensory organs to create generative models that capture the causal structure of the world (inferring from effects to causes). These generative models are in turn used to predict the brain's next state. If the predictions match the sensory input well, a good model has been selected, since it accurately represents the world and consequently is perceived. However, if a discrepancy between the predicted and actual sensory input arises, meaning a high amount of prediction error occurs that exceeds the expected levels of noise, the system (the brain) will either engage in active inference or change its models according to the sensory input. The factors that determine which option is pursued are the precision of the prediction errors and the respective priors. High precision prediction errors lead to model revisions, since the sensory input is regarded as trustworthy. Low precision on the other hand leads to action, the aim being to adjust sensory input according to the model.

## 4 Marr's Levels

Having presented both Bayesian models and predictive processing, the question arises how these two frameworks relate to each other. Here, David Marr's (Marr 1982) levels of analysis can help.

### 4.1 Levels of Analysis

Marr (Marr 1982) proposed that any cognitive system needs to be analyzed at three different levels in order to fully explain it. These are the computational, algorithmic, and implementational levels and each level should answer certain questions about the investigated system (p. 23f). At the computational level the *what* and *why* questions are answered, i.e. *what* does the system do and *why* does it do it? Thus, specifying an optimal behavioral output to a certain perceptual input and stating *why* this output is optimal would be located at the computational level. The algorithmic level answers *how* the system accomplishes *what* it does by concentrating on the underlying processes that lead to the investigated behavior of the system in question. Yet, theories at the algorithmic level need not specify *where* or *how physically* the system is realized. These questions are answered at the implementational level, i.e. *where* in the brain is the system localized and *how* is the system *physically* realized?

Although these levels seem to provide a straightforward approach to investigating a cognitive system, the relationship between levels remains unclear. The relation between the levels of analysis is one of realization, meaning what are the processes that lead to the observed behavior and how are these processes realized by a physical system. Also, Marr advocated that the three levels should be seen as formally independent of one another, i.e. that many algorithms could realize the computational problem and the algorithms could be realized by many different physical parts. However, this formal independence "does not entail that the algorithms used by the human cognitive system are best discovered independently of a detailed understanding of its neurobiological mechanisms." (Colombo and Seriés 2012, p. 17, original emphasis).

We argue that the minimal requirement for an algorithmic-level theory consists in making *causal* claims about the structure of the system, and for an implementational-level theory that its proposed components of the system are *structurally* describable. Computational-level theories do not and need not meet either of these requirements. For a computational-level model it suffices to specify an output to some input. Often, these input-output relations have a mathematical formulation. Yet, as Marr (Marr 1982) has proposed, all levels should be considered if one wants to reach a full explanation of any cognitive system. The alternative is so-called single-level theorizing, which leads to incomplete explanations for several reasons. For example, remaining at the computational level results in such theories being largely "under-constrained and somewhat arbitrary" (Love 2015, p. 233) since the behavior observed and described by computational-level theories could be inconsistent with results from e.g. cognitive psychology or neuroscience (Griffiths et al. 2012, p. 264; Colombo and Seriés 2012). If there are no physical counterparts that can compute what a certain computational-level model presupposes it seems to make little sense to further pursue that particular model.[1] Likewise, remaining at the implementational level may result in having descriptions about the neural hardware, reaction times, or neural networks, yet still being unable to incorporate these insights into higher-order cognitive systems or concepts (Cooper and Peebles 2015). For example, how do neural firing rates inform us about emotions or problem-solving tasks? Such bottom-up driven theories appear unguided by an overarching question and "do no more than mimic in an unenlightening way." (Marr 1982, p. 347, in Cooper and Peebles 2015, p. 2).

Our intuition behind this paper is nicely captured by Kaplan & Craver (Kaplan and Craver 2011) who state that "the line that demarcates explanations from merely empirically adequate models seems

1 This is only the case when one is interested in the *human* brain and *human* cognition, not some artificial system that may achieve human-like behavior, yet is composed of different parts/hardware.

to correspond to whether the model describes the relevant causal structures that produce, underlie, or maintain the explanandum phenomenon" (p. 602). In the case of computational-level theories or models, insight into the causal structure of the system cannot be gained. This is due to the fact that these types of theories or models are performance-oriented, i.e. they aim at specifying an output given some input while "no internal structure is specified within the model" (Colombo and Seriés 2012, p. 10). Algorithmic — as well as implementational-level theories on the other hand are regarded as process-oriented theories since they provide insights into the causal processes that realize the abstract computational problem. Again, the aim of the algorithmic and implementational level is to give an account of *how* and *where* the computational problem is computed/solved.

## 4.2 Marr's Levels, Bayesian Models and Predictive Processing

It is widely agreed upon that Bayesian models are located at the computational level since "[t]hey help researchers understand what a cognitive system does, because they describe and predict its behavior." (Zednik and Jäkel 2014, p. 666, emphasis added) and "attempt to explain why cognition produces the patterns of behavior that [it] does." (Jones and Love 2011, p. 170). To what extent computational-level theories such as Bayesian models can inform and offer constraints at the algorithmic and implementational level is currently being debated in light of the Bayesian program (Zednik and Jäkel 2014; Zednik and Jäkel 2016; Colombo and Hartmann 2015; Bowers and Davis 2012). Zednik & Jäkel (Zednik and Jäkel 2016) for example argue that Bayesian models can constrain theories at the algorithmic level by reverse-engineering from the computational level to the levels below via a number of heuristics. In this paper we will not dive into the details of this discussion since it suffices for our argument that Bayesian models do exhibit the tendency to remain at the computational level and that "mechanism is neglected in favor of a focus on behavior." (Love 2015, p. 233).[2] Bayesian models do not provide sufficient explanations for cognitive systems and provide little insight into the causal structure of the investigated system (algorithmic) nor the physical entities that constitute that system (implementational) by themselves. This poses a problem for so-called top-down approaches that aim at proceeding from the computational level downwards to the algorithmic and implementational levels (Love 2015) as it remains unclear "whether a given probabilistic model is inconsistent with particular cognitive or neural processes." (Griffiths et al. 2012, p. 264).

Predictive processing on the other hand can be considered an algorithmic level theory since it provides an account of *how* cognition may be realized. It's very ambitious in its scope and argues that cognition adheres to one core principle: the minimization of prediction error. It divides a system such as the brain into different subcomponents and provides a theory of how these subcomponents interact with one another. Here predictive processing has been criticized since up to this point the concepts employed are purely functional. They do not make any reference to neurobiological structures in the brain (Rasmussen and Eliasmith 2013).

Yet, there is mounting empirical evidence in favor of this theory. Consequently, as evidence accumulates and physical properties are identified that could realize the functional properties proposed by predictive processing, predictive processing can incrementally be regarded as an implementational theory. This means identifying physical structures that could realize e.g. prediction errors, precision, or how the cortical hierarchy is structured. In fact, some of these concepts have already been identified with neurophysiological entities that can be structurally described. In the remainder of this paper we will present four such cases.

One of the most important aspects of predictive processing is that the brain is structured hierarchically. This entails the distinction between feedforward (prediction errors) and feedback (predictions) connections. Predictive processing then gains 'implementational weight' once these connections, so far only described functionally, are associated with entities in the brain that can be identified and

2 Mechanisms (Craver 2007; Bechtel 1994) are located at the algorithmic and implementational levels.

structurally described. This has been done. It has been hypothesized that the two mentioned types of connections are realized by pyramidal cells in the brain. In particular, feedforward connections are associated with superficial pyramidal cells, and feedback connections with deep pyramidal cells (Mumford 1991; DeFelipe et al. 2002; Friston 2005). Since these types of cells can be e.g. measured and analyzed, and are able to pass on messages as supposed by predictive processing, they seem to be ideal candidates to count as realizers of the structural hierarchy.

The next example concerns how precision could be realized in the brain. Friston et al. (Friston et al. 2012) argue that the realization of precision may consist in the modulation of the synaptic signal-to-noise ratio via synaptic gain. The proposed structural entity responsible for this mechanism is the neurotransmitter dopamine. Since dopamine is involved in many different cognitive functions, and a unifying theory regarding the function of dopamine is still lacking, this approach "provides a novel perspective on the role of dopamine that accounts for its apparently diverse roles in terms of a single mechanism". (ibid., p. 2).

Another strong indication for predictive processing comes from Hosoya et al. (Hosoya et al. 2005) who state that retinal ganglion cells encode changes in the visual field to which an organism (here salamanders and rabbits) has adapted to rather than the raw visual image. Spike trains from ganglion cells in the retinae of salamanders and rabbits were recorded and the visual field was manipulated so that the subject was adapted to that particular environment and then a novel/uncorrelated stimulus was used to probe the ganglion cells' receptive fields. It was seen that the response of the ganglion cells flattened (reduced) once it was adapted to a particular environment while they became more sensitive to novel stimuli. This systematic effect of enhancing sensitivity to novel stimuli and relative to the adapting environment gives a strong suggestion towards a strategy of dynamic predictive coding (Rao and Ballard 1999).

Lastly, predictive processing can also aid in the interpretation of extra-classical field effects. If two neighboring neurons have the same orientation preference and a visual stimulus extends over the boundaries of the receptive field of one of these neurons, the response of that particular neuron will be suppressed. This type of effect has been found in a number of brain areas (V1, V2, V4, MT; (Allman and Miezin 1985)) and been termed as extra-classical receptive field effects. Here, Rao & Ballard (Rao and Ballard 1999) argue that "visual cortical neurons with extra-classical [receptive field] properties can be interpreted as residual error detectors, signaling the difference between an input signal and its statistical prediction based on an efficient internal model of natural images." (Rao and Ballard 1999, p. 79).

Our main argument then consists in the following: by taking evidence from both the computational level (provided by Bayesian models) and implementational level (provided by neurophysiological findings) into consideration, one may, albeit provocatively, conclude that the algorithmic level (predictive processing) can be regarded as the best candidate to form the bridge between behavior and the brain (Love 2015). This means that Bayesian models increase the evidence in favor of predictive processing at the computational level and neurophysiological findings at the implementational level.

By considering how Bayesian models, predictive processing and implementational findings relate to each other in the Marrian framework we can assess how each of them contribute to understanding cognition as a whole. Bayesian models provide no (or at least very little) insight into the causal structure of a system. Implementational level findings on the other hand give accurate physical descriptions of a system's components and their interactions, but fail to provide an over-arching theory about how these single components or processes result in a large-scale system such as the brain. Lastly, predictive processing gives an algorithmic level account of how cognition is functionally realized. As mentioned in the previous paragraphs, predictive processing then seems to be an ideal candidate to bind computational level theories with implementational findings due to being an algorithmic level theory.

## 5 Conclusion

According to the Marrian framework, to reach a full explanation of a target system one must investigate and understand that system at all three levels of analysis. At the computational level we have Bayesian models, that accurately predict human behavior and give strong reasons to interpret the brain as a Bayesian inference machine. However, due to their focus on performance rather than the underlying processes, Bayesian models lack insight into the physical components that lead to the observed human behavior, and the causal relations that hold between them. Yet, this does not mean that they do not contribute to the scientific enterprise of understanding human cognition. They do provide strong evidence that human behavior approximates Bayesian optimality (in line with predictive processing).

Next we have implementational findings provided by the neurosciences that investigate and describe physical structures in the brain. Yet, taken by themselves, they provide little insight into how a complex system, such as the brain, could realize the multitude of behaviors that it does. More importantly, explaining how implementational-level findings lead to the observed behaviors at the computational level seems almost impossible.

Our main argument then states that predictive processing, as an algorithmic-level theory, is an ideal candidate to tie computational-level findings with implementational-level ones together. Bayesian models confirm predictive processing's premise that humans do in fact approximate Bayes' optimal behavior. Predictive processing then provides a functional theory of how such behavior comes about. Lastly, once implementational-level findings are identified that can physically realize the functions proposed by predictive processing, we reach an increasingly detailed account of mind and brain.

## References


Allman, J. & Miezin, F. (1985). Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual Review of Neuroscience*. https://dx.doi.org/10.1146/annurev.ne.08.030185.002203.

Bechtel, W. (1994). Levels of description and explanation in cognitive science. *Minds and Machines, 4* (1), 1–25. https://dx.doi.org/10.1007/BF00974201.

Beierholm, U., Shams, L., Ma, W. J. & Koerding, K. (2007). *Comparing Bayesian models for multisensory cue combination without mandatory integration* (pp. 81–88).

Bowers, J. S. & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychol Bull, 138* (3), 389-414. https://dx.doi.org/10.1037/a0026450.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

Colombo, M. & Hartmann, S. (2015). Bayesian cognitive science, unification, and explanation. axv036. https://dx.doi.org/10.1093/bjps/axv036.

Colombo, M. & Seriés, P. (2012). Bayes in the brain—on Bayesian modelling in neuroscience. *63* (3), 697-723. https://dx.doi.org/10.1093/bjps/axr043.

Cooper, R. P. & Peebles, D. (2015). Beyond single-level accounts: The role of cognitive architectures in cognitive scientific explanation. *Top Cogn Sci, 7* (2), 243-58. https://dx.doi.org/10.1111/tops.12132.

Craver, C. F. (2007). *Explaining the brain. Mechanisms and the mosaic unity of neuroscience.* Oxford: Oxford University Press.

DeFelipe, J., Alonso-Nanclares, L. & Arellano, J. I. (2002). Microstructure of the neocortex: Comparative aspects. *Journal of Neurocytology*.

Desimone, R. & Schein, S. J. (1987). Visual properties of neurons in area V4 of the macaque: Sensitivity to stimulus form. *Journal of Neurophysiology*.

Ernst, M. O. & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*. https://dx.doi.org/10.1038/415429a.

Feldman, H. & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*, 215.

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

Friston, K. J. & Stephan, K. E. (2007). Free-energy and the brain. *159* (3), 417-458.

Friston, K. J., Daunizeau, J. & Kiebel, S. J. (2009). Reinforcement learning or active inference? *4* (7). https://dx.doi.org/10.1371/journal.pone.0006421.

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. & Bestmann, S. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology, 8* (1), e1002327. https://dx.doi.org/10.1371/journal.pcbi.1002327.

Griffiths, T. L., Vul, E. & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science, 21* (4), 263-268. https://dx.doi.org/10.1177/0963721412447619.

Hesselmann, G., Sadaghiani, S., Friston, K. J. & Kleinschmidt, A. (2010). Predictive coding or evidence accumulation? False inference and neuronal fluctuations. *PLoS ONE, 5* (3), e9926. https://dx.doi.org/10.1371/journal.pone.0009926.

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. https://dx.doi.org/10.5096/ASCS200922.

——— (2012). Attention and conscious perception in the hypothesis testing brain. *Front Psychol, 3*, 96. https://dx.doi.org/10.3389/fpsyg.2012.00096.

———(2013). The predictive mind.

Hosoya, T., Baccus, S. A. & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature, 436* (7047), 71-77. https://dx.doi.org/10.1038/nature03689.

Hubel, D. H. & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195* (1), 215-243.

Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34* (4), 169. https://dx.doi.org/10.1017/S0140525X10003134.

Kaplan, D. & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science, 78* (4), 601-627. https://dx.doi.org/10.1086/661755.

Kersten, D., Mamassian, P. & Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol., 55*, 271–304.

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

Körding, K. P. & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427* (6971), 244–247.

——— (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences, 10* (7), 319–326.

Lee, T. S. & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Optical Society of America, 20* (7), 1434.

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science, 7* (2), 230-242. https://dx.doi.org/10.1111/tops.12131.

Marr, D. (1982). *Vision: A computational approach.* San Francisco: Freeman.

Maunsell, J. H. & Essen, V. D. C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. *Journal of Neurophysiology*.

Mumford, D. (1991). On the computational architecture of the neocortex. *Biological Cybernetics, 65* (2), 135-145. https://dx.doi.org/10.1007/BF00202389.

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

Rasmussen, D. & Eliasmith, C. (2013). God, the devil, and the details: Fleshing out the predictive processing framework. *Behavioral and Brain Sciences*. https://dx.doi.org/10.1017/S0140525X12002154.

Shams, L., Kamitani, Y. & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature*.

———(2002). Visual illusion induced by sound. *Cognitive Brain Research*.

Shams, L., Ma, W. J. & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport, 16* (17), 1923–1927.

Tenenbaum, J. B., Griffiths, T.L. & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*.

Triesch, J., Ballard, D. H. & Jacobs, R. A. (2002). Fast temporal dynamics of visual cue integration. *Perception, 31* (4), 421–434.

Yuille, A. & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends Cogn. Sci. (Regul. Ed.), 10* (7), 301-8. https://dx.doi.org/10.1016/j.tics.2006.05.002.

Zednik, C. & Jäkel, F. (2014). How does Bayesian reverse-engineering work? In P. Bello, M. Guarini, M. McShane & B. Scassellati (Eds.) *Proceedings of the 36th annual conference of the cognitive science society* (pp. 666–671). Austin, TX: Cognitive Science Society.

———(2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*. https://dx.doi.org/10.1007/s11229-016-1180-3.

# Literal Perceptual Inference

Alex Kiefer

In this paper, I argue that theories of perception that appeal to Helmholtz's idea of unconscious inference ("Helmholtzian" theories) should be taken literally, i.e. that the inferences appealed to in such theories are inferences in the full sense of the term, as employed elsewhere in philosophy and in ordinary discourse. The argument consists in first defending a minimal conception of inference based on Gilbert Harman's account (Harman 1973), and then arguing that Helmholtzian computational models of perceptual inference such as those proposed in Hinton and Sejnowski 1983, Hinton et al. 1995, and Friston 2005 implement the type of process Harman describes.

In the course of the argument, I consider constraints on inference based on the idea that inference is a deliberate action (Boghossian 2014; Broome 2014; Wright 2014), and on the idea that inferences depend on the syntactic structure of representations (Mandelbaum 2016). I argue that inference is a personal-level but sometimes unconscious process that cannot in general be distinguished from association on the basis of the structures of the representations over which it's defined. I also critique the argument against representationalist interpretations of Helmholtzian theories in Orlandi 2015, and argue against the view that perceptual inference is encapsulated in a module.

**Keywords**

Artificial neural networks | Bayesian Inference | Free energy minimization | Generative models | Induction | Inference | Perceptual inference | Predictive processing | Representation

**Acknowledgements**

I would like to thank the anonymous reviewers of this paper, as well as Grace Helton, Geoffrey Hinton, Jakob Hohwy, Zoe Jenkin, Hakwan Lau, Eric Mandelbaum, Thomas Metzinger, Nico Orlandi, Jona Vance, David Papineau, Jake Quilty-Dunn, David Rosenthal, Wanja Wiese, and the Cognitive Science group at the CUNY Graduate Center, for conversations and suggestions that have informed my thinking on these matters.

## 1 Introduction

Helmholtz proposed the idea, so influential within recent cognitive science, that what we perceive in sensory experience is the conclusion of unconscious inductive inference from sensory stimulation. Less famously, he questioned whether the term 'conclusion' could be applied to the deliverances of perception in the same "ordinary" sense in which it is applied to conscious acts of reasoning (Von Helmholtz 1860/1962, p. 4). This is not, of course, a merely verbal question. If we know that a term such as 'knowledge', 'memory', or 'inference' is being used in an unusual sense, we should be able to articulate the difference between this sense and the usual one, and to that extent we must understand the nature of the corresponding phenomenon.

Helmholtzian theories of perception thus put philosophical pressure on the concept of inference. Such theories include predictive processing models (Friston 2005; Rao and Sejnowski 2002; Huang and Rao 2011; see Clark 2013 and Hohwy 2013 for discussion), as well as many models of perception in machine learning that in part inspired the predictive processing framework, notably those discussed in Hinton and Sejnowski 1983, Hinton et al. 1995, and Hinton 2007, as well as Oh and Seung 1997 and many others.[1] Taken at face value, these theories are committed to the view that the representations underlying perceptual experiences are literally the conclusions of inductive inferences, taking sensory representations and background knowledge or memories as premises (Aggelopoulos

1 By 'model' I mean a formal structure along with its interpretation, which I take to be a type of theory.

Kiefer, A. (2017). Literal Perceptual Inference.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 17. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573185

2015). In this paper, I argue that the literal interpretation of such theories is warranted, and that apart from the lack of conscious awareness that gave Helmholtz pause, which may be regarded as inessential to inferential mechanisms themselves, there is no compelling reason to deny that Helmholtzian perceptual inference is the genuine article.

## 2 Inference

In this section, I attempt to provide a well-motivated account of inference as a psychological process, independently of any role it may play in perception. The account draws heavily on Gilbert Harman's work (Harman 1973), which gives a central role to the notion of coherence. In the second section of the paper, I first argue that this account of inference, though couched in terms of propositional attitudes, can plausibly subsume processes defined over sensory representations. I then describe in detail how I take perceptual inference to be realized in Helmholtzian computational models.

### 2.1 Inference as Reasoned Change in View

A general account of inference must cover all paradigmatic uses of the term 'inference'—most saliently, it should accommodate deductive reasoning as well as various sorts of ampliative reasoning, such as enumerative induction and abduction or "inference to the best explanation". Paul Boghossian, following Harman (Harman 1986), offers an intuitively plausible characterization of inference as a "reasoned change in view": a process "in which you start off with some beliefs and then, after a process of reasoning, end up either adding some new beliefs, or giving up some old beliefs, or both" (Boghossian 2014, p. 2). This first pass must be generalized in certain ways.

First, we should make room for processes of reasoning that alter degrees of belief or subjective probabilities without necessarily resulting in the wholesale dropping or adding of beliefs. Such processes intuitively fit the description 'reasoned change in view', and Boghossian could be viewed as describing a special case in which probabilities are changed to or from zero. Second, in hypothetical reasoning, certainly a paradigm of inference, one may infer $Q$ from $P$ without believing either, for example to explore the consequences of a counterfactual. Thus, the above description should be modified so that it appeals to some weaker attitude than belief, such as provisional acceptance (Wright 2014, p. 29).[2] I alternate between the two formulations in what follows, noting where a generalization to acceptance raises special issues.

Clearly, inference so conceived may involve more than just drawing new conclusions from premises. It may, for example, involve lowering the probability of previously held beliefs based on new evidence, or rejecting the premise of an argument whose conclusion is inconsistent with an entrenched belief. It is thus not clear that inference in general can be modeled on simple syllogistic reasoning.

Consider for example a change in belief that takes $P$ and $P \rightarrow Q$ as premises and, instead of resulting in the belief that $Q$, results in the rejection of $P \rightarrow Q$ (Harman 1973, p. 157). This could perhaps be understood as a chain of inferences in which one first infers $Q$ from the premises, and then infers from the combination of $Q$ and one's prior belief that $\sim Q$ that one must have been mistaken about $P \rightarrow Q$ (a *reductio*). There are at least two concerns about this proposal, however: one must momentarily believe both $Q$ and $\sim Q$ (or assign probabilities to these propositions that sum to more than 1), and some grounds must be supplied for rejecting $P \rightarrow Q$ rather than $P$.

These examples suggest that inference involves more than consideration of relations of logical implication even where such relations are relevant. Harman (Harman 1973, ch. 10, section 4) takes this to show that there are no distinctively deductive inferences, as opposed to deductive arguments. Boghossian (Boghossian 2014, p. 5), similarly, suggests that deduction can be distinguished from induction in terms of the standards for evaluation (logical entailment VS probabilistic support) that (one

2 By 'acceptance' I mean, like Wright , an attitude toward a proposition that, like belief, involves commitment, but in which the commitment may be merely provisional, hypothetical, or temporary. Thus, as Wright claims (Wright 2014, p. 29), supposition is a kind of acceptance—as is belief.

takes to) apply to one's inference, but that this distinction gives us no reason to suppose that there are two intrinsically different types of inference. I'll provisionally assume that Harman and Boghossian are right about this, and return to the point shortly.

Harman (Harman 1973) develops an account of inference designed to handle examples like the above. He conceives of inference not as a serial process akin to mentally traversing a syllogism, but as a parallel process that takes one's total current evidence as input and yields a new total set of beliefs (Harman 1973, p. 159):

> A more accurate conception of inductive inference takes it to be a way of modifying what we believe by addition and subtraction of beliefs. Our "premises" are all our antecedent beliefs; our "conclusion" is our total resulting view. Our conclusion is not a simple explanatory statement but a more or less complete explanatory account.

On Harman's account, all inference is essentially inference to the best explanation (where some "explanations", as in deductive inferences, are arguably trivial). In the following paragraph Harman suggests that this inferential process is constrained by two competing principles, coherence-maximization and change-minimization:

> Induction is an attempt to increase the explanatory coherence of our view, making it more complete, less ad hoc, more plausible. At the same time we are conservative. We seek to minimize change. We attempt to make the least change in our antecedent view that will maximize explanatory coherence.

Harman's characterization of coherence is rather laconic, but Laurence BonJour (BonJour 1985) articulates a serviceable notion that is consistent with Harman's usage, and which I will assume in what follows. According to BonJour, a coherent set of beliefs must minimally be largely logically consistent, as well as enjoying a high degree of consistency between the probabilities and truth-values assigned to its members (e.g. the combination "*p*" and "It is highly improbable that *p*" lessens coherence). Lack of inconsistency is not sufficient for coherence, however, since a set of unrelated beliefs may be internally consistent without intuitively cohering. Thus, BonJour supposes in addition that coherence is increased with "the number and strength of inferential connections" between members of a consistent collection of beliefs (BonJour 1985, p. 98). On this assumption, beliefs in (non-trivial) explanatory statements (e.g. scientific laws) greatly enhance the coherence of a system, since they are typically inferentially related to many other beliefs.

Harman's coherence-based account allows hypothesis testing to be viewed as a special case of inference in which one of the premises is an observation that may not cohere with existing belief. Thus, the Duhem-Quine thesis about confirmation holism (Quine 1951) applies also to inference as Harman characterizes it. This holism can explain why the premises that lead to an unacceptable conclusion may be treated differently during belief-revision: they may stand in different evidential relations to the other things one believes.

This account of inference also provides a compelling way to avoid the lottery paradox[3] and similar cases. We do not know that any particular ticket in the lottery will lose because our evidence doesn't favor any one ticket winning over the others. All the alternatives are equally coherent, so no total view

3 The lottery paradox, invented by Henry Kyburg (Kyburg 1961 p. 197) and often discussed in the context of epistemology, arises if one accepts a principle to the effect that any very highly probable hypothesis should be accepted as true. In a large lottery with one winner, "Ticket number *X* will lose" is extremely probable for each *X*. Thus, given such a principle, we should accept that no ticket will win, which contradicts the prior belief that one ticket will win.

containing the belief that a particular ticket will lose can be inferred (Harman 1973, p. 160; see also Lehrer 1986, pp. 155-156).[4]

Generalizing Harman's account to acceptances may seem problematic, because what distinguishes acceptances from beliefs is precisely that they do not depend on, and indeed must be capable of conflicting with, one's beliefs. For this reason, I do not take it to be a constraint on all inference that it be holistic in the sense that any inference must take as input one's total body of evidence (though some inferences may do so). Harman's account can be extended to acceptances and thus to hypothetical reasoning by supposing that the same type of holistic process that governs rational change in belief governs rational change in acceptance, where the range of acceptances taken as input to the process depends on the context and is distinct from one's set of beliefs. Given this extension, syllogistic reasoning such as that involved in deduction can be construed as reasoning in which only a very small set of acceptances is considered.

## 2.2 Rationality

Thus far inference has been characterized as a process that modifies (degrees of) belief (or acceptance, more broadly), such that coherence among beliefs (or acceptances) is maximized. Implicit in this account is the idea that changing one's view in a way that results in greater coherence is rational. In this section I discuss how standards of rationality for inductive and deductive reasoning may be taken on board by this type of account.

Logical implication is of course a paradigm of rational transition between representations. The rationality of a process in this sense is a matter of its tendency to preserve truth (necessarily, in the case of deductively valid inference). The rationality of induction may arguably be characterized in essentially the same way, as on BonJour's (BonJour 1985, p. 96) view that inference must be "to some degree" truth-preserving. Hume 's skepticism about induction aside, inductive inference is a process that tends to yield true belief in worlds like ours.

The preceding assumes that modification of degrees of belief or subjective probability can be truth-preserving. There is precedent for this view in formal treatments of induction such as those offered by Reichenbach (Reichenbach 1949). Carnap (Carnap 1950), similarly, generalizes deductive consequence relations to "partial implications" (degrees of inductive support). A detailed treatment of this topic is beyond the scope of this paper, but correspondences between logical truth-functions and analogous rules for probability suggest that at least simple truth-functions of propositional logic can be viewed as special cases of the axioms of probability theory in which the relevant probabilities are Boolean truth values, assuming independence for the probabilities:

| Expression | Boolean truth function | Probability |
|---|---|---|
| $A \wedge B$ | $A * B$ | $p(A) * p(B)$ |
| $A \vee B$ | $A + B - (A * B)$ | $p(A) + p(B) - p(A)p(B)$ |
| $\sim A$ | $1 - A$ | $1 - p(A)$ |

**Figure 1.** Simple truth functions in propositional logic as special cases of rules relating probabilities.

More systematically, an obvious way of understanding probability in terms of truth is to adopt a generalized frequentist account according to which probabilities are measures of frequency relative to sets of situations, actual or hypothetical. From this perspective conditional probabilities, Bayesian

4 Harman offers one more reason to accept a holistic account of inference: arguably, one may not rationally infer that $P$ without also believing that there is no evidence against $P$ of which one is ignorant. To respect this principle while avoiding a regress of inferences, we must suppose that the belief that there is no undermining evidence against $P$, call it $\sim UE(P)$, is inferred along with $P$ (as part of a total most explanatory account) rather than antecedently, since inference to $\sim UE(P)$ would itself require a similar belief $\sim UE(\sim UE(P))$, which would require a previous inference, etc. (Harman 1973, p. 153). I mention this argument only in a footnote, because it depends on many assumptions tangential to the topic at hand.

subjective probabilities, and modal claims (including implicitly modal claims such as subjunctive conditionals) can be treated similarly. Subjective probabilities may be regarded as frequencies relative to a set of possible worlds consistent with one's evidence.

Instantiation of a psychological process can thus be considered an exercise of (theoretical) rationality[5] to the extent that we have reason to believe it will yield true outputs given true inputs (both actually and counterfactually), where truth may be evaluated in a single situation only or simultaneously in a range of situations. Inferences that result in dropping beliefs may fit this description, since dropping a belief may be a way of avoiding believing something false. And when large collections of representations are involved, truth-preserving transition amounts to maximization of coherence.

The most pressing objection to the account sketched so far is that, since it picks out inferential processes as just those processes that conform to a standard of rationality, no inference can fail to conform to this standard. As Boghossian (Boghossian 2014, p. 4) puts it in discussing a similar proposal, "if one is reasoning at all, one is reasoning to a conclusion that one has justification to draw". Of course, a process may tend toward truth-preservation but fail to preserve truth in a particular case. But simple appeal to a type-token distinction cannot meet the objection, because there seem to be *types* of reasoning that fail to yield true beliefs—systematic failures of rationality such as base rate neglect and denial of the antecedent (Wright 2014, p. 37).

An advocate of the present account of inference needn't deny, however, that the latter are species of reasoning. One possible response is to appeal to a hierarchy of types, rather than a simple type-token distinction. Systematic errors in reasoning may be due to the application of heuristics that are effective in one domain to analogous domains in which they fail. Consistent with this suggestion, Gerd Gigerenzer claims that agents often approach cognitive problems not by performing optimal inference in conformity with strict canons of rationality, but instead by in effect substituting for the target problem a simpler one that is easier to solve and, at least in the given context, may be just as likely or more likely to succeed than the ideally rational method (Gigerenzer calls this the "ecological rationality" of heuristics—see, e.g. Goldstein and Gigerenzer 2002).[6] Though Gigerenzer appeals to this mechanism to explain successful inference, it may also explain cases of cognitive failure in a way that preserves their rationality, as suggested here.

A more substantial response, one which I favor but which needs elaboration and defense beyond what can be given here, is that apparent systematic failures of rationality are systematic failures to represent the target problem correctly. Base rate neglect, for example, is a failure to take base rates into account in inferring probabilities, but corresponds to optimal Bayesian inference *given* that the base rate is neglected, i.e. if the actual base rate is replaced with 0.5 in Bayes's theorem. In the case of denial of the antecedent, we may suppose that reasoners ignore the deductive validity of the argument and simply consider whether $Q$ is likely to be the case, given only the information that $P \rightarrow Q$ and $\sim P$. Absent any other reason to believe $Q$, denial of the antecedent is rational and amounts to assuming $Q$ only if $P$, i.e. $Q \rightarrow P$. This is a misrepresentation of the original argument in question, but is correct reasoning given that misrepresentation. To take one more example: the reasoning that goes on when one affirms the consequent may be regarded, simply, as an inference to the best explanation of $Q$, where $P$ is the only contextually available explanation.

5 Practical rationality will not be discussed here, but the notion of "active inference" (Friston 2011) suggests that a unification of theoretical and practical reason is also conceivable along these lines, since active inference makes transitions between high-level hypotheses and sensory predictions truth-preserving and is thus a rational process in the sense defined here.

6 I thank an anonymous reviewer for suggesting this connection to Gigerenzer's work. In the bigger picture, of course, there is potentially some tension between that work and the Bayesian perspective afforded by predictive processing theories (though one may consider approximate Bayesian inference a heuristic of sorts).

John Anderson suggests just such a treatment of apparent irrationality, which is worth quoting at length:

> Many characterizations of human cognition as irrational make the error of treating the environment as being much more certain than it is. The worst and most common of these errors is to assume that the subject has a basis for knowing the information-processing demands of an experiment in as precise a way as the experimenter does. What is optimal in the micro-world created in the laboratory can be far from optimal in the world at large (Anderson 1991, p. 473).

Similar remarks apply to cases such as the latter two discussed above, in which a piece of reasoning "in the wild" that is sensible given partial information is evaluated against canons of deductive rationality that the subject may not be imposing. [7]

## 2.3 Reasoning as Deliberate Action

The foregoing conception of inference as rational change in view is rather minimal compared with recent accounts that emphasize the deliberate, conscious, personal-level character of paradigmatic reasoning. In this section, I consider such alternative views of what inference consists in, and argue that they either boil down to the minimal account or impose unwarranted constraints on inferential processes.

Boghossian argues that "Inferring necessarily involves the thinker taking his premises to support his conclusion and drawing his conclusion because of that fact" (Boghossian 2014, p. 5). 'Taking' here may be interpreted in a variety of ways: as having a further belief to the effect that the premises support the conclusion, as following an implicit or explicit rule of inference (Boghossian's preferred interpretation), or as being disposed to give the premise(s) as reason(s) for the conclusion. The motivation for this "taking" condition is twofold: it distinguishes inferential processes from other causal transitions between representations, and it ensures that inference is something that one does with a goal, rather than something that simply happens, perhaps within a cognitive subsystem (see 3.2 below for further discussion of the latter possibility).

The minimal account defended above may satisfy the "taking" condition on a weak enough reading. Taking a conclusion to follow from some premises may simply be a matter of regularly inferring the conclusion from the premises, or being disposed to do so, *ceteris paribus*, where inferring is already distinguished from mere causal connection (and from irrational or arational association, should there be such a thing) by the rationality, in the sense of subjunctively truth-preserving character, of the transition.

One concern here is that, while Harman's account of reasoning rules out the possibility of its occurring within a module due to its holism, the modification of the account to include reasoning over acceptances seems to sacrifice this feature. But it's plausible that reasoning over acceptances relies on the same implicit background beliefs as ordinary reasoning, and so still requires access to general world knowledge. The inferential models of perception discussed in the next section of the paper are consistent with this possibility.

Still, the minimal account would likely not do justice to "taking" as Boghossian intends it, because such inference may for all that's been said occur automatically and unconsciously, and Boghossian is explicitly concerned to give an account of deliberate, conscious, "System 2" reasoning. But it's not clear that all inference is done deliberately (consider drawing a conclusion one doesn't like in spite of oneself). And it begs the question against the minimal account to suppose that reasoning that occurs consciously differs, *qua* reasoning, from reasoning that occurs without consciousness. This is so even if one grants that inference is by definition a personal-level, goal-directed activity, since one may do

7 Thanks to Wanja Wiese for suggesting this connection to Anderson's work on the method of "rational analysis" (Anderson 1991).

things purposefully but without awareness of doing them. One may, for example, take a break from working on a cognitively demanding problem and wake from a nap with the solution. The obvious explanation of cases like this is that one has been reasoning unconsciously.[8]

John Broome (Broome 2014) agrees with Boghossian that reasoning is a matter of rule-following, and offers an account of following a rule in terms of a complex disposition: to follow a rule in doing *X* is to have a disposition both to do *X* and for doing *X* to seem "right" to you with respect to the relevant rule. The second clause distinguishes reasoning from mere causation (Broome 2014, p. 21). Broome suggests that this account satisfies Boghossian's "taking" condition in something like the weak way suggested earlier.

But it is not clear why the simple disposition to conclude *Q* from *P* isn't sufficient for inference, where the connection between *P* and *Q* is rational. Animals that employ fewer higher-order monitoring mechanisms than humans do can plausibly still reason, at least in limited ways. Such animals may have no disposition for their inferences to seem correct to them. While Broome doesn't require that inference involve conscious "taking", his account still requires meta-representation that needn't be supposed essential to inference, especially if a more minimal account is viable.

Crispin Wright rejects Boghossian's "taking condition" and supposes simply that inference is a matter of accepting a conclusion for the reason that one accepts the premises (Wright 2014, p. 33), which is a case of acting for reasons more generally. Wright is clear that on his view acting for reasons is not a matter of meta-representation but rather a matter of acting in a way that conforms to constitutive constraints on rational action (p. 35). In particular, he claims that acting "in accordance with basic rules of inference is constitutive of rational thought" (p. 36).

Since preserving truth is arguably *the* constitutive norm on rational thought, the minimal account plausibly satisfies Wright's description. If one arrives at a set of acceptances *C* as a result of a rational process that takes another set *R* as input, one may be said to accept *C* for the reason that one accepted *R*. It is less obvious whether the minimal account can be extended to action generally, unless action can be said to preserve truth in the way that inference can.[9] Wright does not, in any case, offer an analysis of acting for reasons, so the minimal proposal has no clear competitor with respect to this issue.

## 2.4 Inference, Association and Compositional Structure

I have so far said nothing about what, if anything, distinguishes inference from association. This issue may seem particularly pressing given the connectionist (and therefore, some would argue, associationist) pedigree of Helmholtzian computational models of perception. Though his primary concern is not to give an account of inference, Eric Mandelbaum (Mandelbaum 2016, p. 8, fn. 14) claims that inference is distinguishable from other sorts of mental transition between propositional contents in terms of its computational profile. In particular, in inference, "The mental transition between the premises and the conclusion occurs not because they were (e.g.) associated through prior learning, but instead because they conform to the logic of thought", where the latter is a system of rules that governs transitions between mental representations, analogous to but likely distinct from classical logic.

What distinguishes this view from the rule-following proposals just discussed is the claim that the rules of the envisaged logic are sensitive to the formal or syntactic structure of the representations they govern: "A mental inference is a transition in thought that occurs because the argument structure instantiated the germane cognitive rules of inference." Mandelbaum supposes that such rules operate on "Structured Beliefs": representations whose relevant core features are that they (a) have compositional

8 To anticipate a bit, it may be objected that the automatic inference supposed by Helmholtz to occur unconsciously is certainly not goal-oriented behavior. But this depends on how one defines the latter notion. Perception is arguably often goal-oriented. And an inference may seem to be 'automatic' because it is drawn immediately in the face of compelling evidence, as may be the case with the sensory evidence involved in perceptual inference.

9 Active inference may provide a route to making sense of this, since predictions include intentions, motor commands, and other representations with a "world-to-mind" direction of fit. Another possibility would be to construe any action as constitutively involving an intention-in-action (along the lines of Searle 1983), whose content represents the satisfaction-condition of a desire and is true when the action succeeds.

structure, (b) relatedly, have a proprietary syntax and semantics, and (c) sometimes enter into causal relations that mirror the "implicational structure" of their contents.

Clause (c) is clearly compatible with the view defended above, according to which syntactically specified rules in effect play no essential role in inference. Transitions likely to preserve truth can be picked out by appeal to a formal logic, but in order for a mental transition to be rational and thus inferential it is (on the minimal view) sufficient for it to in fact conform to the pattern of inferences defined by the logic.

What (a) and (b) add is, in effect, a psychologically realist interpretation of the logic, the internal structure of whose formulas is attributed to the corresponding mental representations by an inference to the best explanation. Presumably, this form of explanation amounts to the assumption that the relevant representational vehicles (neurons and populations of them in this case) contain reasonably concretely specifiable parts, possession of which explains why the causal interactions among the vehicles mirror the relevant structure of implications (see, e.g., Fodor 1975).

But appeal to such compositional structure cannot be used to define inferential transitions in general unless it is also taken to cover inductive inference. And it seems highly unlikely that any syntactic rules governing inductive inferences will serve to distinguish them from associations. This is not to say that structure has no role to play in understanding induction. It may be theoretically useful, for example, to bring simple enumerative inductions under the sway of syntax by treating their premises as conjunctions of the relevant evidential claims. But this requires only propositional logical structure as mediated by connectives, not subject-predicate structure internal to atomic sentences. This suggests that while the internal structure of representations may be the best explanation of certain types of inference, it is not a requirement on inferentially related representations generally.

Moreover, assume that the conditional probability of *Q* given *P* is the proportion of situations in which *P* obtains that are also situations in which *Q* obtains, and also that degree of inductive support in such simple cases as "It's raining, therefore the streets are wet" can be defined in terms of conditional probability. In paradigm cases of associative learning, the strength of the association from *P* to *Q* depends in the same way on conditional probability, where the relevant set of situations is a sample consisting of observed cases. Thus, simple cases of induction may be indistinguishable from association between propositions, as Hume in effect contended.

Mandelbaum notes the difficulty of distinguishing inductive inferences from associations, but claims that associations are inherently symmetric (i.e. if *P* is associated with *Q* then *Q* is associated with *P*), which would serve to distinguish them (p. 6-7).[10] Presumably, the symmetry of association is supposed to arise from the fact that strength of association depends on previous co-activation of representations, and co-activation is a symmetric relation.

Hebbian models of neural plasticity, however, suggest that association should be intrinsically asymmetrical, since the efficiency of a synapse from neuron *a* to neuron *b* depends on the extent to which firing of *b* is contingent on *a*'s having fired very recently, and not on the reverse relation. Symmetrical associations may as a matter of empirical fact be likely to occur as a result of associative learning, but on a Hebbian story this symmetry would be implemented via a pair of reciprocal synaptic connections, each of which would mediate a distinct associative link.[11] This does not suggest a mechanism for association distinct from that subserving inductive inference. It may suggest also that though there is

10 In making the point about symmetry, Mandelbaum is discussing what distinguishes associative *structures* in memory from propositional structures, not what distinguishes associative from non-associative transitions. But he also claims (reasonably, I think) that the only difference between associative structures and transitions is that the former involve co-activation of representations while the latter involve one representation activating another after some delay.

11 This issue is somewhat complicated in the context of Mandelbaum's discussion. On the one hand, he treats it as sufficient for association that one concept activate a second without a further mediating computational relation (p. 10), and this description would be satisfied by a one-way synaptic connection between two neural representational vehicles of the appropriate sort. On the other hand, he distinguishes between (a) associative connections between mental states and (b) associative neural network "implementation bases" for such states. But the latter distinction, if read as a dichotomy, would beg the question against those who take connectionist models to be both models of neural dynamics and theories of mental architecture, as I in effect do.

a clear distinction between paradigmatic rational belief revision and modification of associative structures by counter-conditioning or extinction, the representational changes caused by conditioning and by inductive inference are rational in the same sense.

While the foregoing discussion has been necessarily brief, I take it that the above considerations are sufficient to defend the minimal account of inference against pressing objections. In the next section I discuss perceptual inference and how it is modeled in several computational theories.

## 3 Perceptual Inference

The recent wave of computational models based on Helmholtz's theory propose that perception is a matter of inferring the best explanation of sensory input by inverting a generative model. A generative model (for present purposes) is a causal model that structurally mimics the process by which sensory input is generated (or more generally, any model capable of generating states of the input channel similar to those caused by the external world). Its inverse is a recognition model that maps from sensory input to explanations or, more narrowly, causes. This section of the paper explores how this idea is implemented in three distinct models, after consideration in the first two subsections of some putative reasons to doubt that perception could be based on inference.

### 3.1 Truth-Evaluability

It is uncontroversial that perception allows us to arrive at largely accurate judgments about what is in the environment on the basis of sensory input. This ensures that perceptual judgments are rationally related to *knowledge of* the inputs to the perceptual process. Therefore, what seems most relevant to evaluating Helmholtz's theory is the claim that the inputs and outputs of the perceptual process are themselves truth-evaluable representations. There are challenges to this claim on both ends.

On the input side, it's been argued that the proximal inputs to perceptual processes themselves should not be confused with the knowledge of those inputs that could serve as a basis for inference. Davidson (Davidson 1986), for example, criticizes Quine for conflating the causal intermediaries that lead from distal stimulus to belief with epistemic intermediaries in this way. And Tyler Burge agrees that "sensation does not play the role of data or evidence. Thinking that it does is the primary mistake of the sense-data tradition." (Burge 2010, p. 367).

With respect to output, it's not always clear how representations that are constitutive of perception itself can be distinguished from post-perceptual judgments (see Siegel 2010 for an attempt at providing a method for doing so, but also Quilty-Dunn unpublished for a critique), and it may be doubted whether perceptual experience always involves such judgments. The general project of distinguishing perception from cognition is of course arguably challenged by Helmholtzian theories, and it is difficult to address this issue without begging the question either way. Fortunately, any account of the inputs to perception according to which they are truth-evaluable will likely suffice to show that the same is true for representations further downstream, so that perception will turn out to involve inference no matter where one draws its upper boundary. And if the inputs to perception are not truth-evaluable, this suffices to show that perception is not inferential in the relevant sense.[12]

A possible reason to deny that perceptual representations could figure in inferences is that perceptual experiences seem to involve rich, non-discursive (e.g. iconic (Fodor 2007), qualitative (Rosenthal 2005), analog (Dretske 1981), or nonconceptual (Evans 1982)) contents. But it is not clear why a representation should fail to qualify as truth-evaluable simply because it occurs in an iconic, qualitative or otherwise special format. Arguably, all contents can be expressed in terms of an exhaustive enough (set of) proposition(s), so that format distinctions are orthogonal to the issue of whether a given content is truth-evaluable. This claim is liable to ring false if one assumes that genuinely truth-evalu-

12 One could still suppose in this case that perception involved inference at some later stage of processing, perhaps in some sub-perceptual module. I argue that the inputs to perception are truth-apt so I do not explore this possibility here.

able representations *ipso facto* possess language-like subject-predicate syntactic structure, but as we've seen, nothing about inference as such requires this.[13]

Perceptual contents are in any case clearly specifiable in terms of propositions, as Susanna Siegel (Siegel 2010) notes. Typically, each such proposition partially characterizes what is perceived in terms of one of its aspects—in other words, specifies a part of the overall perceptual content (for example, one may see, among other things, *that the cat is green*). If contents are the accuracy or veridicality conditions of representations (see, e.g., Burge 2010), it seems natural to identify perceptual contents with the sum of these parts, i.e. with the set of contents (truth-conditions) of the individual propositions sufficient to specify them. In Dretske's somewhat idiosyncratic terms, sensory representations may carry many distinct pieces of information in an "analog" format.

More concretely, an iconic representation, for example, may be composed of an array of truth-evaluable representations that mark the light intensities across a visual field, or the relative positions of edges, or more generally, the components of any feature map.[14] It is consistent with the models of perceptual processing discussed below to suppose that this decomposability of iconic perceptual content into propositional contents is reflected in vehicular structure, where each component representation is implemented by a neuron or neural population, corresponding to a node in a connectionist model (see e.g. Hinton and Sejnowski 1983). However, the basic claim about truth-evaluability is independent of this idea.

## 3.2 Modularity and Sensory Representations

Perception is often supposed to be modular in nature, operating without influence from beliefs and responsive only to limited sources of information. If it is, then the causal intermediaries between sensory stimulation and perceptual judgment may at best be "subpersonal" or "subdoxastic" representations[15], or perhaps not representational states at all (Orlandi 2015), in which case the theory of inference defended above would not apply to them (as mentioned already in 2.3).[16]

I cannot comprehensively address the modularity question here (see Zoe Drayson's contribution to the present collection—Drayson 2017), but as mentioned previously, Helmholtzian theories arguably posit architectures that are nonmodular with respect to coarse-grained functional distinctions such as that between perception and cognition. Such theories are supported by mounting empirical pressure against modularity, which has given rise to debates over the "cognitive penetrability" of perception (Jenkin and Siegel 2015; Lupyan 2015), some specifically in the context of predictive coding theories (Newen et al. 2017).

For example, the existence of "extra-classical receptive field effects", cited by Friston as evidence for his model (Friston 2005), constitutes a form of context-sensitive and therefore presumably rational top-down modification of early sensory representations (i.e. in striate cortex—see, e.g., Harrison et al. 2007). Top-down modification in some cases stems from sources in higher cognitive areas (for example, orbitofrontal cortex has been shown to modulate activity in temporal regions during object recognition—see O'Callaghan et al. 2017).

13 Burge (Burge 2010, pp. 230-232) argues in effect that any truth-evaluable representation, at least of the kind involved in perception, must be structured in a way that both allows for repeated application (i.e. exhibits generality) and depends on particulars in a specific context of application. This suggests object-property structure in the semantics, but Burge does not discuss syntactic structure, and allows that the particular element in perceptual representation may be "entirely implicit".

14 This is in some ways similar to an account of mental imagery defended by Michael Tye (Tye 1991).

15 As Zoe Drayson points out, the use of the term "subpersonal" to type-individuate psychological states marks a perhaps dubious extension of its original use to distinguish among types of psychological explanation (Drayson 2012, section 2.5; see also Dennett 1978). I intend the term to pick out states or representations that occur within special-purpose cognitive subsystems or modules, in accordance with contemporary usage, and, as Drayson points out, with Stephen Stich's use of the term "subdoxastic" (Stich 1978), discussed below.

16 Perceptual processing may occur within a cognitive subsystem but still realize Harman's account of inference at the "subpersonal" level, in that coherence is maximized with respect to representations within the module. Such processing would fail to meet the Boghossian/Wright conditions on inference discussed in section 2.3, however, because the "inferences" in question would not be attributable to the subject capable of paradigm (e.g. conscious, deliberate) reasoning.

Drayson suggests that there are senses in which "predictive architectures" may nonetheless implement forms of modularity. In particular, though probabilistic dependence is transitive, she suggests, causal dependence needn't be, so hierarchical models that implement Bayesian inference in their causal structures may exhibit a form of information encapsulation of one level relative to others (Drayson 2017, p. 9).[17] But this sort of limited causal dependence, even if it sufficed for modularity, would not suggest that perceptual inference is encapsulated within a module. Perceptual inference, by hypothesis and as implemented in the computational models discussed below, is simply the process by which incoming sensory data is assimilated into a prior model of the world. This process may comprise many modular operations, but is itself as widely distributed throughout the hierarchy as is needed to facilitate the minimization of prediction error subsequent to the impact of sensory signals. It is therefore potentially holistic in the sense relevant to Harman's account of inference.

Even if the mind is not modular, the representations involved in allegedly modular processes are widely supposed to differ importantly from beliefs. Stephen Stich (Stich 1978), for example, claims that "subdoxastic" representations differ from beliefs in that (a) they play very limited roles in inference, in contrast to beliefs which are "inferentially promiscuous", (b) they are normally not accessible to conscious awareness, and (c) their contents may not correspond to anything the subject can comfortably be said to believe.

Feature (a) in Stich's characterization poses no problem for present purposes, since it assumes that the states in question can play roles in inference. In any case, some beliefs are arguably much more restricted in their inferential roles than others, and it is to be expected on the basis of their contents alone that early sensory representations, which invariably concern concrete, context-bound particulars represented egocentrically, would exhibit comparatively limited inferential roles.

Feature (b) raises complex issues, but in brief, it is not clear that all representations involved in early and intermediate stages of perceptual processing normally occur without awareness. We are often visually aware of complexes of shapes, oriented edges, and other currently instantiated features of the environment represented in "low-level" vision, though not of the many constraints that seem to guide visual interpretation, such as the defeasible maxim "light comes from above".[18] For present purposes, it does not matter whether the latter are considered dispositional beliefs or subdoxastic states in Stich's sense, so long as transitions among the consciously accessible representations are inferential.

Stich's final point (c) concerns the contents of allegedly subdoxastic states. The point may be put this way: suppose someone verbally expresses the belief "There is an edge oriented 20 degrees from vertical in the upper-left quadrant of my visual field." Surely this person's cognitive state differs from that of one who lacks the precise concepts QUADRANT and VERTICAL, even if this characterization captures the information carried by relevant sensory states. And if the verbally expressible state is what we normally attribute when we attribute belief, the sensory representation must be something else. This argument is compelling, but may suggest merely that the verbally sophisticated subject harbors more complex beliefs than, for example, a dog. Characterizing the content of the sensory state itself still poses a challenge, but this difficulty arises with respect to the cognitive states of nonhuman animals in any case.

More radically, Nico Orlandi (Orlandi 2015) argues that the causal antecedents of percepts are "representations" only in a trivial, too-liberal sense, and are better seen as detectors, indicators or biases, responsive only to their proximal causes. She claims that such states "do not model distal or absent conditions" (p. 19), and that they are not "map-like, pictorial or linguistic representations"

17 Drayson is equivocal, however, about whether this highly contingent, possibly transient form of encapsulation suffices for modularity as traditionally understood (p. 10). Conditional independencies between layers in a Bayesian network can be precisely characterized using the concept of a Markov blanket (Pearl 1988, p. 97, 121), but this form of independence does not seem to amount to information encapsulation in Fodor's sense (Fodor 1983), since as Drayson mentions, "modules" defined in such a way may overlap.

18 This distinction may correspond to that between rapidly changing occurrent representations encoded in neural activities and those encoded in synaptic weight matrices, which implicitly represent regularities over longer time-scales. It is to be expected that the latter contents would not be immediately accessible to consciousness.

(p. 25). They may cause the system to behave according to norms of rationality, but on Orlandi's and similar views, transitions between such states, or from them to propositional attitudes, do not involve representations as premises, and so cannot be truth-preserving transitions.

However, many proponents of the theories under discussion understand representation in terms of the notion of structural homomorphism between model and environment (Hohwy 2013, ch. 8, Gładziejewski 2015), which Orlandi does not consider. On this view, the hierarchical model as a whole functions in a map-like way, even if its parts do not. Representations in the system get their contents in virtue of occupying positions in the system's overall structure analogous to the positions occupied by represented facts or situations in the represented system. The intermediate states in perceptual processing hierarchies structurally represent causes at varying time-scales and thus model distal conditions. They may also model absent conditions, when the hierarchy is employed for non-perceptual purposes such as mental imagery or dreaming, or when illusion or hallucination occur. Moreover, such a system may need to be quite complex in order to have structure that can meaningfully correspond to environmental structure, so the relevant notion of representation is not trivial. A full defense of structural representation is firmly beyond the scope of this paper, but it suffices for present purposes to point out that the representationalist options considered by Orlandi are not exhaustive.

My aim in the preceding two sections has been to sketch a route around *prima facie* conceptual obstacles for the view that perception is the result of inference. The arguments offered have been necessarily both wide-ranging and brief, but I hope that those who disagree may nonetheless grant that the case for perceptual inference hinges on the outcome of these ancillary debates.

### 3.3 A Computational Model

I turn now to computational models in the Helmholtzian tradition. Hinton & Sejnowski (Hinton and Sejnowski 1983) pioneered a model of "optimal perceptual inference" inspired in part by John Hopfield's earlier work (Hopfield 1982). Though different from predictive processing models in important ways, it implements the idea that Bayesian inference can be accomplished via the minimization of potential energy within a generative model in a particularly transparent way. I first briefly describe the dynamics of the network, simply as a physical model of the brain, and then discuss the interpretation of the model as performing statistical inference.

The model consists of a large collection of binary stochastic processing units ("neurons")—that is, units that may be in one of two states, 0 and 1, and whose states at any moment are determined probabilistically, using an activation function that computes the probability of a neuron being in the "1" state as a function of its weighted, summed input (plus a threshold or bias term). Each pair of units is reciprocally connected via symmetric synaptic weights, and a subset of the units additionally can receive input directly from an external information source. The network is otherwise unstructured.

Action potentials in a biological neuron are likely to occur in proportion to the voltage difference that has built up across its cell membrane (the "membrane potential"). This suggests that, as Hopfield (Hopfield 1982, p. 2555) puts it, "the mean rate at which action potentials are generated is a smooth function of the mean membrane potential." This function turns out to be nonlinear, and in particular sigmoidal in shape (Figure 2a). Hinton & Sejnowski's activation function (Figure 2b) models this function from potential energy to firing rate stochastically.

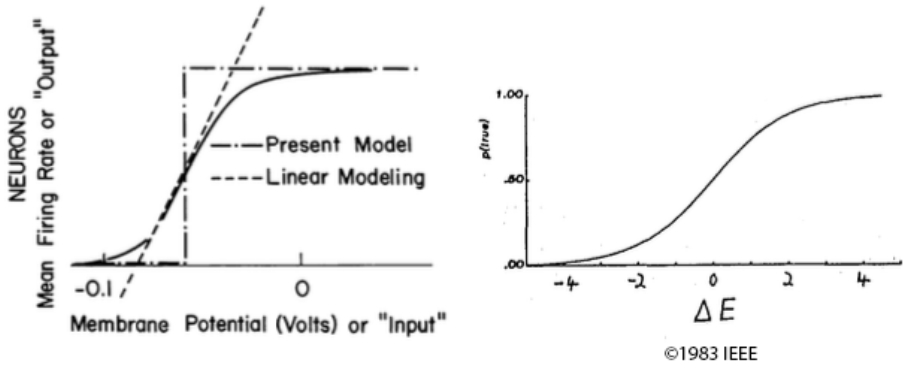


**Figure 2.** (a) The relationship between firing rates and membrane potential (Hopfield 1982, Fig. 1, reprinted with permission of the author). (b) The logistic activation function that maps an artificial neuron's contribution to potential energy to its stochastic state (Hinton and Sejnowski 1983, Fig. 1).

Starting from some initial configuration (choice of "0" or "1" states for all neurons) and assuming a fixed set of weights, the network can be run so as to minimize its potential energy by choosing states for each unit based on the states of its neighbors plus any external input, via the sigmoid function. Neurons receiving large negative (inhibitory) input minimize the potential energy when turned "off", and those receiving large positive input minimize it when firing, as is reflected in the energy equation for the network (Hinton and Sejnowski 1983, Eq. (1)). Significant membrane potentials are local non-equilibrium states, so repeatedly choosing states for the units using the activation function simulates (in a very coarse-grained way) the brain's settling into thermodynamic equilibrium, at least insofar as this process depends on neural spiking activity. If external input is added, the network can simulate the impact of the absorption of sensory signals on this process.

Hinton & Sejnowski propose a simple interpretation of such a network as a representation of the source of its external input signals: the state of each unit is interpreted as the truth-value assigned to a proposition, and the probability of a unit's being in the "True" or "1" state at a given moment corresponds to the probability the system assigns to the corresponding proposition at that moment ("probabilities are represented by probabilities" (p. 448)). The "input" nodes correspond to pieces of sensory evidence, and the rest correspond to hypotheses invoked to explain the evidence. Except for the input units, whose states are overwhelmingly determined by the external information source, the probability of each unit (and thus each hypothesis) depends only on the states of the other units (hypotheses) plus the strength and sign of the connections between them. The network can thus represent complex probabilistic dependencies among hypotheses.

Given this interpretation of the units, the process of updating each unit's state is equivalent to Bayesian inference.[19] Bayes's rule, written in terms of the natural exponential function and the log prior and likelihood ratios, is identical in form to the sigmoid activation function, where the prior ratio for H is implemented by unit *h*'s threshold term and the likelihood ratio of *H* given *E* is implemented by the symmetrical weight between *h* and *e*.[20]

19 Historically, ideas from statistical mechanics were applied to optimization problems on the basis of an analogy between the behavior of physical systems with large numbers of interacting parts and functions with large numbers of interacting variables (Kirkpatrick et al. 1983). The argument pursued here proposes a more direct link between energy minimization and Bayesian inference, exploiting the idea that the mind/brain realizes a complex statistical model in virtue of its complex physical structure.

20 This is a simplified formulation, which ignores direct external input as well as the temperature parameter in the sigmoid function, which is assumed to be set to 1. See Hinton and Sejnowski 1983, Eqs. (2), (3) and (5) and discussion. Note that the sum of weighted input in Fig.3(c) corresponds to using Bayes's rule to update the probability of *H* given multiple pieces of independent evidence.
The authors point out two important issues with this implementation of Bayesian inference: (a) symmetrical weights are required if each unit is to implement inference using only local information, but the relation between evidence and hypothesis described by Bayes's theorem is not symmetrical, and (b) the weights and thresholds must be so designed as to capture the effect of the negation of the evidence on a hypothesis. These issues are surmountable, but I omit the details here (see pp. 450-451 & 453 of the paper).

(a) $$p(H|E) = \frac{1}{1 + e^{-}\left(\ln\left(\frac{p(H)}{p(\sim H)}\right) + \ln\left(\frac{p(E|H)}{p(E|\sim H)}\right)\right)}$$

(b) $$p_h = \frac{1}{1 + e^{-\Delta E_h}}$$

(c) $$\Delta E_h = \sum_e w_{he} s_e - \theta_h$$

(d) $$-\theta_h = \ln\left(\frac{p(H)}{p(\sim H)}\right), w_{he} = \ln\left(\frac{p(E|H)}{p(E|\sim H)}\right),$$
$$p_h = p(H|E)$$

**Figure 3.** Equations adapted from Hinton and Sejnowski 1983. (a) Bayes's rule, expressed in terms of the natural exponential function. (b) The sigmoid activation function used in Hinton & Sejnowski's network (with temperature parameter $T$ omitted), where $p_h$ is the probability of unit $h$ firing, i.e. being in the "True" state and $\Delta E_h$ is the difference between the energy of the network with unit $h$ in the "False" state and the energy with $h$ in the "True" state. (c) $\Delta E_h$ is determined locally for each unit by its weighted, summed input minus its threshold term, where $w_{he}$ is the symmetrical weight between unit $h$ and unit $e$, $s_e$ is the binary state of unit $e$, and $\theta_h$ is the threshold for unit $h$ (the term for direct external input is omitted here for simplicity). (d) Interpreting variables in the model as representing relevant probabilities yields a formal equivalence between energy minimization in the network and statistical inference using Bayes's rule. The weights and biases of the network are interpreted as log probability ratios, while the probability of unit $h$ being "on" is interpreted as the probability assigned to hypothesis $H$.

Since the network's settling into equilibrium can thus be interpreted as massively parallel Bayesian inference, the potential energy of a global state is a measure of the incoherence of the corresponding collection of hypotheses. As Hinton & Sejnowski put it, "The energy of a state can.be interpreted as the extent to which a combination of hypotheses fails to fit the input data and violates the constraints between hypotheses, so in minimizing energy the system is maximizing the extent to which a perceptual interpretation fits the data and satisfies the constraints" (p. 449).

Obviously, there is no reason to suppose that an arbitrary collection of weights will result in a very coherent set of hypotheses. Learning can be accomplished in the network by adjusting the weights so as to minimize the difference between the network's equilibrium states when running independently of external input and its equilibrium states given fixed external input (p. 452). This amounts to fitting a generative model to the data supplied at the input nodes. By supplying sustained input and letting the network settle into equilibrium, the generative model is implicitly inverted.

Since nothing about the content of the units has been assumed beyond bare truth-evaluability, this model supplies a purely formal theory of inductive inference of a different kind than the logical models discussed earlier: rather than appealing to the internal syntax of language-like representations to explain inference, this model appeals to a structurally specifiable notion of coherence, which can be shown to increase as representations are updated. In this respect, it thus realizes Harman's account of inductive inference.

As in Harman's account, the model subsumes confirmation holism as a special case of the holism of inferential processes, since fitting the data is treated as a special case of coherence.[21] Just as observations are in practice accorded proportionally more weight than theoretical assumptions in determining posterior belief, it may be supposed that the incoming sensory signal is generally stronger and more consistent than endogenously generated signals in neural networks like the one discussed here, so that in practice coherence can only be achieved along with empirical adequacy.

There is a loose end concerning acceptances. It was suggested earlier that the same holistic process that results in rational change of belief could result in rational change in acceptance. There are many

21 I thank an anonymous reviewer for pressing me to emphasize this point.

ways in which one might implement this possibility in the sort of model just discussed, but perhaps the simplest is to suppose that hypothetical reasoning makes use of the same representational vehicles that subserve occurrent and dispositional beliefs.

The longer-term knowledge underlying inferential transitions, encoded in synaptic weights, can remain the same in both processes. Entertaining a hypothesis may be a matter of subtly modulating the activity of the units corresponding to occurrent beliefs, which may be supposed to spread activation through the network in a pattern similar to (but perhaps weaker than) one that would occur were one fully committed to the inference. Predictive processing models (as well as a wealth of empirical evidence independent of them) predict in any case that the same neural hardware is used both for perceptual representation and (possibly concurrent) mental imagery. The simultaneous representation of what is believed and what is provisionally accepted may be seen as a generalization of this process to amodal representations.

### 3.4 Extension to Other Models

The model just discussed provides a particularly perspicuous, but highly idealized and abstract, account of the neural implementation of Bayesian inference. Similar mechanisms exist in more sophisticated models, including predictive processing models. For brevity, I'll consider two models with respect to their differences from the one just discussed, and argue that they implement inference in similar ways.

The Helmholtz machine, a model of perception proposed by Hinton, Dayan , Frey & Neal (Hinton et al. 1995), is in some ways an intermediary between the one just considered and predictive processing models. It uses binary stochastic units arranged into layers: an input layer whose states are determined externally and a series of hidden layers. The units in each layer are connected to those in the layer above by "recognition" weights, and to those in the layer below by distinct "generative" weights. These sets of weights (plus biases) implement the corresponding recognition and generative models. Like predictive coding models, this model incorporates hierarchy.

Since the Helmholtz machine encodes a feedforward recognition model in its bottom-up weights, perceptual inference does not require letting the system settle to equilibrium to invert the generative model, but nor does it employ top-down priors dynamically, so it arguably does not implement true coherence-maximizing inference in Harman's sense. Still, the weights can be learned by a process of implicit error minimization, in which the states of the units under the generative and recognition models are used as targets to train one another (Frey et al. 1997, p. 5). This learning process increases the coherence of the sets of representations induced by perceptual input. The creators of the model acknowledge that the lack of a role for top-down influence during perception limits its biological plausibility (while increasing recognition speed) (Frey et al. 1997, p. 21).

The predictive processing model described by Friston (Friston 2005) differs from the Helmholtz machine in at least two major respects: (a) it employs the signature mechanism of online predictive coding, whereby only the difference between top-down predictions and bottom-up error signals is passed up the hierarchy, and (b) it does this by including dedicated nodes that represent prediction errors, as well as recurrent and lateral connections that explicitly encode the variances of the prediction errors.

This model seems to differ markedly from the others in that its nodes represent the magnitudes of environmental quantities and prediction errors rather than binary propositions. This difference may be important in various ways but I argue presently that it does not affect the truth-preserving character and thus the inferential status of the transitions in the model. The difference concerns only what is explicitly encoded, rather than the truth-evaluability of the model's representations.

Bayesian inference in its most general form takes arbitrary probability distributions[22] as inputs and yields a posterior distribution as output. The common use of Bayes's theorem to update the posterior probability of hypothesis *H* given evidence *E* can be assimilated to the more general case by noting that the relevant probabilities yield a Bernoulli distribution over Boolean truth-values of the claims *E* and *H*. In the other direction, a distribution over values of a real-valued variable can be thought of as an assignment of probability to each of a range of hypotheses about the value of that variable.[23] Thus, Bayesian inference in general satisfies the truth-evaluability constraint on inferential processes.

It remains to be seen exactly why we should expect transitions in the more complicated models to preserve truth (and more specifically, to conform at least approximately to Bayesian norms). First, since inference in a hierarchical generative model depends on many coordinated (layers of) nodes, improvement of the generative model guarantees increased coherence among the hypotheses it represents. Second, these models can be understood as variations on the simpler proposal discussed earlier by appeal to their common denominator: the idea that minimization of potential energy is equivalent to improvement of a generative model of the external source(s) of the system's input. The main differences between the models in this respect concern how they represent the distributions that constitute the model.

In brief, the generative distribution in both the Hopfield-inspired model and the Helmholtz machine are encoded in the synaptic weights (plus biases) and activities of the units, since the probability of each unit firing (and therefore the probability assigned to the represented proposition) is determined directly by its weighted, summed input, which also defines the unit's contribution to the energy function for the network. In the latter model, the Helmholtz free energy measures the potential energy of the system in various states, and the two distinct sets of weights make different contributions to this term.[24]

Predictive processing models such as Friston's, as Bogacz 2015 shows, minimize the free energy by adjusting the explicitly represented model parameters (mean and variance) used to define the relevant distributions. As has been widely discussed (see e.g. Clark 2013 and Hohwy 2013, ch. 3), the precision (i.e. inverse variance) of the error units in such models is used to adjust the relative influence of priors and incoming evidence at various levels of hierarchy, so that the posterior means (which implement the inferred hypotheses) are a precision-weighted combination of the prior and likelihood. These models thus in effect use the precision as a way of controlling the tradeoff between the two criteria in Harman's account of induction (conservatism and coherence).[25]

Importantly, this precision-weighting is not a special feature of predictive coding models but falls naturally out of Bayesian inference. Relatedly, as Arnold Zellner (Zellner 1988) shows, Bayes's theorem itself mandates as much conservatism as possible with respect to the total amount of information in the input VS the output of a process. It is an optimal information-processing rule in that it maximally conserves information, subject to the constraint that it produce a probability distribution.[26]

## 4 Conclusion

In summary, I've argued that there is no distinctive sense of 'inference' that covers all uncontroversial, commonsense uses of the term but fails to cover perceptual inference as characterized by Helm-

22 Since I take the difference between continuous and discrete distributions not to affect the main argument here, I use 'distribution' throughout, without meaning to exclude distributions that could only be characterized using a density function.

23 One could interpret a continuous distribution in terms of an infinite range of hypotheses, but in this case the probability assigned to each individual hypothesis would be 0. This is not very useful, but the distribution can be described meaningfully and as precisely as one likes by specifying a set of hypotheses each of which covers an arbitrarily small range of values.

24 This is reflected in equations (3) and (5) in Hinton et al. 1995, which define the cost functions for learning. See Hinton and Zemel 1994 for a general discussion of the connection between Helmholtz free energy and generative models.

25 I owe this point to an anonymous reviewer.

26 Thanks to Wanja Wiese for bringing this point, and the relevant reference, to my attention.

holtzian theories. In the process, I've defended Harman's (Harman 1973) view of inference as coherence-seeking, conservative change in view, as well as a simple descriptive conception of rationality as the tendency to preserve truth in the transitions between one's internal representations. I also argued in some detail that Helmholtzian models of perception implement inference so described. If these models are accurate as rough descriptions of biological cortical networks, and if Harman's conception of inductive inference is defensible, Helmholtz's theory that perception is a form of inference is vindicated.

I close by addressing the extremely general applicability of the Bayesian perspective defended above, which one might fear borders on triviality. Since I've claimed in effect that any system of interdependent variables can be interpreted as a set of propositions, it may be difficult to imagine any regular mental or neural process that would *not* count as inference.

However, the conception of inference on offer here is in the end only as trivial as is the truth-evaluability of the representations over which it is defined. I have suggested that all content is in effect truth-evaluable, and rejected a popular view about what it takes to be a truth-bearer (namely, internal compositional structure), but next to nothing has been said about the determinants of content in this paper, beyond a nod to the notion of structural representation appealed to by predictive processing theorists. That account of representation does not seem to trivialize it, and may even rule out simple feedforward systems as involving genuine representation and therefore genuine inference.[27]

The most radical conclusion one might draw from the line of argument developed here is that sensory representations are to be counted, somehow, among the propositional attitudes. This may seem to strain ordinary usage, but as in the case of inference, a principled claim that a term such as "belief" is being used in an extended sense requires substantive support. Commonsense psychology implicitly defines propositional attitudes in terms of their functional roles, but does not determine *a priori* which representations will turn out to play those roles, and many considerations suggest that representations throughout perceptual and cognitive processing hierarchies function similarly. This is not to say that there are no important differences among such representations. It was conceded above, for example, that the contents of sensory representations are difficult to spell out convincingly. I should also stress that while I see no pressing theoretical need to suppose that Helmholtzian systems exhibit modularity, this is ultimately an empirical question, and the argument here requires only that perceptual inference itself not be encapsulated, not that such systems are in no way modular.

A final reply to worries about triviality touches on a point of philosophical methodology. We are sometimes concerned to preserve the interest of debates by avoiding interpretations of key concepts that render them too broadly applicable, but I am not sure that this avoidance is always desirable. If a relatively deflationary conception of inference is sufficient to capture the uncontroversial features of the commonsense notion, as well as to serve the needs of Bayesian theories in cognitive science (and Helmholtzian theories of perception in particular), it's not clear *a priori* why we should want a more discriminating account. Important generalizations about inference over syntactically structured representations, or conscious, deliberate reasoning, can be captured by specific accounts of those phenomena. A clear conception of the more basic notion of inference can only assist in these endeavors.

27 Since structural homomorphism comes in degrees, it may be doubted whether even representation (so construed) can provide an ironclad criterion for distinguishing inference from simpler mechanisms, or more broadly, minds from non-minds. Perhaps this distinction can be drawn, if at all, by appeal to the sorts of interesting (and increasingly abstract) *contents* that more complex systems with deeper hierarchies can represent.

## References


Aggelopoulos, N. C. (2015). Perceptual inference. *Neuroscience & Biobehavioral Reviews, 55*, 375-392.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences, 14*, 471-517.

Bogacz, R. (2015). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*. http://dx.doi.org/10.1016/j.jmp.2015.11.003.

Boghossian, P. (2014). What is inference? *Philosophical Studies, 169* (1), 1-18.

BonJour, L. (1985). *The structure of empirical knowledge.* Cambridge: Harvard University Press.

Broome, J. (2014). Comments on Boghossian. *Philosophical Studies, 169* (1), 19-25.

Burge, T. (2010). *Origins of objectivity.* Oxford: Oxford University Press.

Carnap, R. (1950). *Logical foundations of probability.* Chicago: University of Chicago Press.

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

Davidson, D. (1986). A coherence theory of truth and knowledge. In E. LePore (Ed.) *Truth and interpretation: Perspectives on the philosophy of Donald Davidson*. Oxford: Basil Blackwell.

Dennett, D. C. (1978). *Brainstorms.* Cambridge: MIT Press.

Drayson, Z. (2012). The uses and abuses of the personal/subpersonal distinction. *Philosophical Perspectives, 26* (1), 1-18.

——— (2017). Modularity and the predictive mind. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Dretske, F. (1981). *Knowledge and the flow of information.* Cambridge: MIT Press.

Evans, G. (1982). *The varieties of reference.* Oxford: Oxford University Press.

Fodor, J. A. (1975). *The language of thought.* Cambridge: Harvard University Press.

———(1983). *The modularity of mind: An essay on faculty psychology.* Cambridge: MIT Press.

———(2007). The revenge of the given. In B.P. McLaughlin & J.D. Cohen (Eds.) *Contemporary debates in philosophy of mind*. Malden: Blackwell.

Frey, B. J., Dayan, P. & Hinton, G. E. (1997). A simple algorithm that discovers efficient perceptual codes. In M. Jenkin & L.R. Harris (Eds.) *Computational and biological mechanisms of visual coding*. New York: Cambridge University Press.

Friston, K. (2005). A theory of cortical responses. *Phil. Trans. R. Soc. B, 360*, 815-836. http://dx.doi.org/10.1098/rstb.2005.1622.

——— (2011). Action understanding and active inference. *Biological Cybernetics, 104*, 137-160.

Goldstein, D. G. & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109* (1), 75-90.

Gładziejewski, P. (2015). Predictive coding and representationalism. *Synthese*, 1-24. https://dx.doi.org/10.1007/s11229-015-0762-9.

Harman, G. (1973). *Thought.* Princeton: Princeton University Press.

———(1986). *Change in view: Principles of reasoning.* Cambridge: MIT Press.

Harrison, L. M., Stephan, K. E., Rees, G. & Friston, K. J. (2007). Extra-classical receptive field effects measured in striate cortex with fMRI. *NeuroImage, 34* (3), 1199-1208. http://dx.doi.org/10.1016/j.neuroimage.2006.10.017.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11* (10), 428-434. https://dx.doi.org/10.1016/j.tics.2007.09.004.

Hinton, G. E. & Sejnowski, T. J. (1983). Optimal Perceptual Inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

Hinton, G. E. & Zemel, R. (1994). Autoencoders, minimum description length and Helmholtz free energy. *Advances in Neural Information Processing Systems, 6*, 3-10.

Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science, 268*, 1158-1161.

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

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA, 79*, 2554-2558.

Huang, Y. & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, n/a-n/a. https://dx.doi.org/10.1002/wcs.142.

Jenkin, Z. & Siegel, S. (2015). Cognitive penetrability: Modularity, epistemology, and ethics. *Review of Philosophy and Psychology, 6* (4), 531-545. https://dx.doi.org/10.1007/s13164-015-0252-5.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, New Series, 220* (4598), 671-680.

Kyburg, H. E. Jr. (1961). *Probability and the logic of rational belief.* Middletown: Wesleyan University Press.

Lehrer, K. (1986). The coherence theory of knowledge. *Philosophical Topics, 14*, 5-25.

Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology, 6* (4), 547-569. https://dx.doi.org/10.1007/s13164-015-0253-4.

Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs, 50* (3), 629–658. https://dx.doi.org/10.1111/nous.12089.

Newen, A., Marchi, F. & Brössel, P. (2017). *Consciousness and Cognition (47): S.I. Cognitive penetration and predictive coding* (pp. 1-112). http://www.sciencedirect.com/science/journal/10538100/47.

O'Callaghan, C., Kveraga, K., Shine, J. M., Adams, Jr. & Bar, M. (2017). Predictions penetrate perception: Converging insights from brain, behaviour and disorder. *Consciousness and Cognition, 47*, 63–74. https://dx.doi.org/10.1016/j.concog.2016.05.003.

Oh, J. & Seung, H. S. (1997). Learning generative models with the up-propagation algorithm. *NIPS 1997*.

Orlandi, N. (2015). Bayesian perception is ecological perception. http://mindsonline.philosophyofbrains.com/2015/session2/bayesian-perception-is-ecological-perception/.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Francisco: Morgan Kaufman Publishers, Inc.

Quilty-Dunn, J. (unpublished). *Phenomenal contrast and perceptual belief.*

Quine, W. V. (1951). Two dogmas of empiricism. *The Philosophical Review, 60* (1), 20-43. http://www.jstor.org/stable/2181906.

Rao, R. P. N. & Sejnowski, T. J. (2002). Predictive coding, cortical feedback, and spike-timing-dependent plasticity. In R. P. Rao, B. A. Olshausen & M. S. Lewicki (Eds.) *Probabilistic models of the brain: Perception and neural function*. Cambridge: MIT Press.

Reichenbach, H. (1949). *The theory of probability: An inquiry into the logical and mathematical foundations of the calculus of probability.* Berkeley: University of California Press.

Rosenthal, D. M. (2005). *Consciousness and mind.* Oxford: Clarendon Press.

Searle, J. (1983). *Intentionality: An essay in the philosophy of mind.* New York: Cambridge University Press.

Siegel, S. (2010). *The contents of visual experience.* Oxford: Oxford University Press.

Stich, S. P. (1978). Beliefs and subdoxastic states. *Philosophy of Science, 45* (4), 499-518.

Tye, M. (1991). *The imagery debate.* Cambridge: MIT Press.

Von Helmholtz, H. (1860/1962). *Treatise on physiological optics.* New York: Dover.

Wright, C. (2014). Comment on Paul Boghossian, 'What is inference'. *Philosophical Studies, 169*, 27-37.

Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician, 42* (4), 278-280.

# (Dis-)Attending to the Body

## Action and Self-Experience in the Active Inference Framework

Jakub Limanowski

Endogenous attention is crucial and beneficial for learning, selecting, and supervising actions. However, deliberately attending to action execution usually comes with costs like decreased smoothness and slower performance; it may severely impair normal functioning and, in the worst case, result in pathological behavior and self-experience. These ambiguous modulatory effects of attention to action have been examined on phenomenological, computational, and implementational levels of description. The active inference framework offers a novel and potentially unifying view on these aspects, proposing that actions are enabled by attentional modulation based on expected precision of prediction errors in a brain's hierarchical generative model. The implications of active inference fit well with empirical results, they resonate well with ideomotor action theories, and they also tentatively reflect many insights from phenomenological analysis of the "lived body". A particular strength of active inference is its hierarchical account of motor control in terms of adaptive behavior driven by the imperative to maintain the organism's states within unsurprising boundaries. Phenomena ranging from movement production by spinal reflex arcs to intentional, goal-directed action and the experience of oneself as an embodied agent are thus proposed to rely on the same mechanisms operating universally throughout the brain's hierarchical generative model. However, while the explanation of movement production and sensory attenuation in terms of low-level attentional modulation is quite elegant on the active inference view, there are some questions left open by its extension to higher levels of action control—particularly about the accompanying phenomenology. I suggest that conceptual guidance from recent accounts of phenomenal self- and world-modeling may help refine the active inference framework, leading to a better understanding of the predictive nature of embodied agentive self-experience.

**Keywords**

Active inference | Action | Attention | Ideomotor theory | Intentional action | Lived body | Minimal phenomenal selfhood | Motor control | Precision-modulation | Self-model

**Acknowledgements:**
I would like to thank Felix Blankenburg, Ryszard Auksztulewicz, Thomas Metzinger, and Wanja Wiese for their helpful comments.

### 1 When Attending to the Body Impairs Performance

A centipede was happy – quite!
Until a toad in fun
Said, "Pray, which leg moves after which?"
This raised her doubts to such a pitch,
She fell exhausted in the ditch
Not knowing how to run.
*Kathrine Craster (1871)*

The "Centipede's dilemma" nicely captures the fact that I am usually not paying attention to my body as I interact with the world. The poem also suggests that this may be a good thing, as such attention—triggered, for example, when one is asked how one coordinates one's many legs—can severely impair one's normal functioning in the world: the centipede certainly *wants* to move, yet it fails to do so because it directs its attention towards its body (i.e., to how the body should execute movements) instead of forgetting about it and just moving as usual. The suppression of the body from experience has been of central interest to classical phenomenology, which considers it a necessary means for

Limanowski, J. (2017). (Dis-)Attending to the Body - Action and Self-Experience in the Active Inference Framework.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 18. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573192

interacting with the world as a "lived body". The lived body concept was proposed by Merleau-Ponty (Merleau-Ponty 1945/1962, and was developed by others, see Gallagher 1986, for a review) to explain, without resorting to Cartesian dualism, the dual role of the body as both an object belonging to the world, and our means (the "vehicle") of being an experiencing and acting subject in this world. In brief, the lived body *is* our being and acting in the world; it therefore is a "lived body-environment" (Gallagher 1986, p. 162). In such equilibrium with the environment, the body is not an object in my phenomenological field—it is absent from my experience.[1] Naturally, the body can be experienced via the senses—but this explicit, often "analytic" access to the objective body reveals "its belongingness to the physical realm" (Legrand 2011, p. 15; cf. Merleau-Ponty 1945/1962; Liang 2015). Husserl , the founder of phenomenology, described this as a "self-objectivation of the lived body" (Zahavi 1994, p. 70). While this need not necessarily imply a total loss of the body's subjectivity (cf. Zahavi 1994), it may lead to an experienced "doubling of the body, the 'splitting of the phenomenon' into two abstractions" (Gallagher 1986, p. 140).

An important postulate of classical phenomenology of the body is that "it is never our objective body that we move, but our phenomenal body" (Merleau-Ponty 1945/1962, p. 106). If we subscribe to this postulate, we can see why directing attention to the body may be detrimental to (inter)acting in the world: self-directed attention presents the body *also* as an object of experience, thus interfering with the normal experience and performance as a lived body-environment under experiential suppression of the physical body. This can happen in two ways: the body can suddenly appear as an object in my phenomenological field, such as when I bump into something, when I am exhausted, or when I am injured (i.e., in "limit situations", Gallagher 1986, p. 148). In these cases, the body-as-object captures my attention. But similar self-objectivation can also be induced deliberately via endogenous self-directed attention, as in the case of the centipede. An extreme example of this is illustrated by the "analytical, decomposing effect" (Fuchs 2010, p. 241) of self-directed attention in schizophrenic hyperreflexivity, where "every action, however trifling, requires targeted attention and action of the will, as it were, a 'Cartesian' impact of the Ego on the body" (p. 247) and "the self is, so to speak, no longer at home in its body" (p. 251). Thus some forms of mental illness may be understood as an extreme case of experiencing the body as an object, which may result in a vicious cycle[2] of increasing "estrangement from oneself" (Fuchs 2010, p. 239)—estrangement from oneself as a lived body. Less extreme, but similar cases include directing attention towards automatic behavior that has not been learned, like falling asleep or being sexually aroused (Fuchs 2010). Such an impairment of performance by self-directed attention also underlies numerous reports of professional athletes who suddenly become unable to perform certain long-mastered movements. A prominent case is former baseball pitcher Steve Blass, who had to quit his career after suddenly and inexplicably losing his ability to throw accurately. Presumably, just like the centipede such athletes start focusing too much on the movement execution itself.

Of course, how attention affects action has also long been a central empirical research question. Research on motor control has demonstrated that endogenous attention is essential for learning, selecting, and supervising actions. However, experiments have also shown that deliberate attention to action execution usually comes with costs like lack of smoothness and slower, step-by-step performance (Norman and Shallice 1986; Diedrichsen and Kornysheva 2015). For example, people perform and learn motor tasks worse when they attend to their execution, whereas performance increases and

1 This has nothing to do with the awareness that one actually *is* a physical body per se, as this fact can be implicitly experienced without directing attention to it, just as I can be aware that I am walking without directing attention to my walking (Merleau-Ponty 1945/1962; Norman and Shallice 1986; Gallagher 1986). There have been some attempts to clarify why being aware of oneself as a physical body does not necessarily imply a suspension of the body's subjectivity (cf. Zahavi 1994). For example, Legrand (Legrand 2011) proposes a distinction between analytic and subjective access to the self-as-object, where only the former implies a disruption of the body's subjectivity by reification. Alternatively, Liang (Liang 2015) distinguishes the first-personal from the third-personal sense of body ownership, where the latter treats the body as an object.

2 There may be a relation to sustained attention directed from the meaning to the carrier of that meaning, as in the case of semantic satiation, where continued fixation or verbal repetition of a word causes the word to lose meaning (Fuchs 2010; cf. Hohwy 2012; Clark 2015).

movement is much smoother when attention is directed away from execution (e.g. Wulf et al. 2001). Detrimental effects are particularly evident when attention is deliberately directed to already well specified (learned) movements (which may be described as a "reinvestment in movement", Brown et al. 2013, p. 421), where such an internal focus of attention may enslave resources and interfere with automatic motor control processes or schemata (Wulf et al. 2001).

In sum, the ambiguous modulatory effects of attention to action—a necessary control mechanism on the one hand, a potentially substantial impairment on the other—have been examined from various perspectives, spanning phenomenological, computational, and implementational levels of description. In the remainder of this paper, I will argue that all of these levels can in principle be accommodated by the active inference framework (Friston et al. 2009), a recent mechanistic account of adaptive behavior as being driven by hierarchical prediction error minimization which is ultimately aimed at occupying unsurprising states, and which appeals to a theory of brain function based on a universal free energy principle (FEP, Friston 2010; cf. Hohwy 2013; Clark 2015). I will first present an explanation of the aforementioned ambiguous effects of attention to movement in terms of active inference, i.e., as attentional modulation at low levels of the motor control hierarchy of the central nervous system. I will then examine the claim that active inference can in principle be extended to *all* levels of action and behavior—thus mapping, for instance, onto concepts like intention and cognitive control. I will argue that the active inference framework may help bridge the various levels at which attention to the own moving body has been investigated, and thus constitutes a very promising basis for an interdisciplinary investigation of embodied agentive self-experience.

## 2 Active Inference: Moving by Attentional Modulation

The FEP is built around the claim that biological agents must maintain homeostasis and must therefore occupy a limited range of states defined by their phenotype. Thus avoiding "surprising" states is the common principle underlying all behavior and cognition (Friston et al. 2009; Friston et al. 2010; Friston 2010). However, the state of the environment (including the organism itself) is hidden from the agent and must be inferred from incoming sensory information. The FEP proposes that the brain performs such inference via probabilistically mapping hidden causes to sensory data in a hierarchical generative model (HGM), where each level encodes conditional expectations ("beliefs") about information in the level below, with the overall hierarchy ultimately modeling the generative process in the environment that causes the current sensory data. By inverting this model, surprise approximated in the form of prediction error[3] can be minimized via model (parameter) update, which is known as *predictive coding* (Friston et al. 2009): ascending data (at the lowest level, actual sensory input) are compared with descending predictions at each level, and only unpredicted data—the prediction errors—are communicated upwards. These errors can be then minimized by changing the model's higher-level beliefs about the causes of this input, which corresponds to perceptual inference (Friston 2010). Predictive coding in the brain therefore emerges as a consequence of the imperative to maintain homeostasis, whereby priors may be acquired and optimized by learning, or be innate and optimized by natural selection (Friston et al. 2010; Pezzulo et al. 2015).

For inference to be optimal, the brain needs to decide which prediction errors are currently most relevant, and it needs to assign these errors relatively more weight in determining inference. According to predictive coding, this is implemented by adjusting the gain of prediction error units according to their expected precision (which corresponds to reliability or inverse variance). Thus there are two types of descending predictions: those of input, inhibiting error units in the level below, and those of precision, optimizing the gain of error units (i.e., changing the postsynaptic response of error units to their presynaptic inputs, presumably via NMDA-dependent plasticity, dopaminergic modulation,

3 "Surprise" corresponds to the negative log-likelihood of the sensory data under the model and can be approximated via free energy, which, under some simplifying assumptions made by the predictive coding scheme, corresponds to prediction error (Friston et al. 2010).

or other classical neuromodulators; Friston et al. 2012a; Adams et al. 2013). Precision-modulation is thus also a Bayes-optimal mechanism that minimizes free energy. Weighting prediction error signals by their expected precision determines their relative impact on inference, i.e., on updating prior beliefs at higher levels of the model (Friston et al. 2009). Applied throughout a HGM's hierarchy, this mechanism allows for delicately balancing the relative influence of sensory evidence and prior beliefs on (active) inference. This top-down modulation is a contextual one that will vary depending on the current circumstances and requirements: "When higher levels have greater precision, their contextual influence dominates; whereas, when expected sensory precision is high, inference and subsequent behavior is driven by sensory evidence" (Pezzulo et al. 2015, p. 24). Note that the top-down, context-dependent selection and weighting of (sensory) prediction errors, based on their expected precision, is nothing other than weighting specific sensory channels according to (expected) signal-to-noise ratios, the function generally attributed to *attention* (Feldman and Friston 2010; Auksztulewicz and Friston 2016). Under active inference, precision-modulation is therefore described as an attentional modulation (e.g. Edwards et al. 2012; Brown et al. 2013). Crucially, such attentional modulation also determines whether an agent resorts to perceptual inference as described above—or whether it acts.

Of course, not only can we suppress prediction error by changing our model so that it better reflects the state of the world, we can also change the state of the world so that our sensory input corresponds to our current predictions. By acting on the environment, i.e., by intervening with the generative process itself, we can directly suppress surprise (i.e., free energy) and thus also minimize prediction error. *Active inference* (Friston et al. 2009) thus extends the principles of predictive coding (as descdribed above for sensory systems) to the motor system—the difference is that in the motor system, the predictions and errors are proprioceptive, i.e., they are about the posture and position of the body's joints and the forces applied to them (Adams et al. 2013; Friston et al. 2012a). Active inference therefore explains motor control in terms of predicting states of the body as part of the "environment", i.e., of the hidden process generating the current sensory data.[4] Movement accordingly occurs because high-level multimodal or amodal beliefs predict counterfactual exteroceptive and proprioceptive states (sensory consequences that would ensue if the movement were performed), and the proprioceptive predictions generate a prediction error in the spinal cord where they meet afferent information about the current proprioceptive state, i.e., movement is predicted but not sensed. Unlike sensory systems, where the predictions would now be revised to explain away prediction error, the motor system can use an alternative strategy to suppress errors: the fulfillment of proprioceptive predictions by activation of alpha motor neurons of classical reflex arcs in the spinal cord, i.e., by performing the predicted movement (Friston et al. 2010; Friston et al. 2011; Adams et al. 2013; Brown et al. 2013; Edwards et al. 2012). Thus movement results from predictions about its sensory consequences rather than from motor commands in the classical sense.

Under active inference, goal states for action and behavior are defined by prior expectations. Motivated or adaptive behavior can therefore be described as based on the minimization of interoceptive prediction error (which informs about deviance from optimal homeostatic levels) and proprioceptive and exteroceptive prediction error (which specifies the external goal state to be attained by action, Pezzulo et al. 2015). An important implication of this, one which distinguishes active inference from previous approaches, is that goal states are not desirable because they are "valuable" in themselves, but because they are states that the organism *expects* to occupy (under the assumption that it will always minimize free energy). In other words, we selectively sample sensory input that is expected, based on the predictions of our current HGM, to be precise—thus action and perception are intimately coupled (Friston et al. 2009; Friston et al. 2011; Friston et al. 2012b; Hohwy 2012; Pezzulo et al. 2015).

4 One could say, more specifically, that defining the environment in this sense includes the physical body but not the brain that actually employs the HGM, as the brain's states are (to our knowledge) not accessible to itself via sensory organs. Some interesting questions that follow from this are discussed by Metzinger (Metzinger 2017).

Crucially, whether or not action occurs is determined by an attentional balancing act, i.e., attention weights prediction errors to optimize not only perceptual inference, but also action. Movement only occurs if the proprioceptive prediction errors at the spinal cord level generated by confident high-level sensorimotor expectations are expected to be very precise, and if simultaneously the expected precision of ascending sensory prediction error (which conveys evidence against the prediction that one is moving) is attenuated. Only then are proprioceptive prediction errors at the spinal level acted out instead of being accommodated by perceptual inference, i.e., only then are the counterfactual predictions about the body's state fulfilled rather than updated (Friston et al. 2009; Friston et al. 2011). Thus, in addition to confident beliefs about the sensory consequences of the intended movement, "action requires […] targeted dis-attention" away from current sensory evidence that one is actually not moving (Clark 2015, p. 217).

Within this framework, the detrimental effects of attention towards movement execution—remember the centipede—are readily explained in terms of low-level precision-modulation (Brown et al. 2013; Edwards et al. 2012): attending to the sensory input generated by my body increases the precision of the corresponding sensory prediction errors, which are conveying evidence contra the descending predictions of the sensory consequences generated by movement. These errors have now more influence on higher-level beliefs, which are therefore adjusted to accommodate the fact that I sense no movement. Consequently, no sufficiently precise proprioceptive prediction errors are generated, and no (or abnormal) movement results (Adams et al. 2013). Therefore, a system following active inference is only capable of producing movement under an appropriate balance between precision at high versus low levels; under abnormal precision-estimation, pathological behavior ensues, with effects varying according to the hierarchical site of the imbalance (Brown et al. 2013; cf. Edwards et al. 2012; Friston et al. 2012a).

In sum, the act of balancing expected precision at various levels of the generative model determines whether a system operating on such a model resorts to perceptual inference or to action. An important (and prima facie counterintuitive) implication of active inference is that such attentional control is not an action but a part of perceptual inference—it is optimization of precision in a HGM that "has no notion of action; it just produces predictions that action tries to fulfil" (Friston et al. 2009, p.4). Action or behavior—a change of external states—emerges only at the lowest level of the motor hierarchy as a suppression of precise proprioceptive prediction error by peripheral neurons (the central nervous system is only concerned with perceptual inference) and an attenuation of the expected precision of ascending sensory prediction error. Describing the underlying precision-modulation as (endogenous) attentional modulation implies the Jamesian characterization of attention as something selective that "implies withdrawal from some things in order to deal effectively with others" (James 1890, p. 404). Specifically, to be able to interact with the world, I need to withdraw attention from my body's current state and focus it on what I predict sensing in my desired state. This conclusion is very similar to that of classical phenomenology, namely, that the "experiential absence" of the body is necessary for action in the world—being a lived body-environment—and that attention directed towards the objective body is detrimental to normal performance. So active inference intuitively explains why the centipede cannot move in terms of specific effects of low-level attentional (precision) modulation. But does this mechanism likewise explain why the centipede can normally move as it wishes?

## 3 Attentional Modulation throughout the Hierarchical Generative Model: A Motor Control Hierarchy

One of the greatest strengths and boldest claims of the active inference framework is its proposed universal mechanism operating across all levels of the HGM, which is neurobiologically implemented via predictive coding in the brain (and action via reflex arcs). It acknowledges the hierarchical nature of motor control which spans from kinematics to conceptual knowledge about the world; it integrates

distinct control systems (Friston 2011; Pezzulo et al. 2015), and it avoids the pitfalls of describing action control either as purely stimulus-driven, or in purely "perceptuo-motor" or "associative" terms (Ondobaka and Bekkering 2012; cf. Kilner et al. 2007). Active inference thereby fundamentally relies on the top-down contextualizing effect of higher levels on lower ones, enabled by attentional modulation based on expected precision, where a context can be a selected action, a goal, or even agency. Thus it aims at explaining phenomena across all levels of the motor hierarchy, from the reflex arcs that produce movement to intentional action and cognitive control (Pezzulo and Cisek 2016).

### 3.1 Sensory Attenuation and Agency

A particularly interesting implication of the active inference account is its explanation of sensory attenuation, which can be observed during movement (Blakemore et al. 1998; Brown et al. 2013) and even during movement preparation (Voss et al. 2006). The attenuation of self-generated sensory signals during movement has previously been proposed in terms of forward models that predict and thus cancel out the sensory consequences of one's movements based on the body's current state and corollary discharge (Blakemore et al. 1998; cf. Friston et al. 2012b for a more detailed comparison of these accounts). However, the implications of attentional balancing across the motor hierarchy as assumed by active inference go beyond this: as noted above, sensory attenuation is a necessary dis-attention away from sensory input, which would otherwise bias perceptual inference and potentially preclude movement (as is likely the case in the centipede's dilemma). Active inference even postulates that sensory attenuation and its effect on perceptual inference underlies certain forms of self-consciousness, including the experience of self-other distinction in action execution versus observation. Similarly to previous accounts of the mirror neuron system, active inference assumes that the brain uses the same HGM and thus the same action control hierarchy to model and predict the intentions, goals, actions, and kinematics of both one's own and other bodies (Kilner et al. 2007). This means that high-level beliefs encoding action goals "do not assign agency to any particular agent" (Friston et al. 2012b, p. 539): these beliefs generate amodal, multimodal, and unimodal predictions throughout the motor hierarchy for one's own and for others' actions.

According to active inference, self- or other-agency—whether I perform a movement or whether I instead perceive someone else performing the movement—is a *context* determined by precision-modulation of (i.e., selective attention to) proprioceptive and visual information in one and the same HGM (Friston et al. 2011). If I observe an action, the visual prediction error generated by the seen movement will update multimodal beliefs at higher levels in the motor hierarchy, which predict visual and proprioceptive action consequences. This means I must attenuate the expected precision of the prediction errors generated by *proprioceptive* predictions—otherwise I might move myself. With this attenuation, updating my model's beliefs by visual prediction errors allows me to infer the cause of the observed movement and thus ultimately to understand the other's intentions (Kilner et al. 2007; Friston and Frith 2015). Conversely, recall that increased high-level proprioceptive precision is necessary to produce movement via spinal reflex arcs. Thus "active inference presents in one of two modes; either attending to sensations or acting during periods of sensory attenuation" (Friston and Frith 2015, p. 398), where attentional modulation is fundamentally involved in realizing both of these modes.[5]

Correspondingly, misattributions of agency, as in schizophrenia, have been explained by aberrant attention. Here, inference about the hidden causes of sensations fails because the precision of high-level beliefs is (falsely) increased to compensate for a failure to attenuate sensory prediction error during action. These overconfident beliefs generate additional, incorrectly confident, predictions about external causes—the agent is not able to infer whether it caused its sensations itself, or whether someone or something else caused them (Brown et al. 2013). In sum, under active inference, agency is grounded

5 However, the *sense* of agency in action certainly also depends on behaving in accordance with (confidently) expected states, i.e., when precise proprioceptive prediction error is resolved in line with our predictions (Hohwy 2007; Hohwy 2013; Friston et al. 2013; Clark 2015).

in the contextual influence of high-level beliefs on lower levels, which manifests itself in attentional modulation, i.e., in adjusting the relative gain of vision and proprioception.

## 3.2 Intentional Action as Adaptive Behavior

So far, we have seen that active inference provides an elegant explanation for the role of precision-modulation (attentional biasing) in movement initiation and production. However, active inference aims to explain all facets of behavior. Therefore even complex phenomena like the conscious selection of actions based on goals and intentions should be explained as driven by beliefs about behavior and modulated by expected precision. The proposed answer that active inference offers to these questions is partly reminiscent of that of the classical ideomotor theory (IMT) of action, which was developed as an explanation for how intentions might drive actions (James 1890; Stock and Stock 2004; Kunde et al. 2007). Here, I will briefly outline some commonalities and differences between IMT and active inference, which will reveal the novel contribution and explanatory power of active inference, but also some questions that it leaves open.

Most people would probably agree that an intentional action is always accompanied by a conscious goal representation (cf. Hommel 2015). IMT[6] proposes that this conscious goal representation is in fact driving the action. Movement is accordingly brought about by an "idea" or "effect image" of the anticipated sensory consequences of that movement, which is itself the result of previous associative learning between movements and their sensory consequences (Hommel et al. 2001; cf. Stock and Stock 2004, for a review). Consequently, IMT states that, rather than there being separate perceptual representations and motor commands, perception and action share a common representational format (Prinz 1997; Hommel et al. 2001), just as the active inference view does not distinguish between perceptual and motor representations in the classical sense. An interesting conclusion of IMT is that even the simplest actions are goal-directed, as they are always aimed at reaching an anticipated sensory effect (the "goal representation", Kunde et al. 2007; Hommel 2015). The same holds for active inference, where goal representations are the result of perceptual inference and correspond to (counterfactual) beliefs about sensory states that elicit corresponding prediction errors. Like active inference, IMT emphasizes that actions can only be brought about by ideas if one ignores "competing" ideas—most notably, the fact that one is currently not moving (James 1890; Clark 2015). In sum, both IMT and active inference state that a withdrawal of attention from movement execution and a focus onto the action goal is essential for action, thus nicely explaining the Centipede's dilemma.[7]

Active inference, however, specifies its claim that movement relies on both confident beliefs and attenuated sensory input by suggesting an underlying attentional modulation, implemented by increasing high-level precision and decreasing low-level precision. The universal role of attention proposed by active inference, however, seems at odds with some extensions of IMT. For example, Hommel et al. 2001 differentiate between attentional and intentional weighting in perception and action:

> With reference to perception, feature weighting may be called an *attentional* process, inasmuch as it selectively prepares the cognitive system for the differential processing of relevant (i.e., to-be-attended) and irrelevant (i.e., to-be-ignored) features of an anticipated perceptual event. With reference to action planning, however, the same kind of feature weighting could rather be called an *intentional* process, because it reflects the perceiver/actor's intention to bring about a selected aspect of the to-be-produced event. (Hommel et al. 2001, p. 864)

6 There are of course many variations of IMT; here I will only present the basic assumptions of its classical form.

7 Experimental work has shown that even visual space is attentionally structured in this way: whereas attentional processing is facilitated in peri-hand space, it is impaired on the hand's surface (Taylor and Witt 2014). The explanation may be the same: the brain prevents attention to the body to assist goal-directed action.

Active inference, in contrast, specifies the intentional process as *attention to* intention (Edwards et al. 2012), where an intention is specified by a high-level goal representation. In fact, attention to action intention increases brain activity in supplementary motor areas that under active inference encode intentions (Lau et al. 2004). Active inference thus subscribes to James' proposal that "attention *creates* no idea; an idea must already be there before we can attend to it" (James 1890, p. 450). However, it puts attention (i.e., precision-modulation) at the center of intentional action selection. In conclusion, under active inference, goal states and intentions are *defined* by high-level beliefs, and *selected* by attention (i.e., certain beliefs are assigned more precision, cf. Friston et al. 2011; Pezzulo and Cisek 2016).

In this light, I tentatively propose, precision-optimization at higher levels of the HGM maps onto concepts like "will" (defined as "the direction of action by direct conscious control through the supervisory attentional mechanism", Norman and Shallice 1986, p. 24) or "cognitive control" (defined as the "ability to guide one's behavior in line with internal goals", Jiang et al. 2014, p. 31). In fact, active inference's tenet that attentional allocation is based on predictions of precision is similar to the proposals of some Bayesian accounts of cognitive control, where "the regulation of cognitive control should be considered as a process of *predicting* the optimal amount of cognitive control required in a given context" (Jiang et al. 2014, p. 35). Intuitively, concepts like will or cognitive control imply an important function of attention in directing intentional behavior in line with one's goals—sometimes, for example under distraction or uncertainty, such direction of actions will be notably harder. Classical theories of the relationship between attention and action have correspondingly suggested that "will varies along a quantitative dimension corresponding to the amount of activation or inhibition required from the supervisory attentional mechanisms" (Norman and Shallice 1986, p. 24). Conversely, put in the vocabulary of IMT, in situations where there is no competing "idea", there is also no need for "will" (James 1890).[8] Active inference likewise proposes that high-level precision is especially important under "cognitive conflict", for example, in situations where multiple representations have high precision (e.g., at high and low levels simultaneously, Pezzulo et al. 2015). In such cases, the brain needs to weight a certain belief (goal representation) more strongly than other beliefs and/or more strongly than sensory evidence. The voluntary allocation of attention against the "resistance" of some other precise belief or sensory evidence could explain why we experience an accompanying sense of effort in these situations (Metzinger 2017).

So on the one hand, conscious experience (of will and effort) could co-vary with computational cost of attentional allocation. On the other hand, however, conscious experience and top-down attentional control need not always correspond: in certain functional motor symptoms, for example, movements are executed but feel involuntary. Active inference accounts of such pathological behavior (Edwards et al. 2012) explain it as resulting from the generation of abnormally confident intermediate-level beliefs. These beliefs are sufficiently high-level to generate complex movements, but are still below the levels associated with representing the intention to move. Thus movements are induced, however, are not inferred to have been intended because the resulting percepts are not predicted by higher levels. Hence, although these movements are produced by voluntary (top-down) attention, they do not feel voluntary. This explanation aligns with previous observations that there are "cases in which one experiential sense of 'automatic' does not correspond to 'automatic' in the operational sense" (Norman and Shallice 1986, p. 19), and so some action may seem automatic while actually involving volitional attentional top-down control. For the centipede, the reverse case seems to be true: it does not *want* to be immobile—it wants to move!—but its voluntary attentional allocation prevents this.

Like any other account of action, active inference now faces the challenge to explain which aspects of motor control are accessible to conscious experience, and why. Recent extensions of the IMT have, for example, dropped the assumption that the action-driving ideas or goal representations must be conscious, and do not consider conscious experience to play a causal role in action control (Hommel et

8 James explicitly distinguished between "ideo-motor" and "willed" acts (cf. Norman and Shallice 1986).

al. 2001; Prinz 1997). Their conclusion is that voluntary action may well be possible without conscious experience (Hommel 2015). Active inference offers a convincing mechanistic theory of attention as precision-optimization during perceptual inference and action. Early accounts linking attention and motor control suggested that "the phenomenology of attention can be understood through a theory of mechanisms" (Norman and Shallice 1986, p. 25). However, while attentional modulation as part of active inference very elegantly explains low-level phenomena like sensory attenuation, its extension to higher-level phenomena such as intentional action does not (yet) immediately accommodate the *phenomenology* of attentional allocation in action control. Therefore, some explanatory work remains to be done if active inference is to fully explain all aspects of agentive self-experience: which aspects of volitional behavior are accessible to consciousness, and how does the phenomenology associated with, for example, attentional agency and conscious volition emerge from the proposed brain mechanisms? As one starting point, a valuable contribution, in the form of conceptual guidance, can come from analytical approaches to phenomenal self- and world-modeling.

## 4 Active Inference and Phenomenal Self-Modeling

One such candidate complementary account is self-model theory (SMT, Metzinger 2004; Metzinger 2009). SMT is based on the assumption that the experience of being a self emerges in organisms or systems because they possess an internal model of the world that includes and is centered on the organism itself, which, through identification of the model with its content, experiences phenomenal selfhood (Blanke and Metzinger 2009). Such a model is therefore called a phenomenal self-model (PSM, Metzinger 2004; Metzinger 2009). There are striking commonalities between the assumptions of SMT and active inference (Limanowski and Blankenburg 2013; Hohwy 2013; Metzinger 2014). Most notably, SMT suggests a hierarchy of phenomenal self-modeling, ranging from pre-reflective, "minimal" self-representations like a first-person perspective, body self-identification, or spatiotemporal self-location (Blanke and Metzinger 2009) to complex cognitive self-representations (Metzinger 2017). Such self-modeling can be well described in terms of active inference, whereby the "self" (in all its cognitive-to-minimal dimensions) is a sophisticated hypothesis about the organism's environment which is generated by the brain's HGM, and which tries to maximize evidence for its own existence (Limanowski and Blankenburg 2013).

The SMT account also proposes an important universal "second-order" function of attention operating on a PSM, but a slightly different one than on the active inference view: the attentional absence or inaccessibility of certain processing stages of self-modeling determines the phenomenal transparency of the respective conscious mental representations[9] and the associated experience of presence or realness (Metzinger 2004). Thus mental representations are transparent because only their content, and not their vehicle (e.g., the brain processes at earlier stages underlying this representation) is accessed by attentional introspection. However, not all mental representations are fully transparent. Rather, they can become more or less transparent: the more a system in possession of a PSM can attentionally access earlier processing stages, the less transparent (or more opaque) the representation becomes. This means the representation is recognized *as* modeled: as an internal, self-generated and mind-dependent construct, rather than as an invariant property of the world (Metzinger 2004; Metzinger 2009). Transparency is thus also a "phenomenal signature of epistemic reliability" (Metzinger 2014, p. 124; cf. Seth 2015), i.e., it is a sign of the system's certainty that it has identified something that is real. Conversely, if parts of one's PSM become opaque, this indicates the need to question their realness, and a possible need to revise one's self-model.

Importantly, the SMT thereby assumes a "gradient of realness in the human self-model, with the bodily self being perceived as real and present while the cognitive self-model is experienced as comprised of representations" (Metzinger 2014, p. 123). So whereas I am (or can be) attentionally aware

9 This does not apply to unconscious representations; in the following, only conscious representations are referred to.

that my conception of myself as an industrious person is actually made (up) by my mind, I usually do not conceive of minimal aspects of myself as an embodied self in this way—these representations are in this sense transparent to me. In other words, while I can easily change some cognitive conceptions of myself, changes to pre-reflective representations at lower levels of my PSM like body self-identification are far more difficult to make, and I believe they may have far more severe consequences. Note that although it is certainly possible to update such lower levels of bodily self-representation, as for example one's perceived arm position in the rubber hand illusion, this does not imply that the *realness* of the content of the underlying (still transparent) self-representation is questioned—even in the rubber hand illusion, the assumption of body-self identification holds: I still feel like a normal body with just one, not two right arms (Limanowski 2014; Hohwy 2013). However, I would speculate that even minimal self-representations, i.e., those aspects of minimal phenomenal selfhood eventually constituting the basic, bodily-founded self-experience (Blanke and Metzinger 2009) can (partly) lose their transparency. I further think that such a loss of transparency at low levels of the PSM may result in (usually temporary and reversible) pathological experience—in the worst case, I would cease to "be" a self (Metzinger 2004).

In this way, SMT offers another conception of the detrimental effects of self-directed attention onto the bodily foundations of being a self (a lived body). The SMT view proposes that transparent self-modeling gives us the feeling of "being there" in the world (Blanke and Metzinger 2009; Metzinger 2004; Seth et al. 2011; Limanowski and Blankenburg 2013; Limanowski 2014). Hence, speculatively, the phenomenal absence of the body-as-object can in SMT be conceived of as a form of transparence of the representations underlying minimal phenomenal selfhood. Conversely, once I attend to the body, these phenomenal representations may gradually (but will not necessarily[10]) become opaque—one could rephrase this in classical phenomenological terms as a partial loss of the body's experiential absence. The result could be a state in which the system experiences a certain representation of the self as opaque, because it recognizes, for example, that the link between itself and this particular physical object that is the body is actually "made up" by itself. Following Metzinger's theory, the phenomenological prediction for such states is an experience of "de-identification" as occurring, for example, in depersonalization disorder (Metzinger 2004; Metzinger 2009). A similar experience is also tentatively suggested by the phenomenological description of schizophrenic hyperreflexivity as an abnormal "reification" of the embodied self by self-directed attention, whereby "the transparency of the bodily medium gets lost" (Fuchs 2010, p. 242). Under such lost transparency of the lived body, "aspects of oneself are experienced as akin to external objects" (Sass and Parnas 2003, p. 427). Perhaps one could say that under pathological self-directed attention the body becomes more present and "real" as a physical object, just as pain becomes more present when attended to. Less dramatically, this could apply to cases of attention to movement execution, where transparent, or perhaps even unconscious representations become conscious and (partly) opaque due to endogenous attention, which interferes with normal, fluent performance.

However, what SMT can most notably contribute to active inference is an analysis of the phenomenology accompanying various levels of action control. For instance, the transparency gradient assumed in the PSM may help us understand why some of active inference's proposed attentional modulations are very intuitive (e.g. it is very intuitive that attending to the action goal is necessary to act; goal representations are high-level, and may even be opaque, cf. Gallese and Metzinger 2003), whereas others are not so easy to grasp (e.g. low-level sensory attenuation: do I really volitionally ignore my body during movement initiation?). This could also help explaining why, although attentional modulation is functionally the same across various levels of the HGM, there is *phenomenologically* a substantial difference between whether I attend to the external world or to my bodily self (Metzinger

10 Of course, SMT also entails the classical notion of attention as sharpening the representation of what is currently relevant; attending to a sensation can potentially increase transparency and also the "realness" of the resulting percept. However, in this case attention is directed to the content of the representation, not to the fact *that* the sensation is the content of a representation implemented in the brain (Metzinger 2004).

2017). SMT likewise tells us why agents act as if there are desirable goals in the world where there are really just goal representations in the agent's HGM: via transparent phenomenal modeling, the agent arrives at the conclusion "that goals, actions, and intending selves actually belong to the basic constituents of the world that it is internally modeling" (Gallese and Metzinger 2003, p. 366; however, certain goal representations can also be opaque). Recall that according to active inference, the self-other agency distinction relies on a contextual manipulation of the influence of complex higher-level beliefs on visual and proprioceptive modalities. By assuming that such high-level representations as well as the representation of the attentional allocation process itself may be transparent, SMT provides an explanation of why a system operating via hierarchical inference *experiences* itself as an agent—or conversely, why it experiences another agent as the cause of its current sensory data. Thus conscious volition emerges when an agent integrates a goal representation as an object within a "model of the phenomenal intentionality relation", a representation of an asymmetric subject-object relationship, i.e., of "a system being *directed* at a goal state" (Metzinger 2017; Gallese and Metzinger 2003; Metzinger 2004). Attentional agency, on the other hand, is a fully transparent representation "of the process of selecting the object component for attention" (Gallese and Metzinger 2003, p. 374); it is the experience that results from identification of the agent as a whole with a particular self-representation as an "epistemic agent" (Metzinger 2017).

In sum, SMT accommodates many overlaps between the active inference framework and phenomenological analysis of bodily self-experience, and with its conceptualization of phenomenal transparency versus opacity of conscious mental representations offers a compelling complement. Therefore SMT also opens up alternative ways of addressing some open questions within the active inference framework. Active inference, conversely, offers a neurobiologically plausible implementation of hierarchical self- and world-modeling, including specific testable hypotheses about recurrent message-passing and precision-modulation in the brain. Phenomenological questions have already been addressed using this approach, for example, explaining the loss of a sense of presence due to imprecise interoceptive prediction errors (Seth et al. 2011; Seth 2013; see also Limanowski 2014; Liang 2015 for related discussions of phenomenological implications of experimental paradigms that rely on the direction of attention to specific features of the bodily self). A joint effort of active inference and SMT could be extremely useful in understanding how and why certain aspects of volitional action are conscious, and in the long run, understanding the embodied agentive self-experience in general.

## 5 Conclusion

Most of us will have experienced beneficial and detrimental effects of attention to action to some degree. Not surprisingly, the role of attentional modulation in action control—more generally, in the experience of being an embodied agent in the world—has attracted the interest of philosophers, phenomenologists, psychologists, and neuroscientists alike. This interest has resulted in many hypotheses being proposed, but has also opened up many questions. Active inference, as implemented in the brain via predictive coding, offers a very elegant mechanistic- and implementational-level explanation of adaptive behavior—ultimately, as the result of a system trying to maintain its states within unsurprising boundaries. Active inference describes what happens in the brain of the centipede when it, despite wanting to move, cannot, due to increased attention to sensory prediction errors that preclude fluent movement generation. Thereby active inference proposes attention as a mechanism that balances between the relative impact of prior beliefs and current sensory evidence on inference, thus explaining a range of empirical and phenomenological observations of both normal and pathological behavior. This explanation acknowledges the fundamental role of the body for being an agent in the world, while also emphasizing the body as being part of the to-be-predicted environment. This is very much in line with classical phenomenology's interpretation of the experiential absence of the body-as-object in the subjectively lived body-environment. The extension of the active inference account to higher levels

of action control, however, leaves open some questions about the accompanying agentive self-experience, i.e., the phenomenology of, for instance, volition or attentional agency. Here, a joint application of active inference-based views and analytical accounts of phenomenal self- and world-modeling can lead to conceptual refinement and a correspondingly enhanced understanding of the predictive nature of action and self-experience.

## 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.

Auksztulewicz, R. & Friston, K. (2016). Repetition suppression and its contextual determinants in predictive coding. *Cortex, 80*, 125–140.

Blakemore, S.-J., Wolpert, D. M. & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience, 1* (7), 635–640.

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

Brown, H., Adams, R. A., Parees, I., Edwards, M. & Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing, 14* (4), 411–427.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

Diedrichsen, J. & Kornysheva, K. (2015). Motor skill learning between selection and execution. *Trends in Cognitive Sciences, 19* (4), 227–233.

Edwards, M. J., Adams, R. A., Brown, H., Pareés, I. & Friston, K. J. (2012). A Bayesian account of 'hysteria'. *Brain, 135* (11), 3495–3512.

Feldman, H. & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4*, 215.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11* (2), 127–138.

——— (2011). What is optimal about motor control? *Neuron, 72* (3), 488–498.

Friston, K. & Frith, C. (2015). A duet for one. *Consciousness and Cognition, 36*, 390–405.

Friston, K. J., Daunizeau, J. & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PloS One, 4* (7), e6421.

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

Friston, K., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics, 104* (1-2), 137–160.

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. E. & Bestmann, S. (2012a). Dopamine, affordance and active inference. *PLoS Comput Biol, 8* (1), e1002327.

Friston, K., Samothrakis, S. & Montague, R. (2012b). Active inference and agency: Optimal control without cost functions. *Biological Cybernetics, 106* (8-9), 523–541.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T. & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience, 7*.

Fuchs, T. (2010). The psychopathology of hyperreflexivity. *The Journal of Speculative Philosophy, 24* (3), 239–255.

Gallagher, S. (1986). Lived body and environment. *Research in Phenomenology, 16*, 139-170.

Gallese, V. & Metzinger, T. (2003). Motor ontology: The representational reality of goals, actions and selves. *Philosophical Psychology, 16* (3), 365–388.

Hohwy, J. (2007). The sense of self in the phenomenology of agency and perception. *Psyche, 13* (1), 1–20.

———(2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3*.

———(2013). *The predictive mind.* New York: Oxford University Press.

Hommel, B. (2015). The sense of agency. In P. Haggard & B. Eitam (Eds.) (pp. 307-326). New York: Oxford University Press.

Hommel, B., Müsseler, J., Aschersleben, G. & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24* (05), 910–926.

James, W. (1890). *The principles of psychology.* New York: Holt.

Jiang, J., Heller, K. & Egner, T. (2014). Bayesian modeling of flexible cognitive control. *Neuroscience & Biobehavioral Reviews, 46*, 30–43.

Kilner, J. M., Friston, K. J. & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing, 8* (3), 159–166.

Kunde, W., Elsner, K. & Kiesel, A. (2007). No anticipation–No action: The role of anticipation in action and perception. *Cognitive Processing, 8* (2), 71–78.

Lau, H. C., Rogers, R. D., Haggard, P. & Passingham, R. E. (2004). Attention to intention. *Science, 303* (5661), 1208–1210.

Legrand, D. (2011). Oxford handbook of the self. In S. Gallagher (Ed.) (pp. 204-227). New York: Oxford University Press.

Liang, C. (2015). Self-as-subject and experiential ownership. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570030.

Limanowski, J. (2014). What can body ownership illusions tell us about minimal phenomenal selfhood? *Frontiers in Human Neuroscience, 8*.

Limanowski, J. & Blankenburg, F. (2013). Minimal self-models and the free energy principle. *Frontiers in Human Neuroscience, 7*.

Merleau-Ponty, M. (1945/1962). *Phenomenology of perception (trans. Colin Smith)*. London & New York: Routledge & Kegan Paul.

Metzinger, T. (2004). *Being no one: The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

———(2009). *The ego tunnel: The science of the mind and the myth of the self.* New York: Basic Books.

———(2014). How does the brain encode epistemic reliability? Perceptual presence, phenomenal transparency, and counterfactual richness. *Cognitive Neuroscience, 5* (2), 122–124.

———(2017). The problem of mental action. Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Norman, D. A. & Shallice, T. (1986). Consciousness and self-regulation. In R. J. Davidson, G. E. Schwartz & D. Shapiro (Eds.) *Consciousness and self-regulation* (pp. 1-18). New York: Springer.

Ondobaka, S. & Bekkering, H. (2012). Hierarchy of idea-guided action and perception-guided movement. *Frontiers in Psychology, 3*, 579.

Pezzulo, G. & Cisek, P. (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in Cognitive Sciences, 20* (6), 414–424.

Pezzulo, G., Rigoli, F. & Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology, 134*, 17–35.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology,* 9 (2), 129–154.

Sass, L. A. & Parnas, J. (2003). Schizophrenia, consciousness, and the self. *Schizophrenia Bulletin, 29* (3), 427–444.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17* (11), 565–573.

——— (2015). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570108.

Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology, 2*.

Stock, A. & Stock, C. (2004). A short history of ideo-motor action. *Psychological Research, 68* (2-3), 176–188.

Taylor, J. E. T. & Witt, J. K. (2014). Altered attention for stimuli on the hands. *Cognition, 133* (1), 211–225.

Voss, M., Ingram, J. N., Haggard, P. & Wolpert, D. M. (2006). Sensorimotor attenuation by central motor command signals in the absence of movement. *Nature Neuroscience, 9* (1), 26–27.

Wulf, G., McNevin, N. & Shea, C. H. (2001). The automaticity of complex motor skill learning as a function of attentional focus. *The Quarterly Journal of Experimental Psychology: Section A, 54* (4), 1143–1154.

Zahavi, D. (1994). Husserl's phenomenology of the body. *Etudes Phénoménologiques, 10* (19), 63–84.

# The Problem of Mental Action

## Predictive Control without Sensory Sheets

Thomas Metzinger

In mental action there is no motor output to be controlled and no sensory input vector that could be manipulated by bodily movement. It is therefore unclear whether this specific target phenomenon can be accommodated under the predictive processing framework at all, or if the concept of "active inference" can be adapted to this highly relevant explanatory domain. This contribution puts the phenomenon of mental action into explicit focus by introducing a set of novel conceptual instruments and developing a first positive model, concentrating on *epistemic* mental actions and epistemic self-control. Action initiation is a functionally adequate form of self-deception; mental actions are a specific form of predictive control of effective connectivity, accompanied and possibly even functionally mediated by a conscious "epistemic agent model". The overall process is aimed at increasing the epistemic value of pre-existing states in the conscious self-model, without causally looping through sensory sheets or using the non-neural body as an instrument for active inference.

**Keywords**

Attentional agency | Cognitive affordance hypothesis | Cognitive agency | Epistemic agency | Epistemic agent model | Epistemic goal states | Epistemic self-control | Epistemic value | Interactive inference | Interoceptive inference | M-autonomous | M-autonomy | Mind wandering | Phenomenal self-model | Predictive control | Veto control

### 1 Introduction: The Problem of Mental Action

There is no obvious way to accommodate mental action within the framework of predictive processing (PP). Examples of mental action are the volitional control of endogenous attention (as, for example, in deliberately focusing one's attention on a perceptual object or attaching it to an abstract goal-representation), trying to retrieve a series of images from episodic memory, using semantic memory to bind an object as a token to its type, active categorization or the construction of part-whole relationships, as well as engaging in mental calculation, the "building" of an argument from premises or high-level reasoning. Mental actions are a large and relevant subset of the domain of mental events, but it is unclear if they can be made amenable to scientific explanation using the conceptual instruments offered by PP. The latter approach holds out the promise of uniting perception, attention, and bodily action in a single formal framework. Yet if we take as our starting point a reasonably well-established notion such as "embodied active inference" (Friston et al. 2014; Fabry 2015; Fabry 2017), then while perhaps *bodily* actions can be explained in terms of "self-fulfilling motor fantasies" cancelling out proprioceptive prediction errors, *mental* actions cannot thus be appropriated.

In mental actions there is no motor plant to be controlled, no sensory manifold that could be manipulated by bodily movement. It is not easy to assimilate mental actions to the idea of active inference (Pezzulo 2012), simply because they do not necessarily involve any relevant changes in the non-neural body or the prediction of sensory events. From a metatheoretical perspective, mental action poses the interesting challenge of describing the deeper principles of goal-state selection and action initiation while subtracting the non-neural body and abstracting from issues of motor implementation.

My main claim in this chapter is that mental action is the predictive control of effective connectivity, where what is predicted is the epistemic value of states integrated into the phenomenal self-model under counterfactual outcomes. I will also claim that the circular causality constituting genuine men-

Metzinger, T. (2017). The Problem of Mental Action - Predictive Control without Sensory Sheets.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 19. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573208

tal action does not embrace events on any sensory sheet,[1] that mental action is a rare event, and that the specific phenomenal signature of mental action can be explained by a new content layer in the conscious self-model, the "epistemic agent model" (EAM), which may sometimes transiently emerge in the brains of human beings. In addition, I present the cognitive affordance hypothesis, which proposes that a central function of autonomous activity in the mind wandering network is to create a constant stream of *affordances for cognitive agency*, a continuing internal competition among possible cognitive actions.

Section 2 lays some conceptual foundations, Section 3 presents four building blocks of a future theory, and Section 4 draws some interim conclusions towards a first positive model for the new target phenomenon of mental action. I end by sketching a model of action initiation as a functionally adequate form of self-deception and draw attention to a metaphysical dilemma constituted by what I see as the three most important issues for future research.

## 2 What *Is* Mental Action?

I will briefly introduce some conceptual distinctions and tools in this section. Later I will situate these tools in the logical context of current theorizing on predictive processing, to prepare for a brief application at the end.

### 2.1 Mental Action versus Mental Behavior

Philosophers have thought long and hard about what distinguishes "action" from other kinds of event in the physical world (Davidson 2001; Dretske 1988; Wilson and Shpall 2016). As a matter of fact, "action theory" can be considered a small subfield within the discipline of academic philosophy. However, one of the major deficits of "action-oriented" views in cognitive science (e.g. Engel et al. 2013; Engel et al. 2016) is that there is no shared or even clearly defined concept of "action" in the background, no framework which could unite the new field from a metatheoretical perspective.

For the purposes of this paper, let us distinguish between "actions" and "behaviors" as follows. Actions and behaviors are a subset of the overt output of information-processing systems which are conceptually distinguished from other outputs by their conditions of satisfaction, being directed at goal states. Actions and behaviors can be successful, or they can fail. For actions, however, *conscious* goal-representation plays a central causal role. Actions can be terminated, suspended, or intentionally inhibited, and they exhibit a distinct phenomenological profile involving subjective qualities such as agency, a sense of effort, goal-directedness, global self-control, and ownership. Arguably, there is also a phenomenal quality of "ultimate origination", the more-or-less implicit appearance of a robust ability to do otherwise (see Section 2.3). Behaviors, on the other hand, while purposeful, do not entail explicit, conscious goal-representation. They are functionally characterized by automaticity, decreased context-sensitivity, and low self-control. Although they can be more rapid than actions, we may not even notice their initiation. While their phenomenological profile can at times be completely absent, behaviors typically involve the subjective experience of ownership without agency, where the introspective availability of goal-directedness varies and meta-awareness is frequently absent.

We can add a second conceptual distinction: there are not only bodily actions but also *mental* actions. Mental actions belong to the internal, *covert* output of some information-processing systems. Deliberately focusing one's attention on a perceptual object and consciously drawing a logical conclusion are examples. As with physical actions, mental actions possess satisfaction conditions (i.e. they

1 A sensory sheet is a collection or population of receptors. Sensory receptors absorb physical energy from a stimulus; in this way they can also function as *transductors* by transforming physical energy into electrical energy in the form of neural firing. Examples are the photoreceptors on the retina, which hyperpolarize in response to electromagnetic energy, or the olfactory receptor cells enabling odor perception by forming a spatially discontinuous olfactory sheet in different nasal cavities and airflows. The example of mechanoreceptors across the body surface shows that there can be great variations in the ability for tactile discrimination on the skin, because the minimal interstimulus distance required to perceive two simultaneously applied stimuli as distinct can vary between 1 and 45 mm.

are directed at a goal state) and, although they mostly lack overt behavioral correlates, they can also be intentionally inhibited, suspended, or terminated. In addition, however, they are characterized by their temporally extended phenomenology of ownership, goal-directedness, a subjective sense of effort, and the concomitant conscious experience of agency and *mental* self-control. For the purposes of this paper I will also assume that mental actions are typically directed at epistemic goal-states. Examples of such states are "seeing this visual object more clearly and in greater detail", "knowing the sum of 2 + 3", and "having arrived at a valid conclusion".

Mental action is a specific form of flexible, adaptive task control with proximate goals of an *epistemic* kind: in consciously drawing conclusions or in guiding attention there is always something the system wants to *know*, for example the possibility of a consistent propositional representation of some fact, or the optimal level of perceptual precision. There may also be relevant classes of non-epistemic or purely "conative" mental actions, for example those that are directed towards reward events. Here, I would propose that for the very large majority of mental actions "reward expectation" can be conceptually reduced to "epistemic value", for example as a relevant fitness-enhancing information gain under counterfactual outcomes. But let me keep things simple by limiting the investigation to epistemic mental actions. If irreducibly non-epistemic mental actions exist, they are not covered by the main arguments in this chapter.

Not only are there mental actions, however; there are mental *behaviors* too. These also belong to the *covert* output of some information-processing systems. "Mind wandering", or spontaneous, task-unrelated thought, is a paradigm case of unintentional mental behavior (Metzinger 2013a; Metzinger 2015; Metzinger 2017; Smallwood and Schooler 2015). Some mental activities are not autonomously controllable, because one centrally important defining characteristic does not hold: they cannot be inhibited, suspended, or terminated. Mental behavior may often be purposeful, but it exhibits no conscious goal-representation nor overt behavioral correlate. It is characterized by an unnoticed loss of mental self-control and a high degree of automaticity, plus a lack of sensitivity to the external situational context. The phenomenological profile is marked by ownership without agency, a variable or null capacity for introspective availability of goal-directedness, and — frequently — lack of any meta-awareness (Schooler et al. 2011). To be sure, involuntary mental behavior may serve many important epistemic functions, for example creative incubation, the consolidation of long-term memory (Mooneyham and Schooler 2013), or the continuous updating and maintenance of an autobiographical self-model (Metzinger 2013a). It may also assist the refinement and consolidation of goal-representations, by gradually descending from the abstract level of their general satisfaction conditions towards concrete, embodied motor intentions (Medea et al. 2016), and in this way it may enable further epistemic benefits in the future (Bortolotti 2015). In addition, low levels of cognitive control can boost epistemic processes in open-ended tasks relying on the use of diverse sources of information and involving temporal delays (Amer et al. 2016, p. 911), and a considerable number of mind wandering episodes are even intentionally controlled in their onset (Seli et al. 2016). There is a difference between "zoning out" and deliberately "tuning out", and there are interesting and fine-grained phenomenological nuances connecting both phenomena: some episodes are intentionally initiated, and sometimes even sustained by wilful "rebooting", but as they unfold over time they typically become unintentional mental events — inner behaviors, not inner actions. As I will propose, unintentional mental behavior is interesting, because it helps to *constitute* mental action.

## 2.2 Mental Autonomy: Variable Degrees of Epistemic Self-Control

We have just seen that mental actions are typically directed at epistemic goal states. Interestingly, these states, if successfully brought into existence, are often "self states", because it is the organism itself which has acquired new (epistemic) properties. It is the system *itself* (and not the world) which now knows something new, has optimized the depth of perceptual object representation, has arrived at a

novel conclusion, and so on. There are obvious exceptions. For example, whenever human beings actively contribute, via more complex forms of social interaction, to the knowledge their *group* possesses, they change epistemic group properties as well. In the very large majority of cases, however, mental action is a process by which an individual changes their own epistemic properties. If we assume the conceptual distinctions above, general intelligence can be seen as the capacity for adaptive epistemic self-control and *mental* self-control plays a central role. Self-control comes with different and variable degrees of autonomy. This raises the question as to the minimal degree of functional autonomy which enables an information-processing system to become an agent, in the sense that it crosses the threshold from mental behavior to mental action.

A simple and empirically tractable concept is "M-autonomy" (Metzinger 2015). In general, autonomy is often framed as the capacity for rational self-control of overt behavior, whereas the term M-autonomy refers to the specific ability to control one's own mental functions, like attention, episodic memory, planning, concept formation, rational deliberation, decision-making, and so on. One route to a richer conceptual analysis is to describe it as the capacity for *second-order mental action*, i.e., vertical, intra-mental, top-down control. This can be decomposed into the following attributes:

- the imposing of rules on one's own mental behavior;
- explicit goal-selection and commitment, establishing goal-permanence;
- satisfying the constraints of rationality or rational guidance; and
- the ability intentionally to inhibit, suspend, or terminate a process.

This last condition, "veto control", is the central semantic element in defining M-autonomy: if one cannot terminate one's own activity, one cannot be said to be autonomous in any interesting sense. This element can be empirically grounded and gradually refined, and it may prove heuristically fruitful in guiding research. Veto control is a manifestation of the ability to suspend or inhibit an action voluntarily, and from a logical point of view it is a functional property which we do not ascribe to the brain but to the person as a whole. Let us call the capacity "intentional inhibition".[2] During a mind wandering episode, we do not have this capacity because we cannot actively suspend or inhibit our own mental activity (Metzinger 2013a; Metzinger 2017). Therefore, our degree of epistemic self-control is low.

How would one go about isolating the neural basis of autonomous epistemic self-control? From a conceptual point of view, every representation of agency must have three logical components: a model of an entity exerting control (the "self"), a model of the satisfaction conditions of the specific action (the "goal state"), and an asymmetric relationship dynamically connecting and transiently *integrating* the first two components (the "arrow of intentionality"). For the special case of mental action this has three implications. First, we must avoid any homunculus fallacy with regard to the first component. Second, we must do justice to the fact that the number of possible goal states is extremely large, because at any given point the number of possible targets for introspective attention — as well as the number of potential contents for the control of abstract, symbolic thought — is much larger than that for bodily action. For human beings, the set of *cognitive* target objects and states is much larger than what they could reach for, grasp, or run to — simply because our inner environment has become much richer and more complex than the concrete space of causal interaction in which our physical bodies are situated. Third, the dynamic, relational component connecting the first two elements has to be extremely fast and flexible, and it must be able to adapt to a complex task domain in a fluid and highly context-sensitive fashion. This already puts valuable constraints on possible architectures and means of realization.

For example, given the first conceptual constraint above, for every individual mental action, one would expect task-independent and task-dependent components to become integrated. Functional

2 In adopting this terminological convention, I follow Marcel Brass (Brass and Haggard 2007); an excellent and helpful recent review is (Filevich et al. 2012).

connectivity analyses point to a combination of intrinsic and task-evoked connectivity patterns, reflecting a global network architecture composed of an intrinsic part which is also present during rest, and task-general as well as task-specific patterns of connectivity evoked by specific demands (Cole et al. 2014). It is important to avoid implicit homunculus fallacies and to dissolve the "cognitive agent" component into a statistical analysis which assigns flexible sets of interacting cognitive subsets to every specific instance of task execution, but in a way that still reveals functional clusters or "network roles" (Mattar et al. 2015, p. 3). The problem is not to find some mysterious little "man in the head", but to develop a formal understanding of how the brain manages to continuously approximate an optimal balance between global integration and local flexibility, to develop a framework explaining how a stable state can be maintained while transient coalitions of network units generate specific cognitive behaviors. Such an understanding is gradually emerging. "Network roles" will be very complex functional states — context-sensitive, non-encapsulated, determinate — which can conceptually be defined as very large clusters of causal relations. If human beings belong to the class of probabilistic automata (Putnam 1967; Putnam 1975), however, then this engenders the challenge of how even to begin reliably mapping such complex probabilistic roles onto brain regions.

Recent empirical work reveals the dorsal fronto-median cortex (dFMC) as a candidate region for the physical realization of this very special form of purely mental second-order action.[3] It does not overlap with known networks for external inhibition, and its computational function may lie in predicting the social and more long-term individual consequences of an unfolding action, that is, in representing the action's socially and temporally more distant implications for the organism.[4] There is a considerable amount of valuable neurobiological data on the physical substrates of intentional inhibition in human beings, and a number of studies have already led to more abstract computational models of volitional control, action selection, and intentional inhibition itself (Filevich et al. 2012; Filevich et al. 2013; Campbell-Meiklejohn et al. 2008; Kühn et al. 2009; Brass and Haggard 2007). These data are valuable not only for understanding the "back end" of many mind wandering episodes — the transition from mental behavior to mental action — but also for a more comprehensive theory of mental autonomy (for more, see Metzinger 2013a, Section 3.3). From a philosophical perspective, the functional property of M-autonomy is interesting for a wide range of reasons, including its relevance to our traditional notions of a "first-person perspective" (1PP) and "personhood" (Metzinger 2015). If one cannot control the focus of one's attention, then one cannot sustain a stable perceptual first-person perspective, and for as long as one cannot control one's own thoughts, one cannot count as a rational individual.

What could be a first, empirically plausible candidate for a neural realization of the complex functional demands posed by selective cognitive self-control? The fronto-parietal network (FPN) may be a good candidate for this cluster of functional properties (Cole and Schneider 2007; Niendam et al. 2012; Cole et al. 2013). Kalina Christoff and colleagues (Christoff et al. 2016, p. 721) hypothesize that it supports the deliberate constraining of the contents of thought. The FPN plausibly plays a central role in mental health: impaired cognitive self-control and disrupted processes of goal-representation are markers of disease across a large spectrum of neuropsychiatric conditions (Cole et al. 2014), and domain-general measures of fluid, culturally invariant and knowledge-independent intelligence have been found to be specifically correlated with the lateral prefrontal cortex, a circumscribed region within the FPN (Cole et al. 2012). Recent models of cognitive control support the idea of a "flexible hub" which can at the same time monitor and causally influence a large variety of task-relevant information sources (Cole et al. 2012, p. 8997). Relative to other known networks, the FPN is especially active during phases of highly adaptive task control, exploiting global variable connectivity by flexibly

3 See (Kühn et al. 2009), (Brass and Haggard 2007), and (Campbell-Meiklejohn et al. 2008). A helpful recent review of negative motor effects following direct cortical stimulation, listing the main sites of arrest responses and offering an interesting discussion, is (Filevich et al. 2012).

4 This passage draws on (Metzinger 2013a). See also (Filevich et al. 2012, Filevich et al. 2013).

shifting the connectivity pattern across many different brain regions and a wide variety of tasks. It has also been proposed that the FPN employs principles of "compositional coding", which would allow for certain connectivity patterns and representational contents to be reused and recombined (Anderson 2015) in order to transfer existing knowledge across tasks, thereby enabling the rapid learning of novel tasks in new functional contexts (see Cole et al. 2013, fig. 1). While the FPN's variable connectivity is truly global, it has also been identified as one of the ten major functional networks which partition the brain into intrinsic functional clusters, independent of particular task- or goal states. In the current context, it may be relevant to investigate the FPN's causal interaction with the default-mode network (DMN). This is a distinct network of interacting brain regions, whose activity is highly correlated, which automatically activates when a person is not involved in any task — such as at wakeful rest, daydreaming, when the individual is simulating social situations, or during autobiographical rumination. Interestingly, the variable connectivity of the FPN has been found to be significantly greater than that of the DMN with the entire brain (cf. Cole et al. 2013, fig. 5) — a point to which I shall return in Section 3.2.

## 2.3 Mental Action Type 1: Volitional Attention

Many authors have recently begun to ask the question of whether predictive processing (PP) can be extended to a genuine model of high-level cognition (Barsalou 2016; Butz 2016; Pezzulo et al. 2016; Spratling 2016). Let us briefly distinguish the two main types of mental action, which have to be accommodated under the PP approach. I will return to them in Sections 3.1 and 3.2.

Type 1, volitionally controlled attention, brings about a specific set of phenomenal properties, as is the case for pain or the subjective quality of "blueness" in a visual color experience (Metzinger 1995). Attentional agency (AA) is the conscious experience of actually initiating a shift of attention, of controlling and fixing attentional focus on a certain aspect of reality. AA involves a sense of effort, and it is the phenomenal signature of our functional ability actively to influence what we will come to know, and what, for now, we will ignore. As with all other forms of agency, it also involves the subjective quality of "ultimate origination" mentioned above — from the first person perspective, it seems that one could have done otherwise; in the way one experiences the overall process from the first-person perspective, any unconscious causal precursors are necessarily unknown such that the first subjective event carrying the phenomenal character of control (say, determining the focus of attention) necessarily appears as spontaneous and uncaused, emerging "out of the blue" as it were. On this level of the hierarchy, it is an unpredicted internal event. What we call "agency" refers to an interpretation of this fact as "initiation" or "origination": it is the activation of an internal self-model trying to explain away, by creating an explicit, supra-modal representation of an entity capable of ultimate origination and spontaneous self-causation, the surprise involved in suddenly achieving autonomous self-control. I return to this point in Section 4.3.

Consciously experienced AA is theoretically important because it is probably the earliest and simplest form of experiencing oneself as a *knowing* self, as an epistemic agent. To consciously enjoy AA means that one (the cognitive system as a whole) currently identifies with the content of a particular self-representation, that one operates under an "epistemic agent model" (EAM; see Section 2.5 below and Metzinger 2013a; Metzinger 2013b) active in one's brain. Being a phenomenological entity, an EAM can always be a hallucination, but typically it will be a window of self-knowledge, telling the system that M-autonomy has been achieved — we must always be careful to keep functionalist, mech-

anistic, and epistemological readings of this new term apart (see Section 3.1). AA is fully transparent:[5] the content of one's conscious experience is not a self-representation or a process of self-modeling, of depicting oneself as a causal agent in certain shifts of "zoom factor", "resolving power", "resource allocation", and so on. Rather, one directly experiences *oneself* as, for example, actively selecting a new object for attention. During nocturnal dreams and mind wandering episodes we do not have AA, although these episodes can of course be *about* having been an attentional agent in the past, or *about* planning to control one's attention in the future. Other examples of situations in which this property is selectively missing are non-lucid dreams and non-REM-sleep (rapid eye movement) mentation (Metzinger 2013b; Windt 2015), and also infancy, dementia, and severe intoxication.

## 2.4 Mental Action Type 2: High-Level, Symbolic Cognition

What is reasoning, logical thinking, or mathematical cognition from a PP perspective? Can prediction-based mechanisms be fully detached from overt sensorimotor loops (Pezzulo 2016, p. 33)? We can conceptualize cognitive agency (CA) as an abstract mental simulation of embodied actions, first executed using the physical, non-neural body. Such actions could have been the manipulation of discrete symbolic tokens in the external world, the use of gestures or bodily sign language, or even full-blown speech acts. But we can also frame them as abstract, inclusively internal versions of adaptive action control, involving predictive loops (see Section 3.2).

Again, there is a distinct phenomenology of currently being a cognitive *agent*, which can lead to experiential self-reports such as "I am a thinking self in the act of grasping a concept", "I have just actively arrived at a specific conclusion", "I am attempting to build an argument", and so on. We should be careful not to confuse the level of functional analysis ("autonomous cognitive self-control") with phenomenological readings based on verbal self-reports. What AA and CA have in common is that, in both cases, we consciously represent ourselves as epistemic agents: according to subjective experience, we are entities that actively construct and search for new epistemic relationships to the world and ourselves. We are information-hungry; there is something we want to *know*.

## 2.5 The Phenomenology of Epistemic Self-Control: The EAM

If the EAM is what explains the unifying phenomenal signature of mental action, then it is the common denominator on the phenomenological level of analysis. Phenomenologically, for a conscious cognitive system to operate under an epistemic agent model means for it to *know that it knows*. According to subjective experience, some of its own states have an epistemic value. They seem to present information about the world, and they often do so in a stable and counterfactually rich way — we can imagine many situations or possible worlds in which they would stay the same. As yet, we are only operating on a phenomenological level of description, but from a computational perspective this additional feature could serve as an introspective indicator of the fact that there exist many counterfactual manipulations under which the information presented by these states would remain largely invariant because there is a specific, abstract form of robustness or stability characterizing their informational content. Knowledge in this sense is the possession of *reliable* information.[6]

5 "Transparency" is a property of conscious representations, namely that they are *not experienced* as representations. Therefore, the subject of experience has the feeling of being in direct and immediate contact with their content. Transparent conscious representations create the phenomenology of naïve realism. An opaque phenomenal representation is one that is experienced *as* a representation, for example in pseudo-hallucinations or lucid dreams. Importantly, a transparent self-model creates the phenomenology of identification (Section 3; Metzinger 2003; Metzinger 2008). There exists a graded spectrum between transparency and opacity, determining the variable phenomenology of "mind-independence" or "realness". Unconscious representations are neither transparent nor opaque. See (Metzinger 2003) for a concise introduction.

6 The classical philosophical issue here is at what point rich and reliable information turns into semantic information, something that could be true or false (Dretske 1988). It is important to distinguish this metatheoretical question from the empirical question of how and at what level of reliability and counterfactual invariance a given type of cognitive system *internally models* information it possesses as semantic information.

Having an EAM is an instance of self-consciousness. For a human being it means that, subjectively, it now possesses a specific kind of self-knowledge — *knowing* that one knows and that one is able actively to control certain epistemic states. An EAM is a model of a single entity capable of autonomous epistemic self-control and ultimate origination. As this happens on the level of conscious processing, it also creates the phenomenology of ownership for certain states of perceptual or cognitive knowledge. Whatever can be autonomously controlled generates the subjective experience of ownership — a general principle which holds not only for the sensorimotor control of one's own body but also for a wide range of phenomena, ranging from simple tool-use to virtual re-embodiment in avatars and robots (Cohen et al. 2014; Cohen et al. 2014; Lenggenhager et al. 2007; Blanke and Metzinger 2009; Metzinger 2008). The conscious processing does not only enhance adaptivity, flexibility and context-sensitivity in active acquisition of knowledge. It also adds the property of *globality*: because states with high epistemic value are subjectively "owned" and integrated into a phenomenal self-model, possessing them becomes a property of the system as a whole — something it ascribes to itself, a unified space of hypothesis generation.

An EAM is also a new unit of identification (Metzinger 2017; Metzinger 2013a, p. 10). It enables self-reports of the type "I *am* this knowing self!" Phenomenologically, identifying with the content of an EAM typically also refers to a high probability that this system will come to know further novel aspects of the world (bringing about "epistemic optimism", i.e. a high expected rate of error reduction). This also means that the system will be *motivated* to bring about a lot of those changes in its own global state of knowledge (it has a "capacity for epistemic self-control", and the related phenomenal sense of control is expressed as positive affect). Perhaps it also has an "autobiographical" model of past fluctuations in prediction error minimization and therefore expects to repeatedly initiate new epistemic actions after a certain time has passed (see Van de Cruys 2017, for an interesting discussion with regard to curiosity and affective value). This is what having an internal model — not only of some passively knowing self but of an epistemic *agent* — means. There are goal states plus a possibility of failure, there is a corresponding high-level capacity, and often this capacity is not just an abstract feature but something that is exerted — a concrete process consciously experienced.

In this way a running EAM is a phenomenal self-model containing an ongoing prediction of future epistemic states, plus an explicit representation of the capacity to bring these states about. That is, there is a specific, high-level capacity for self-control and there are variable rates of prediction error minimization and degrees of autonomy related to this capacity. These degrees of autonomy are always linked to a certain probability which can itself become the target of a predictive model. For example, one could in principle measure autonomy from the outside, by counting, relative to a specific situation, the number of times an agent is able to suppress a spontaneously occurring impulse for motor action. This relative frequency would yield a probability, which we could use to express the degree of autonomy. The same also could be done from the inside, for epistemic actions, with the help of a specific layer in the phenomenal self-model.

A unified self-model does not necessarily entail that a distinct entity such as "a" self exists as well: the phenomenology of "knowing that one knows" could be constituted by something other than the possession of reliable information, as in self-deception. That second-order state ("knowing that one knows") might not be epistemic at all, or, as for example in certain states of belief and subjective certainty caused by epileptic seizures or direct electrical stimulation of the insula (Picard 2013; Picard et al. 2013), the first-order state might be a physical artefact or based on highly unreliable information. The same goes for the phenomenal qualities of "autonomy" and "ultimate origination": the agency component characterizing the subjective experience of being an active, knowledge-seeking self could be a first-person phenomenon only. I will therefore return to non-phenomenological levels of analysis in Section 3.3, to discuss potential constitutive, causal, or explanatory relationships between predictive control and the conscious EAM.

## 3 Building Blocks for a Positive Model

We now have a set of conceptual distinctions and tools that permit us to develop a positive model of mental action. My aim, however, is more modest than this may sound: I merely want to arrive at a conceptual nucleus for the notion of "mental action", a first — and hopefully heuristically fecund — working concept which can be refined and empirically enriched as we go along. The claim in this section will be that mental actions are a specific form of EAM-mediated predictive control of effective connectivity, aimed at increasing the epistemic value of pre-existing states in the self-model without using the non-neural body as an instrument for active inference.

### 3.1 Building Block 1: Attentional Agency under PP

In his seminal and ground-breaking book, *The Predictive Mind*, Jakob Hohwy draws our attention to the central issue:

> Endogenous attention often seems to come with a volitional element that is not quite captured by […] relatively mindless cue directing. We decide to attend to some feature, and then act on that decision. In prediction error parlance, this aligns volitional attention with active inference. The question is, can we conceive of active inference such that it accommodates expectations for precisions? (Hohwy 2013, p. 197)

This is the right question. To find an answer, we need a better understanding of the epistemology of attention, a plausible computational story, and a representationalist analysis of its specific phenomenal character, i.e. the introspectively available conscious content activated along with attentional agency.

Epistemologically, every form of volitional attention is a form of introspection. External scaffolding by the non-neural body is not a necessary condition. We can visually attend to hallucinations with closed eyes or to the hypnagogic imagery preceding sleep onset, even after the body has entered sleep paralysis. During a lucid dream, we can wilfully attend to the dream environment as well as to sensations in the dream body (Metzinger 2008; Windt 2015). A clearer, and perhaps even more provocative, way to bring out this distinction is by saying that, functionally, all high-level, volitional attention is active introspection: it always operates on aspects of an internal world model. If I attend to a visual object in front of me, then I am actively optimizing the second-order statistics of this (inclusively internal) object model. I claim that this is not only true during a lucid dream, but also during waking life. If I attend closer to some visceral sensation in my stomach or to the sensations going along with my breath, then I actively optimize the second-order statistics of my interoceptive self-model. Functionally, both forms of mental action are exclusively internal, because their satisfaction conditions describe a goal state which is a state of an internal model. Phenomenologically, however, these situations are *very* different: in the first case I experience myself as attending to the environment; in the second case I experience myself as actively introspecting certain aspects of bodily self-consciousness (see below).

Jakob Hohwy writes:

> All this seems no different in principle from the active inference we have come across before: a counterfactual hypothesis induces a prediction error causing us to change our relation to the world. It is a slightly unusual instance of action because the way we change our relation to the world is to increase our sensory gain in one region of space. The difference is then that the active inference is driven not by which prediction error we expect, but by the precision of the expected precision error — just as we can sample prediction error selectively, we can sample their precisions selectively. (Hohwy 2013, p. 198)

I think the general idea is correct, but two details are not quite right. Firstly, in being a volitional attentional agent, the system changes its relationship not to "the" world, but to an internal world *model*. Of course, this neurally realized model is a part of the physical world too, but for the system this means that it actively changes its relationship not to the environment, but to a part of itself. Its epistemic perspective is one of active introspection. Secondly, the gain that is increased is not a gain relative to an extra-organismic "region of space", because gain is really only increased for one region of *internal* computational space.[7] The "signal" already is an exclusively internal event, because it is a non-sensory "signal".

Computationally, attentional agency (AA; Metzinger 2003, 6.4.3; Metzinger 2006, Section 4; Metzinger 2013a; Metzinger 2015; Metzinger 2017) is the ability to control precision expectations. However, it is important to make the following distinction: note that precision is a property of the *signal*. If attentional agency is conceptualized as a form of prediction only, then what is predicted is a property of the raw sensory manifold, which is not a representational entity in itself. Its statistics just reflect the causal structure of the world. If AA is conceptualized as a form of *control*, however, then what is controlled is a property of a representational entity, namely the precision expectation embodied by an internal model. This is what makes attentional agency in this richer sense an inherently *mental* kind of action: it optimizes second-order statistics by changing properties of an internal model. The standard version of active inference at most optimizes the precision of a non-representational entity; AA is mental because it optimizes a representational entity.

What about the phenomenology? Clearly, attentional agency is not just any form of controlling precision expectations: it is the subset accompanied by (and possibly functionally mediated through) a conscious model of the self as an epistemic agent — what I have called an EAM in Section 2.5 above. It is the best currently available global hypothesis about certain future epistemic states, in this case states of *perceptual knowledge*. Because this hypothesis originates in what I will call "EAM-space", it automatically predicts a specific form of *self-knowledge* as well: probably, there will not only be a richer and more detailed experience of some specific perceptual object in the near future, but this will also be modelled as the perception of a "knowing self", a self which will be phenomenally aware of this very fact. Recall how an EAM is a model of a single entity, capable of autonomous epistemic self-control, and how, on the level of conscious processing, this creates the phenomenology of ownership for certain states of perceptual or cognitive knowledge.

Being an attentional agent therefore is always a special form of *self*-consciousness: one becomes aware of one's *own* capability for — and the actual process of — controlling the quality of one's own perception, its depth and resolution. At the same time, while quashing prediction error about one's expected precision, one experiences a specific sense of effort, a sort of resistance — and, in combination with the subjective quality of ultimate origination described above, it is this which engenders the relevant, exclusively mental sense of agency. Perhaps we can describe this experiential tension[8] as the way in which active, top-down precision control "collides" with an automatically arising precision prediction error. If so, one might speculate that the temporal curve in which the phenomenal sense of agency unfolds over time exactly reflects this "collision" and its eventual resolution.

The possibility of relating the phenomenology to a fully internalist computational analysis of volitional attention, as the EAM-mediated control of second-order statistics for perceptual states, reveals what is perhaps the philosophically most important point: the phenomenology and epistemology of environment-directed versus introspective attentional agency can diverge dramatically. Functionally, both forms of mental action are exclusively internal. Phenomenologically, however, in the former I

7 Jakob Hohwy discusses volitional spatial attention here, but the point would also hold for volitional feature attention. The relevant processing stream is exclusively internal.

8 A much more detailed analysis of subjectively experienced tension and its relationship to self-deception can be found in (Pliushch in press). See also (Pliushch 2017).

will experience myself as attending to the world out there, while in the latter I will experience myself as introspecting aspects of my own mind, my emotional state, or my bodily self.

### 3.2 Building Block 2: Cognitive Agency under PP

In *Surfing Uncertainty*, Andy Clark reintroduces and develops the notion of an "affordance competition" (Cisek 2007; Cisek and Kalaska 2005; Cisek and Kalaska 2010; Clark 2016, p. 179). This approach could be extremely helpful in understanding what high-level symbolic cognition really is, and in situating it in a PP-framework. I will extend this idea further, into the domain of unintentional mental behavior and mental action. The interim goal will be to create a second building block to arrive at a positive model for the second main category of mental actions. Clark writes:

> One powerful strategy […] involves rethinking the classical sense-think-act-cycle as a kind of mosaic: a mosaic in each which shard combines elements of (what might classically be thought of as) sensing and thinking with associated prescriptions for action. At the heart of this mosaic vision […] lies the simultaneous computation of multiple probabilistically infected 'affordances': multiple possibilities for organism-salient action and intervention.
> The idea here is that the brain is constantly computing — partially and in parallel — a large set of possible actions and that such partial, parallel, ongoing computations involve neural encodings that fail to respect familiar distinctions between perceiving, cognizing, and acting. (Clark 2016, p. 177; 180)

Could there be something like an exclusively *internal* affordance competition? Let me propose that mind wandering, the almost continuous appearance of task-unrelated thoughts, may be exactly this — the creation of a constant flow of possible *mental* actions, a dynamic *inner* environment constituted by non-sensory events which need to be predicted and controlled. Call this the "cognitive affordance hypothesis".

We have already seen that apparently spontaneous, apparently task-unrelated thought can conceptually be described as a form of unintentional behavior. The cognitive affordance hypothesis states that a central function of autonomous activity in the mind wandering network is to create a constant stream of *affordances for cognitive agency*, a competition among possible cognitive actions. It is empirically plausible to assume that large parts of this pattern overlap with activity in the default mode network (DMN; Buckner et al. 2008; Christoff 2012; Christoff et al. 2009; Weissman et al. 2006; Stawarczyk et al. 2013; Andrews-Hanna et al. 2010; Mantini and Vanduffel 2012; Buckner and Carroll 2007; Mason et al. 2007; Spreng et al. 2009), but that it also extends to other functional structures such as the rostrolateral prefrontal cortex, dorsal anterior cingulate cortex, insula, temporopolar cortex, secondary somatosensory cortex, and lingual gyrus (for a recent meta-analysis, see Fox et al. 2015).

For an intuitive grasp, let us imagine a situation in which we had as little control over our bodily behavior as we have over our own minds. Imagine we were beings which in the awake state were plagued by recurrent involuntary motor twitches and behaviors for roughly 30-50 per cent of the time, but that we could sometimes "seize" one of these bodily behaviors by bringing them under conscious control, thereby turning them into proper, goal-directed actions. Or imagine a sleep-walker, unconsciously navigating the world on autopilot, reactively driven by low-level affordances. Sometimes, however, the sleep-walker briefly "comes to", wakes up and becomes capable of autonomous bodily self-control. After each such episode he loses conscious sensorimotor control again, automatically returning to autopilot mode. I propose that what we call "thinking", "reasoning", or even volitionally attending to some object is exactly like this: first we are mental sleep-walkers; then we episodically turn into genuine mental agents by "seizing control", perhaps with the help of a conscious EAM, namely, by re-instantiating the functional property of M-autonomy. We navigate a self-created, internal affordance

landscape, but only rarely do we achieve autonomous self-control (roughly for only one third of our conscious lifetime, cf. Metzinger 2015, fig. 2).

Mind wandering is mental sleep-walking. But unlike embodied somnambulism it actually *sets up* the inner world which later allows us to autonomously perform mental actions during a state of "full consciousness". One central function of the mind wandering network is to provide us with an internal environment populated by competing possible mental actions — unintentionally occurring mental events which have the potential to become part of a stable control loop, thereby turning into genuine inner actions. A rich inner action repertoire enables an equally rich landscape of possible interactions with our inner world; it continuously opens a whole range of new functional windows. There is no inner world containing static mental objects, seen through a single, rigid window frame. What mind wandering does, rather, is to create a fluid and highly dynamic task domain. Every spontaneously occurring "task-unrelated" mental event is a potential task in itself, a cognitive affordance, a dynamic state which has the potential to be selected and transformed from unintentional mental behavior into mental action. The mind wandering network sets up this task domain, and sometimes frontal regions of the brain latch onto the best candidate, and create a transient functional integration with the EAM, "pursuing" a possible thought further, now in a consciously controlled fashion, "trying to stay one step ahead" on the cognitive level. Perhaps we can imagine this process as the frontoparietal control network being flexibly engaged with either the default or dorsal attention network in support of goal-directed cognition (Spreng et al. 2010). Perhaps, in this specific case, we could even speak of mental "patterns of action readiness" and then describe the episodic emergence of cognitive control as the appearance of an "optimal grip", this time on an *internal* situation (Bruineberg et al. 2016; Bruineberg 2017).

I will not enter into a discussion of candidates for neural implementation in this paper, but an empirically plausible model could involve the posterior cingulate cortex (PCC) as a driver of this specific type of activity, with the medial prefrontal cortex (MPFC) modulating and directing the flow of high-level cognition, for example by acting as a gateway in selecting semantic self-representations (cf. Davey et al. 2016, p. 935). Please note that such a "semantic" self-model would already be an exact conscious representation of the system itself as a *carrier of meaning*, with not only physical but also epistemic properties. And this is what mental action is all about — the control and creation of new epistemic properties. Put differently, if conscious perception can be described as a form of controlled hallucination, conscious thinking is a process of controlled mind wandering.

This is very much in harmony with the recently developed idea of adaptive action control as the "navigation of an affordance landscape" (Pezzulo et al. 2016). What it adds are three major aspects. First is the mind wandering network as an automatic, subpersonal mechanism constantly creating a dynamic internal affordance landscape. Second is the phenomenal self-model as characterized by a hierarchy of predictive horizons (or nested timescales) generating different types of representational content. And third is the concept of an "epistemic agent model" as one specific content layer in the phenomenal self-model — the transient conscious correlate of the event of achieving autonomous epistemic self-control.

On the mental level, the mind wandering network is the system that enables the rapid switching of actions (as Pezzulo et al. 2016, p. 415 demand), by continuously processing alternative potential mental actions and creating *expected* cognitive affordances even during phases of goal-directed activity. As the human self-model integrates different predictive horizons, different timescales generate different types of nested self-representational content: the body-model predicts somatosensory events within very small time-windows of only a few hundred milliseconds, whereas the autobiographical self-model predicts events which may lie years ahead, including the organism's eventual death. The cognitive layer not only biases affordance competition at lower hierarchical levels, it crucially allows an agent to adaptively destroy and create new cognitive affordances in order to realize long-term goal-representations. M-autonomy involves veto control and refers to the functional property enabling the system not

only to "reactively pick up currently available affordances" (Pezzulo et al. 2016, p. 416), but to actively sculpt its own internal affordance landscape in accordance with the process sketched above. The EAM is its transient conscious correlate, making the current fact of successful predictive control in this task domain introspectively available.

So the core idea of the cognitive affordance hypothesis is that mental action is the control of spontaneous mental behaviour, which can be described as an internal affordance competition mediated by a large neural network overlapping with the MPFC. Recent connectivity analyses reveal the DMN as consisting of regions at the top of a representational hierarchy which describe the current representational landscape in the most abstract terms (Margulies et al. 2016). Cognitive agency, critically involving the MPFC, then becomes the active sampling of an internal, highly abstract, and non-interoceptive environment. There are no sensory signals involved; hence their causes cannot be revealed. Physical events on sensory sheets play no proximal causal role. This leads to the conclusion that what is revealed can only be non-sensory, abstract properties of ongoing neural dynamics. But what exactly is it that could be predicted by this form of mental action?

### 3.3 Building Block 3: The Convergence of Instrumental and Epistemic Action for the Special Case of Mental Action

In his reply to Wanja Wiese's commentary on his target article in the Open MIND collection, Anil Seth writes:

> 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 employed with an emphasis on predictive control (instrumental), or on revealing the causes of sensory signals (epistemic). (Seth 2015b, p. 7)

Let us extend this idea to the domain of mental action. The target paper itself (Seth 2015b, cf. Section 2.3) had drawn attention to a deep connection between PP and mid-20th century cybernetic approaches which focused on the prediction and control of behavior. Clearly, for any system that has begun to manifest unintentional *internal* behavior (such as spontaneous, task-unrelated thought) it becomes necessary to predict and control exactly this internal behavior as well, because it is a constant source of uncertainty. If this behavior is an *epistemic* activity (as mind wandering seems to be), the new task would have to be described as "predictive epistemic self-control". Seth writes:

> [R]ather than seeing PP as originating solely in the Helmholtzian notion of "perception as inference", it is fruitful to see it also as a process of model-based predictive control entailed by a fundamental imperative towards internal homeostasis. (Seth 2015b, p. 9)

What is predictively controlled in cognitive agency is the spontaneous flow of subpersonal proto-thought, and we now begin to see how the explicit model on which this control is based could possibly be what was introduced on the phenomenological level as the EAM in Section 2.5. At this point we should be extremely careful, because the questions we confront are mainly empirical, not philosophical. For example, currently available data on mind wandering and cognitive control underdetermine possible metaphysical interpretations: on one end of the spectrum we might treat the EAM as a mere epiphenomenon devoid of any functionally relevant properties; at the other end it could be exactly the missing computational element that *constitutes* epistemic, model-based predictive control on the mental level. An intermediate position could be that most mental actions begin without an EAM (for example, by subpersonal processes of mental affordance competition), but *end* by transiently culminating in a conscious model of agentive second-order knowledge. Here, the EAM would not constitute, but *represent*. It might causally enable relevant second-order functional properties like veto

control (see Section 2.2), but it would not constitute the scientific target property of predictive control. Currently available empirical data do not help to decide between these three metaphysical options.

The same point holds for potential explanatory and causal relationships between the computational notion of "predictive control" as described by Anil Seth and the new concept of an EAM. There clearly is something like an EAM; it is a theoretically relevant part of our phenomenology which has been largely ignored in the past. But when I introduced the cognitive affordance hypothesis in the preceding section by saying that we are "seizing control", perhaps with the help of the FPN and a conscious EAM, I left at least two possibilities open: the EAM could be what causally *explains* predictive control, or it could be a mere phenomenal artefact *generated* by predictive epistemic self-control (I am very grateful to Jona Vance for critical discussion here). And when in Section 3.1. I introduced attentional agency (AA) as a type of epistemic self-control not involving sensory manifolds but exclusively aimed at precision expectations embodied in an exclusively internal model, I did not provide any concrete idea of the *form* of control that was presupposed. If it can be applied to processes not involving the non-neural body, then Anil Seth's work on predictive control is relevant, because we cannot solve the problem of mental action by appealing to an ill-understood and unexplained kind of control. Currently, the explanatory priorities are unclear. If I am on the right track, then novel computational models for mental action are needed, firmly grounded in empirical data and entailing testable predictions. To make a start, here are three.

From a third-person perspective, it remains true that the very large majority of all processes and hierarchical levels having to do with constructing a viable model of reality remain completely unconscious and are not part of the organism's phenomenological life-world. But it is conceivable that, in standard situations, an EAM could causally enable a new feature — *global* epistemic self-control plus *knowing* that one possesses this feature. Epistemic states are now represented as states of a single entity that functions as the system's current unit of identification, in this case as the currently active model of a "knowing self". This model makes the content of those states available for introspective attention, and for selective action control on the whole-system level. But it also possibly predicts and controls the acquisition of epistemic properties like the formation of a concept or the having of a belief. As with any other model, we can treat the EAM as an evolving space of possible hypotheses. EAM-space is a novel space, a new subregion or partition of the organism's global self-model, constituted by all those hypotheses exclusively predicting the organism's *epistemic* properties. Integration into the phenomenal self-model in turn creates new functional properties — the content of this space is now globally available to many other processing levels at the same time: for introspective attention and the flexible control of behaviour, be it motor or mental. The level of conscious processing is the level on which a unified ontology first emerges, and epistemic properties such as the possession of knowledge become elements of this ontology, parts of the system's reality model.

These are admittedly speculative observations, but a number of testable hypotheses can already be derived. For example, if an EAM is in place, then selectively blocking its introspective availability should suffice to trigger a mind wandering episode — a bout of spontaneous, task-unrelated thought (H1; see Axelrod et al. 2015; Broadway et al. 2015 for early examples of relevant studies). Conversely, if we find an experimental procedure to reliably re-establish predictive self-control in the relevant level of the hierarchy, then we should be able to turn a non-lucid dream into a lucid dream in a sleep laboratory (H2; Voss et al. 2014; Metzinger 2008), or experimentally to end a mind wandering episode in waking subjects, letting them unexpectedly "come to" again (H3).

Let us take another example: visually searching for a perceptual object by directing and sustaining the focus of visual attention, say, by "looking for" a lemon on the table in front of me. There are three very general expectations involved: that a lemon exists, that this lemon will soon be seen more clearly, and that it will be seen more clearly by *myself*. There is a hyperprior for "perceptual objecthood" (or, perhaps, only a mutual specification of counterfactually rich and hierarchically deep predictions), we have a high-level expectation of successful precision control in the visual domain, and there is a hy-

perprior for "epistemic agency". What is predicted is the holding of a relation between a subject and an object, in the near future, with the relation transiently uniting these elements being of a specific type: the category of "perceptual knowledge" or sensory-driven object representation.

Elsewhere, I have called the conscious correlate of this specific type of transient, dynamic, subject-object-representation the "phenomenal model of the intentionality relation" (or PMIR, see Metzinger 2003; Metzinger 2004; Metzinger 2006), because I claim that it is the conscious brain's subsymbolic and naturally evolved answer to the problem that Brentano and Husserl wanted to solve, much later, on a theoretical level. This model, in a fluid and highly context-sensitive manner, describes the temporal evolution of the "arrow of intentionality", that is, the asymmetric relationship of a system being *directed* at a goal state. For epistemic action, this goal state is one of possessing knowledge — for example, of being related to the lemon through the active, embodied process of "seeing" the lemon, of first establishing and then successfully sustaining a stable causal loop passing through one specific sensory sheet, namely, the receptor system constituting the outer statistical boundary of the visual system. The PMIR is the conscious model of a currently standing loop. Here, my empirical prediction would be that selectively blocking this and only this loop, say, by transcranial magnetic stimulation, would selectively destroy the phenomenology of epistemic goal-directedness (H4). It would leave the epistemic subject self-conscious and situated in a consciously experienced visual world, but without the phenomenal element of "epistemic seeing", because the quality of being a "visually-perceptually knowing self" would be selectively eliminated (e.g. Lamme 2003; Vandenbroucke et al. 2014).

So there are many ways in which Building Block 3 for the first working concept of "mental action" I am trying to construct could be empirically investigated. For something to be a representation means that misrepresentation is always possible (Dretske 1986). An EAM, if more than a mere phenomenal artifact, could always be suboptimal or false, a *bad* model, for example in cases of self-deception or delusion where we find many examples of human beings exhibiting a robust phenomenology of "knowing that they know" while from a third-person perspective they clearly don't (Section 2.2; Picard 2013). This point brings out another important aspect: if the relevant layer in the human self-model exists, then we can now think of it as something that is *itself* continuously optimized via hierarchically structured, context-sensitive forms of prediction error minimization. What it predicts are not so much events on a sensory sheet, but expected outcomes of epistemic self-control — its own epistemic properties as represented by the epistemic value of future states of the organism's self-model (see the next section).

Subpersonal events, if integrated with the EAM, can afterwards be attributed to the person as a whole. In this way we could perhaps say that mind wandering is model-free epistemic self-control, whereas mental agency is model-based epistemic self-control. If future research shows that the EAM is a real and distinct part of a cortical-control hierarchy characterized by causal powers and a functional role of its own, then the conscious model of the "knowing self" could be interestingly described as an internal representation of a good epistemic self-regulator. Mental action is exactly the special case of cybernetic self-regulation for which there is no non-neural body or "world in between", where what counts is the epistemic content of cognitive affordances, and the selective exploitation of pre-existing internal models created in an ongoing process of spontaneous activation. For this reason, "instrumental" and "epistemic" inference almost converge for EAM-optimization: the causes revealed in this process are only causes of hierarchically deep perturbations; the causal structure to be found comprises only pre-existing internal models.

On a purely functional level of analysis, we can then conclude that mental action is predictive control not of events on some sensory sheet, but of effective connectivity in deeper hierarchical levels. It is an intelligent way of sculpting patterns of effective connectivity. As mental actions are M-autonomous conscious events, they are marked out by a capacity for veto control plus explicit goal representation. What is still lacking is their "intentional object" — the satisfaction conditions to which they are direct-

ed. From a more analytical, third-person perspective we should now ask: what could be the abstract property that is represented as the "goal" of mental actions?

### 3.4 Building Block 4: Epistemic Value

Here is what Karl Friston and colleagues write about the exploitation/exploration distinction:

> In this setting, action reduces the difference between current and (unsurprising) goal states that are defined much like cybernetic formulations (Miller, Galanter, & Pribram, 1960). This difference can be reduced in two ways. First, by executing a pragmatic action, that fulfils goals directly (i.e., exploitation); for example, by visiting a known reward site in the context of foraging. Second, by performing an epistemic action (i.e., exploration) to disclose information that enables pragmatic action in the long run; for example, exploring a maze to discover unknown reward sites (Kirsh & Maglio, 1994). Clearly, most behavior has both pragmatic and epistemic aspects. (Friston et al. 2015, p. 2)

It is very tempting to think of high-level symbolic cognition, as well as introspective attention, as a process of "epistemic foraging in one's own inner world" (pragmatic *mental* action), cyclically re-visiting the well-known stream of internal affordances. And perhaps some types of mind wandering can be conceptualized as a form of epistemic but unintentional explorative behavior, creating ever new trajectories through the maze of one's own mind — for example, in an attempt to discover "unknown reward sites" by mental data-mining. Obviously, for any system which, as with human beings, already *has* a very rich and hierarchically deep internal model of reality, this model itself becomes an important resource, something that can be continuously exploited and explored — at least whenever there are no pressing problems presented by the organism's physical environment. In such situations our internal task domain becomes causally dominant, as it were, immediately driving us into epistemic foraging or into an automatic exploration of potentially unknown aspects of our inner reward landscape, of possible positive hedonic experiences. Of course, CA, AA and mind wandering can sometimes be more reward-seeking and sometimes more knowledge-seeking. But for mental action, "epistemic" and "pragmatic" aspects converge, because the "exploitation-exploration dilemma" is very small:[9] here, we find an exclusively brain-based form of active, epistemic self-control.

Recall that according to PP, the brain has no direct access to the causal structure of its own body or the external environment — it has to infer them via interoception (i.e. visceroception and proprioception) and exteroceptive sensation. This raises the following question: what parts of the world can be accessed by *neither* exteroceptive *nor* interoceptive predictive processing? In principle, what relevant epistemic target lies outside transductor space and the space of causal interaction via active inference? One general answer is: the brain itself; the neural body. The brain is blind to itself because it has no self-directed receptor system. The brain is not part of any receptive field and, in inferring properties of itself, it cannot use representational hierarchies that bottom out into sensory sheets. Therefore, should it ever be necessary or adaptive for the brain to access more abstract properties of its *own* causal structure, then active inference using the non-neural body cannot help. To be more precise, above a certain level of complexity there will be "inferentially encapsulated"[10] but epistemically relevant causal

9 In more traditional philosophical terminology, if we take the concept of "theoretical intentionality" as referring to all mental states whose content-specifier consists of truth conditions (i.e. directedness towards an epistemic goal, as in mental states of thinking, believing, etc.) and "practical intentionality" as constituted by all states where the content-specifier consists of satisfaction conditions (i.e. directedness towards an action goal, as in mental states of willing, desiring, etc.), then mental action is special in that it falls under both concepts at the same time.

10 I owe this term to Anil Seth. I propose that this type of inferential encapsulation from the non-neural body is very close to the classical Cartesian idea of "the mental". What Descartes really discovered was that there are two fundamentally different layers in the phenomenal self-model of *Homo sapiens*, one in which events are coded as having spatial *and* temporal properties (the internal model of the non-neural body) and one in which events are coded as temporal, but *not* related to a body-centered spatial frame of reference any more (i.e. the internal model of the inferentially encapsulated parts of the neural body).

properties of the neural body which are not characterized by direct interaction with sensory or motor surfaces. On the control or "instrumental" side, this fact becomes a problem if the brain needs to control its own functional profile. Mental action is exactly what achieves this, because it is a form of self-control aimed mainly at effective connectivity, and not at properties of the non-neural body plant. If you will, it is an attempt to solve the problem of "neural disembodiment", an attempt to take autonomous self-control to a new level.

The good regulator theorem (Conant and Ashby 2007) says that the relevant structure must instantiate a model of the system to be controlled, where the system includes the body plus causal interaction space in the world. What exactly would be the system to be controlled in the special case of epistemic agency? It is the level of knowledge the system possesses and the epistemic quality of its world model, in the human case as physically realized by its neural body. Generally speaking, we could say that for epistemic self-control, the *goal state* of the regulator is the optimal level of knowledge the system could possess, the overall epistemic quality of its world model. For the special case of mental epistemic self-control, however, we could redefine the question, and ask: what is the "good epistemic regulator" or non-sensory "model of the system" that is instantiated? It is tempting to say that this would have to be the EAM, because what matters is the overall epistemic quality of the cognitive *self*-model. We should be careful, however, not to over-interpret the conclusion of Conant and Ashby's argument (I am grateful to Michael Anderson for pointing this out). All that follows is that the regulator models its plant, in the sense that its control structures bear systematic isomorphisms to its states. There is no implication that the control system actually uses an explicit model *of* the plant as a separate element of its control architecture. For the special case of epistemic self-regulation as stated by the cognitive affordance hypothesis, this would mean that mind wandering is self-regulation on the mental level *without* an explicit model of the "epistemic plant", whereas mental action is precisely the rare, special case where we actually operate under a conscious and explicit model of ourselves as epistemic agents.

For individual mental actions, we could then say that their *target* is the expected epistemic value of own cognitive states, under counterfactual outcomes. Karl Friston and colleagues write, "Epistemic value is the expected information gain under predicted outcomes" (Friston et al. 2015, p. 6). This, then, would have to be what is maximized. Geoffrey Hinton (2005, p. 1765 quoted from Clark 2016, p. 196) said that a perceptual state can be seen as nothing but "the state of a hypothetical world in which a high-level internal representation would constitute veridical perception". Analogously, we can now say that a *cognitive* state is the state of a hypothetical knowing *self* in which a high-level internal *self*-representation (namely, the EAM) would itself carry high epistemic value because it would constitute veridical self-knowledge. If so, this would explain how EAMs could serve a causal role in a control hierarchy: they could function as explicit representations of the current level of optimization, of more or less successfully achieved epistemic goal states. I propose that for all non-reward-oriented forms of mental action, epistemic value may be the abstract property that constitutes their goal state (what philosophers call the "intentional object"). This leaves us with a first working concept of "epistemic mental action" under the PP approach: it is the predictive control of effective connectivity aimed at optimizing the epistemic value of attentional and cognitive states embedded in the system's phenomenal self-model, accompanied by the emergence of a new content layer on introspectively available levels of the hierarchy, namely, the EAM.

## 4 Open Questions

Let me end by sketching what I take, on a conceptual level, to be the three most important unresolved issues, perhaps the most relevant targets for future research. They refer to the homunculus problem, the metaphysics of goal-state selection and action initiation, and the possibility of eliminating the concept of "mental action" altogether.

### 4.1 Problem 1: The Bubble in the Bubble

As long as we model mental action analogously to active inference, we will always introduce a second statistical boundary into the brain's model of reality, a new evidentiary boundary potentially leading to a vicious regress of inferential seclusion (see Hohwy 2016; Hohwy 2017 and Clark 2017 for an excellent discussion). If, as I have proposed, what is predicted and actively controlled via changes of effective connectivity is the epistemic value of attentional and cognitive states as embedded in the self-model, then the *causes* of such shifts in epistemic value will lie hidden behind a Markov blanket — they will constitute an inner environment from which the mental agent is secluded. For example, if the cognitive affordance hypothesis points in the right direction, then this internal environment is constituted by the dynamic affordance landscape set up by the mind wandering network. What we call "thinking" is the process of trying to predict and control the epistemic dynamics unfolding in this network, of "surfing inner uncertainty" (Clark 2016) — the constant mental entropy brought about by the stream of spontaneously occurring thought, unbidden memories, task-unrelated goal-state simulations, and so on.

This threatens to create an acerbated version of the homunculus problem, because it is now tempting to posit a little "epistemic surfer" in our head, a truly disembodied entity that always tries to stay ahead of the next cognitive wave and maximizes evidence for its own existence. Mental action would therefore create a "bubble in the bubble" in the brain's model of reality, where all we need to predict our little mental agent's behavior is information about events from *their* side of the statistical veil, because an observation of events within the smaller bubble of mental agency and prior expectations about the cognitive system will suffice to predict his or her epistemic behavior. From a scientific third-person perspective, all causes beyond the smaller Markov blanket would then be uninformative for the purpose of predicting mental actions and for a scientific understanding our own cognitive phenomenology.

Problem 1 consists in avoiding new versions of the homunculus problem by proliferating evidentiary boundaries, action types, and nested explanatory-evidentiary circles (Hohwy 2016, pp. 5 & 12). We do not want to build epistemological Matryoshka dolls dressed in Markov blankets, and we also want to avoid setting the stage for bizarre, new, and *internalist* versions of the "extended mind debate" (Menary 2010). Therefore, the number of agents should not be multiplied without necessity.

### 4.2 Problem 2: Dropping Naïve Realism About "Goals"

In the conceptual framework I am proposing, action initiation becomes a functionally adequate form of self-deception. One open question is if this point holds for bodily and mental actions in the same way. Let me explain.

Many neuroscientists use the term "goal" in a naïvely-realistic manner, as if "goals" were something we could in principle find out there in the world. But goals are not given. It is important to understand that, from a scientific third-person perspective, all goal-representations are misrepresentations. Viewed from the philosophical perspective of a thoroughgoing naturalism, there are no intrinsically normative facts, quite simply because there are no states in the physical world that could count as "goal states" in any more rigorous, metaphysical sense. There are no essences, no intrinsic values which remain invariant across all contexts and situations. If you will, the harshness of naturalist metaphysics exactly consists in the point that nothing has *intrinsic* value, because any possible or actual fact is only normative *relative* to a certain organism, biological population, self-sustaining robot, or other such entity. Of course, a naturally evolved information-processing system may *represent* its own procreative success or the sustaining of its own existence as an intrinsically valuable goal state. It may generate an internal model of the world and of itself in it, in which organismic integrity and homeostatic stability *are* ultimate normative facts, and as this model is physically grounded any such process will have direct causal consequences — for example, the system may turn into a "crook-

ed scientist" (Bruineberg et al. 2016, Section 3) and begin maximizing evidence for its own existence (Friston 2011, p. 117; Limanowski and Blankenburg 2013). Clearly, any self-model successfully integrating (mis)representations of (non-existing) goal states as possessing intrinsic normativity will lead to the instantiation of new functional properties, which can then be selected for in the process of natural evolution (Metzinger 2003).

Call this "the self-deception model of goal selection and action initiation". This may refer to an important step in the biological history of mindedness, because to manifest higher levels of intelligent behavior any naturally evolved agent has to solve two major problems: autonomous self-control and autonomous self-motivation. Biological organisms like ourselves solve the second problem by *hallucinating* goals. One aspect that makes the theory of predictive processing attractive from a philosophical perspective is that it offers a concrete mechanism for what has in the past been called "conscious volition" or "deliberate action initiation": we can dis-attend to the current state of the body, thereby enabling proprioceptive predictions — which are currently *false* representations of body and limb position in space — to "become real", initiating a new cycle of active inference. I think that attenuating sensory input by reallocating expectations for prediction-error precisions across different levels of the generative model (see also Limanowski 2017) is exactly what allows us to "hallucinate goals" — blocking out ascending sensory-prediction error while simultaneously setting a dynamic, top-down context in hierarchical motor control by maximizing the precision of (and hence confidence in) a more *abstract* type of belief. That belief may be normative, and it may be false.

Whenever we begin to hallucinate a goal, we are actively optimizing a high-level, multi- or amodal model describing a possible path into counterfactual states of the organism, where "optimizing" means conferring intrinsic value on it. The self-model theory (Metzinger 2003) would say that this crucial step happens when an internal representation of some state of the world as organism-salient — as a positive state of affairs *relative* to homeostasis, procreative success, long-term self-sustainment and so on — turns into an *intrinsic* (i.e. context-invariant) value for the system whenever it is functionally embedded in its intrinsic self-representation, when it becomes integrated with the deeper, phenomenally transparent, and more context-invariant levels of the phenomenal self-model (Blanke and Metzinger 2009). Before movement onset, sensorimotor estimates are misrepresentations. The moment of action initiation is the moment in which the counterfactuality disappears from the content. What begins as an imagined bodily movement gradually becomes a real bodily movement; phenomenal opacity is transformed into phenomenal transparency; an allocentric representation of a successfully terminated action in a larger predictive horizon is gradually transposed into a shorter timescale, until it becomes "embodied" in an egocentric frame of reference leading to overt action. The intrinsic normativity in the original model's content, however, remains a hallucination — it is not an objective property of the successfully terminated motor pattern depicted by the original allocentric representation. In addition, if we hold on to the notion of active inference, then even the low-level sensory precision estimates generated in the process are systematically misrepresentational *and* beneficial for the agent, simply because they enable action. I believe this general conceptual point may be of considerable philosophical relevance, for example, because, in the words of Wanja Wiese, it draws our attention to the fact that "systematically beneficial misrepresentations may lie at the heart of our neural architecture" (Wiese 2016, Section 8).

Problem 2 consists in developing a fine-grained phenomenological analysis of the process just sketched without assuming a realist metaphysics of "goals" and "intrinsic value" — or reintroducing nested Matryoshka dolls and mental homunculi (Problem 1). Obviously, arriving at a deeper understanding of goal-state selection and action initiation is of particular relevance for the special case of *mental* actions and *epistemic* goal states. It would therefore be of importance to the construction of computational models that can later explain the emergence of the specific phenomenological profile going along with transiently "hallucinating intrinsic value".

### 4.3 Problem 3: Phenomenology and the Metaphysics of Mental Action

It might turn out that the problem of mental action just cannot be solved in a satisfactory manner which simultaneously does justice to all its relevant aspects — the first horn of the dilemma. One classical and respectable option is to *eliminate* the concept of "mental action" from our scientific taxonomy (Churchland 1981; Churchland 1989). "Mental action" would then have to be dissolved into a set of empirically grounded successor concepts. But even if this could be done without shrinking the explanatory scope and predictive power of future theories developed under the PP-approach, it would still leave us with an extremely counterintuitive perspective on our own phenomenology — the second horn of dilemma. Let us take a look at this second horn first.

Here is the sketch of a three-phase model, by which an eliminativist could try to account for our robust phenomenology of mental agency:

- Phase 1 is termed "**Deep Social Embodiment**" (indirectly alluding to the terminology developed by Lisa Quadt, see also Quadt 2017), and it begins with the early prenatal development of the fetus' brain. We often ignore the fact that the human brain finds itself in a situation of "nested double embodiment" right from the origin. The very first social interactions it has to predict and control are shaped by this special, convoluted form of embodiment. Some of these interactions may be motor, but many take place on a molecular level. They are special variants of what Quadt calls interactive inference: the fetus' brain has to predictively control *two* interoceptive environments, namely, a) the stimulus landscape internal to the fetus body itself, and b) the chemical landscape surrounding it, which is in turn predictively controlled by the mother's brain in a process Anil Seth has termed "interoceptive inference". Viewed from the epistemic perspective of the mother's brain, the fetus' body is a robust and new interoceptive signal source. Computationally, there are two overlapping task domains, and we can now view pregnancy as an escalating conflict between two control systems leading to a fundamentally competitive dynamic, ultimately based on the antagonistic coevolution of fetal and maternal genes (e.g. Haig 2015). This leads to the emergence and consolidation of a second self-model in what is still a single, metabolically autonomous organism — but this time in the brain of the fetus — ultimately achieving and culminating in the instantiation of a new phenomenal property — an individual "sense of presence" (Blanke and Metzinger 2009; Seth et al. 2011).

- Phase 2 is "**Amnestic Enculturation**". First, it is important to note that CA as well as AA are enculturated processes (Fabry 2015; Fabry 2017). One relevant cultural practice is that human beings educate their children by ascribing the capacity for veto control ("You *could* have refrained from doing this!") and ultimate origination ("You *could* have done otherwise!") right from the very beginning, and at a stage where, from a scientific perspective, young children almost certainly do not possess these capacities. According to the PP approach, this cultural practice will automatically become internally modelled, and the overall process will influence the gradual development of the child's EAM. All socially embedded PP systems have to extract and exploit the causal structure of their sociocultural niche — they have to predict their caregivers' behavior — and the best resource for the generation and updating of hypotheses they have is their own internal self-model.
  There is empirical evidence demonstrating the social transmission of the experience of agency (Khalighinejad et al. 2016), and not only do children observe adults controlling external and internal events, their social context is constituted by pre-enculturated agents already operating *under the belief* of having the capacities of mental veto control and ultimate origination themselves (Hauf and Prinz 2005). In this way, socially constituted norms for autonomy and successful self-control become internalized via social interactive inference. There is, however, an important second aspect which the eliminativist adherent of the PP approach could highlight: It is only much later, between the ages of 2 and 4 years, that the *autobiographical* self-model comes

onstream. This leads to the well-studied effect of infantile amnesia: the inability of adult human beings to retrieve episodic memories from this period. For the grown-up human being, the process of amnestic enculturation itself will never be introspectively available via the phenomenal self-model.

- I call Phase 3 the period of the "**Agentive Narrative**", because it is dominated by an inner life narrative: the existence of an autobiographical self-model creating an illusion of transtemporal identity (Metzinger 2013a, p. 5). Of course, there really is no such thing as a "narrative self"; our inner life narrative has no author, no mysterious self-behind-the-self that could function as the narrator, it too emerges out of a continuous process of dynamic self-organization and social interaction, leading to an ever-changing autobiographical self-model. But the process just sketched has an important phenomenological effect: when autobiographical memory begins to consolidate, a strong hyperprior for autonomy and moral responsibility is already firmly established. We believe and automatically predict that we (and others) have the capacity for veto control and ultimate origination, because "it has always been like that, since we can remember". We are embodied models of our early social environment. Our phenomenal self-model reflects implicit expectations of our primordial sociocultural niche and says that we have always been autonomous, morally responsible agents. This robust, self-related hyperprior has been installed early through highly reliable social feedback and it is a major factor in determining the phenomenology of self-consciousness and M-autonomy.

Now let us look at our target phenomenon of mental action against this background. According to subjective experience, and as pointed out in Section 2.3, the sudden emergence of M-autonomous epistemic self-control is an unpredicted internal event. Strictly speaking, what we call "agency" refers to an *interpretation* of this fact as "initiation" or "origination": it is the activation of an internal self-model trying to explain away the surprise involved in suddenly achieving autonomous self-control, by creating an explicit, supramodal representation of an entity capable of ultimate origination, veto control, and spontaneous self-causation. It likely has internal and external aspects, because it will involve internal simulations of cultural practices, as explained above. It is therefore something we learn before our conscious inner life-history has even begun. In addition, as Patrick Haggard has pointed out, the phenomenology of intending to act is actually quite thin and evasive, and lacks the vivid subjective quality of, for example, visual experience (Haggard 2005, p. 291). One might certainly argue that if we introspect as closely as we can and on a very fine-grained timescale, there really only is a phenomenological element of "surprise" going along with action initiation. The human adult's conscious model of the "self" involving ultimate origination and self-causation could therefore be an enculturated *post-hoc* confabulation. It could be computationally described as a causal-inference illusion that has become part of a sociocultural niche-model. Ultimately, it is based on an internalization of social interactions and deeply engrained language games that lead to self-evidencing habits of a) shallow introspection, and b) unreflected but "theory-contaminated" phenomenological self-reports. Therefore, the second horn of the dilemma may not be so counterintuitive after all — if we dare to take a closer look.

However, if we really take the option of eliminating the concept of "mental action", then, under PP, we suddenly face the problem of *bodily* action initiation: conventionally, mental actions explain the initiation of overt bodily action via a reweighting of precision expectations from somatosensory sources to goal-representations by generating descending proprioceptive predictions, which I called, in the introduction, "self-fulfilling motor fantasies". If AA is a causally inert phenomenal artifact and only unintentional mental *behaviors* (like the automatic, subpersonal reweighting of precision expectations just mentioned) do exist, then we have to find a different solution. We might say that a) bodily actions do exist, b) mental actions never existed, and c) our picture of bodily action initiation has changed in a way that only leaves us with a very weak and impoverished notion of the term "action". If

— now returning to the first horn of the dilemma — we want to say that mental actions are respectable citizens of a scientifically grounded ontology, then we need a model of *mental action initiation* for the special case of attentional agency, and we need one that avoids a vicious regress. Introducing higher-order precision shifts, for example by positing third-order precision-control targeting processes of second-order precision optimization ("second-order AA"), would create exactly such a vicious regress (Problem 1). It would also give birth to a lineage of nested Matryoshka dolls.

I have therefore proposed a non-circular model for the special case of cognitive agency: the cognitive affordance hypothesis. Cognitive action initiation is explained by the navigation of an internal affordance landscape which is continuously set up by the mind wandering network (Section 3.2). But at present it is not clear if the mind wandering network could play the same role for the initiation of *attentional* action, an equivalent of providing an automatic stream of goal-representations which compete for motor control by inducing sensory attenuation (Problem 2). So we may have a model of action initiation for CA, but not for AA. I take this to be another highly relevant future target for interdisciplinary research.

Problem 3 lies in deciding which horn of the dilemma to take. The realist option continues to treat mental actions as part of our scientific ontology, but bears the explanatory burden of developing a multi-level theory of goal-state selection and action initiation spanning all levels of the hierarchy — ranging from attentional to overt bodily agency — by solving Problem 1 and Problem 2. The eliminativist option has to dissolve the concept of "mental action" into a series of successor concepts possessing at least equal unificatory, predictive, and explanatory power — while making intelligible why, for centuries and in our very own autophenomenological reports, we have described our inner life so falsely.

## 5 Conclusion

Are there mental actions *at all*? I will not answer this question here. My overarching goal in this article has been to put the problem of mental action into explicit focus. But as we have now seen, this problem has many philosophical and empirical facets. It must be clearly stated that to date, there is no satisfying theoretical account of mental action under the PP approach. I have offered a series of new conceptual instruments and connected them to four central ideas discussed in recent philosophical and scientific research on PP. My aim was to integrate core semantic elements in order to arrive at a first working concept for "mental action", a hypothetical construct which can now be further developed and refined. According to this first notion, mental actions are a specific form of predictive control of effective connectivity, accompanied and possibly even functionally mediated by what I have termed the EAM. They are aimed at increasing the epistemic value of pre-existing states in the conscious self-model, without causally looping through sensory sheets or using the non-neural body as an instrument for active inference. I have sketched four empirically testable hypotheses, assuming that what is most needed at this stage are more fine-grained computational models yielding testable predictions. [11]

11 I am deeply indebted to Jona Vance, Michael Anderson, Carsten Korth, Jakob Hohwy, Giovanni Pezzulo, Iuliia Pliushch, Anil Seth, and Wanja Wiese for critical discussion and a considerable number of extremely helpful comments and Lucy Mayne for equally helpful comments plus excellent editorial help with the English version of this text. I have learned a lot from all of them. They are not responsible for the shortcomings of the final version of this paper.

## References


Amer, T., Campbell, K. L. & Hasher, L. (2016). Cognitive control as a double-edged sword. *Trends in Cognitive Sciences, 20* (12), 905–915. https://dx.doi.org/10.1016/j.tics.2016.10.002.

Anderson, M. L. (2015). Précis of After phrenology: Neural reuse and the interactive brain. *Behavioral and Brain Sciences, 16*, 1–22.

Andrews-Hanna, J. R., Reidler, J. S., Huang, C. & Buckner, R. L. (2010). Evidence for the default network's role in spontaneous cognition. *Journal of Neurophysiology, 104* (1), 322–335.

Axelrod, V., Rees, G., Lavidor, M. & Bar, M. (2015). Increasing propensity to mind-wander with transcranial direct current stimulation. *Proceedings of the National Academy of Sciences of the United States of America, 112* (11), 3314–3319. https://dx.doi.org/10.1073/pnas.1421435112.

Barsalou, L. (2016). Can cognition be reduced to action? In A. K. Engel, K. J. Friston & D. Kragic (Eds.) *The pragmatic turn: Toward action-oriented views in cognitive science*.

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

Bortolotti, L. (2015). The epistemic innocence of motivated delusions. *Consciousness and Cognition, 33*, 490–499.

Brass, M. & Haggard, P. (2007). To do or not to do: The neural signature of self-control. *The Journal of Neuroscience, 27* (34), 9141–9145.

Broadway, J. M., Zedelius, C. M., Mooneyham, B. W., Mrazek, M. D. & Schooler, J. W. (2015). Stimulating minds to wander. *Proceedings of the National Academy of Sciences of the United States of America, 112* (11), 3182–3183. https://dx.doi.org/10.1073/pnas.1503093112.

Bruineberg, J. (2017). Active inference and the primacy of the 'I can'. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

Bruineberg, J., Kiverstein, J. & Rietveld, E. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*. https://dx.doi.org/10.1007/s11229-016-1239-1.

Buckner, R. L. & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11* (2), 49–57. https://dx.doi.org/10.1016/j.tics.2006.11.004.

Buckner, R. L., Andrews–Hanna, J. R. & Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences, 1124* (1), 1–38.

Butz, M. V. (2016). Toward a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology, 7*, 925. https://dx.doi.org/10.3389/fpsyg.2016.00925.

Campbell-Meiklejohn, D. K., Woolrich, M. W., Passingham, R. E. & Rogers, R. D. (2008). Knowing when to stop: The brain mechanisms of chasing losses. *Biological Psychiatry, 63* (3), 293–300.

Christoff, K. (2012). Undirected thought: Neural determinants and correlates. *Brain Research, 1428*, 51–59. https://dx.doi.org/10.1016/j.brainres.2011.09.060.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R. & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. *Proceedings of the National Academy of Sciences, 106* (21), 8719–8724.

Christoff, K., Irving, Z. C., Fox, K. C. R., Spreng, R. N. & Andrews-Hanna, J. R. (2016). Mind-wandering as spontaneous thought: A dynamic framework. *Nature Reviews Neuroscience, 17*, 718–731. https://dx.doi.org/10.1038/nrn.2016.113.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy, 78* (2), 67–90.

——(1989). *A neurocomputational perspective: The nature of mind and the structure of science.* MIT Press.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362* (1485), 1585–1599.

Cisek, P. & Kalaska, J. F. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. *Neuron, 45* (5), 801–814.

—— (2010). Neural mechanisms for interacting with a world full of action choices. *Annual Review of Neuroscience, 33*, 269–298.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

——(2017). How to knit your own Markov blanket: Resisting the second law with metamorphic minds. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

Cohen, O., Koppel, M., Malach, R. & Friedman, D. (2014). Controlling an avatar by thought using real-time fMRI. *Journal of Neural Engineering, 11* (3), 035006.

Cohen, O., Druon, S., Lengagne, S., Mendelsohn, A., Malach, R., Kheddar, A. & Friedman, D. (2014). FMRI-based robotic embodiment: Controlling a humanoid robot by thought using real-time fMRI. *Presence: Teleoperators and Virtual Environments, 23* (3), 229–241. https://dx.doi.org/10.1162/PRES_a_00191.

Cole, M. W. & Schneider, W. (2007). The cognitive control network: Integrated cortical regions with dissociable functions. *NeuroImage, 37* (1), 343–360.

Cole, M.W., Yarkoni, T., Repovš, G., Anticevic, A. & Braver, T. S. (2012). Global connectivity of prefrontal cortex predicts cognitive control and intelligence. *The Journal of Neuroscience, 32* (26), 8988–8999.

Cole, M. W., Reynolds, J.R., Power, J.D., Repovs, G., Anticevic, A. & Braver, T.S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience, 16* (9), 1348–1355.

Cole, M. W., Repovs, G. & Anticevic, A. (2014). The fronto-parietal control system: A central role in mental health. *The Neuroscientist*.

Conant, R. C. & Ashby, R. (2007). Every good regulator of a system must be a model of that system. *International Journal of Systems Science, 1* (2), 89–97. https://dx.doi.org/10.1080/00207727008920220.

Davey, C. G., Pujol, J. & Harrison, B. J. (2016). Mapping the self in the brain's default mode network. *NeuroImage, 132*, 390–397. https://dx.doi.org/10.1016/j.neuroimage.2016.02.022.

Davidson, D. (2001). *Essays on actions and events: Philosophical essays.* Oxford: Oxford University Press.

Dretske, F. (1986). Misrepresentation. In R. Bogdan (Ed.) *Belief: Form, content, and function* (pp. 17–36). Oxford, Oxford University Press.

——— (1988). *Explaining behavior: Reasons in a world of causes.* Cambridge: Cambridge University Press.

Engel, A. K., Maye, A., Kurthen, M. & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences, 17* (5), 202–209.

Engel, A. K., Friston, K. J. & Kragic, D. (2016). *The pragmatic turn: Toward action-oriented views in cognitive science.*

Fabry, R. E. (2015). Enriching the notion of enculturation: Cognitive integration, predictive processing, and the case of reading acquisition. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. https://dx.doi.org/10.15502/9783958571143.

———(2017). Predictive processing and cognitive development. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing.*

Filevich, E., Kühn, S. & Haggard, P. (2012). Intentional inhibition in human action: The power of 'no'. *Neuroscience & Biobehavioral Reviews, 36* (4), 1107–1118.

———(2013). There is no free won't: Antecedent brain activity predicts decisions to inhibit. *PLoS ONE, 8* (2), e53053.

Fox, K. C. R., Spreng, R. N., Ellamil, M., Andrews-Hanna, J. R. & Christoff, K. (2015). The wandering brain: Meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *NeuroImage, 111*, 611–621.

Friston, K. (2011). Embodied inference: Or I think therefore I am, if I am what I think. *The implications of embodiment (Cognition and Communication)*, 89–125.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T. & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 369* (1655). https://dx.doi.org/10.1098/rstb.2013.0481.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T. & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience, 6* (4), 187–214. https://dx.doi.org/10.1080/17588928.2015.1020053.

Haggard, P. (2005). Conscious intention and motor cognition. *Trends in Cognitive Sciences, 9* (6), 290–295. https://dx.doi.org/10.1016/j.tics.2005.04.012.

Haig, D. (2015). Maternal-fetal conflict, genomic imprinting and mammalian vulnerabilities to cancer. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 370* (1673). https://dx.doi.org/10.1098/rstb.2014.0178.

Hauf, P. & Prinz, W. (2005). The understanding of own and others' actions during infancy: "You-like-me" or "Me-like-you"? *Interaction Studies, 6* (3), 429–445.

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

———(2016). The self-evidencing brain. *Noûs, 50* (2), 259–285. https://dx.doi.org/10.1111/nous.12062.

——— (2017). How to entrain your evil demon. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing.*

Khalighinejad, N., Bahrami, B., Caspar, E. A. & Haggard, P. (2016). Social transmission of experience of agency: An experimental study. *Frontiers in Psychology, 7* (974), 313. https://dx.doi.org/10.3389/fpsyg.2016.01315.

Kühn, S., Haggard, P. & Brass, M. (2009). Intentional inhibition: How the "veto–area" exerts control. *Human Brain Mapping, 30* (9), 2834–2843.

Lamme, V. A. F. (2003). Why visual attention and awareness are different. *Trends in Cognitive Sciences, 7* (1), 12–18.

Lenggenhager, B., Tadi, T., Metzinger, T. & Blanke, O. (2007). Video ergo sum: Manipulating bodily self-consciousness. *Science, 317* (5841), 1096–1099. https://dx.doi.org/10.1126/science.1143439.

Limanowski, J. (2017). (Dis-)attending to the body. Action and self-experience in the active inference framework. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

Limanowski, J. & Blankenburg, F. (2013). Minimal self-models and the free energy principle.

Mantini, D. & Vanduffel, W. (2012). Emerging roles of the brain's default network. *The Neuroscientist, 19* (1), 76–87. https://dx.doi.org/10.1177/1073858412446202.

Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G., Eickhoff, S. B., Castellanos, F. X. & Petrides, M. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, 12574–12579.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T. & Macrae, C. N. (2007). Wandering minds: The default network and stimulus-independent thought. *Science, 315* (5810), 393–395.

Mattar, M. G., Cole, M. W., Thompson-Schill, S. L. & Bassett, D. S. (2015). A functional cartography of cognitive systems. *PLoS Computational Biology, 11* (12), e1004533.

Medea, B., Karapanagiotidis, T., Konishi, M., Ottaviani, C., Margulies, D., Bernasconi, A., Bernasconi, N., Bernhardt, B. C., Jefferies, E. & Smallwood, J. (2016). How do we decide what to do? Resting-state connectivity patterns and components of self-generated thought linked to the development of more concrete personal goals. *Experimental Brain Research*. https://dx.doi.org/10.1007/s00221-016-4729-y.

Menary, R. (Ed.) (2010). *The extended mind.* Cambridge, MA: MIT Press.

Metzinger, T. (1995). *Conscious experience.* Exeter, UK: Imprint Academic.

——— (2003). *Being no one: The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

——— (2004). Précis of "Being no one". *PSYCHE - An Interdisciplinary Journal of Research on Consciousness, 11* (5), 1–35. http://psyche.cs.monash.edu.au/symposia/metzinger/precis.pdf.

——— (2006). Conscious volition and mental representation: Towards a more fine-grained analysis. *Disorders of Volition*, 19–48.

——— (2008). Empirical perspectives from the self-model theory of subjectivity: A brief summary with examples. *Progress in Brain Research, 168*, 215–278.

——— (2013a). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology, 4*, 931. https://dx.doi.org/10.3389/fpsyg.2013.00931.

——— (2013b). Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. *Frontiers in Psychology, 4*. https://dx.doi.org/10.3389/fpsyg.2013.00746.

——— (2015). M-autonomy. *Journal of Consciousness Studies, 22* (11-12), 270–302.

——— (2017). Why is mind wandering interesting for philosophers? In K. C. Fox & K. Christoff (Eds.) *The Oxford handbook of spontaneous thought*.

Mooneyham, B. W. & Schooler, J. W. (2013). The costs and benefits of mind-wandering: A review. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 67* (1), 11–18. https://dx.doi.org/10.1037/a0031569.

Niendam, T. A., Laird, A. R., Ray, K. L., Dean, Y. M., Glahn, D. C. & Carter, C. S. (2012). Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cognitive, Affective, & Behavioral Neuroscience, 12* (2), 241–268.

Pezzulo, G. (2012). An active inference view of cognitive control. *Frontiers in Psychology, 3*, 478. https://dx.doi.org/10.3389/fpsyg.2012.00478.

——— (2016). The contribution of pragmatic skills to cognition and its development: Common perspectives and disagreement: Common perspectives and disagreement. In A. K. Engel, K. J. Friston & D. Kragic (Eds.) *The pragmatic turn: Toward action-oriented views in cognitive science* (pp. 19–33).

Pezzulo, G., Cartoni, E., Rigoli, F., Pio-Lopez, L. & Friston, K. (2016). Active inference, epistemic value, and vicarious trial and error. *Learning & Memory, 23* (7), 322–338.

Picard, F. (2013). State of belief, subjective certainty and bliss as a product of cortical dysfunction. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior, 49* (9), 2494–2500. https://dx.doi.org/10.1016/j.cortex.2013.01.006.

Picard, F., Scavarda, D. & Bartolomei, F. (2013). Induction of a sense of bliss by electrical stimulation of the anterior insula. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior, 49* (10), 2935–2937. https://dx.doi.org/10.1016/j.cortex.2013.08.013.

Pliushch, I. (2017). The overtone model of self-deception. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

——— (in press). *Self-deception within the predictive coding framework*.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.) *Art, mind, and religion* University of Pittsburgh Press.

——— (1975). *Mind, language and reality (vol. 2)*. Cambridge: Cambridge University Press.

Quadt, L. (2017). Action-oriented predictive processing and social cognition. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

Schooler, J. W., Smallwood, J., Christoff, K., Handy, T. C., Reichle, E. D. & Sayette, M. A. (2011). Meta-awareness, perceptual decoupling and the wandering mind. *Trends in Cognitive Sciences, 15* (7), 319–326.

Seli, P., Risko, E. F., Smilek, D. & Schacter, D. L. (2016). Mind-wandering with and without intention. *Trends in Cognitive Sciences, 20* (8), 605–617.

Seth, A. K. (2015a). Inference to the best prediction. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. https://dx.doi.org/10.15502/9783958570986. http://open-mind.net/papers/inference-to-the-best-prediction.

——— (2015b). The cybernetic Bayesian brain. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. https://dx.doi.org/10.15502/9783958570108.

Seth, A. K., Suzuki, K. & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology, 2*.

Smallwood, J. & Schooler, J. W. (2015). The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology, 66*, 487–518. https://dx.doi.org/10.1146/annurev-psych-010814-015331.

Spratling, M. W. (2016). Predictive coding as a model of cognition. *Cognitive Processing*, 1–27.

Spreng, R. N., Mar, R. A. & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21* (3), 489–510. https://dx.doi.org/10.1162/jocn.2008.21029.

Spreng, R. N., Stevens, W. D., Chamberlain, J. P., Gilmore, A. W. & Schacter, D. L. (2010). Default network activity, coupled with the frontoparietal control network, supports goal-directed cognition. *NeuroImage, 53* (1), 303–317. https://dx.doi.org/10.1016/j.neuroimage.2010.06.016.

Stawarczyk, D., Cassol, H. & D'Argembeau, A. (2013). Phenomenology of future-oriented mind-wandering episodes. *Frontiers in Psychology, 4*, 425.

Van de Cruys, S. (2017). Affective value in the predictive mind. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*.

Vandenbroucke, A. R. E., Fahrenfort, J. J., Sligte, I. G. & Lamme, V. A. F. (2014). Seeing without knowing: Neural signatures of perceptual inference in the absence of report. *Journal of Cognitive Neuroscience, 26* (5), 955–969. https://dx.doi.org/10.1162/jocn_a_00530.

Voss, U., Holzmann, R., Hobson, A., Paulus, W., Koppehele-Gossel, J., Klimke, A. & Nitsche, M. A. (2014). Induction of self awareness in dreams through frontal low current stimulation of gamma activity. *Nature Neuroscience, 17* (6), 810–812. https://dx.doi.org/10.1038/nn.3719.

Weissman, D. H., Roberts, K. C., Visscher, K. M. & Woldorff, M. G. (2006). The neural bases of momentary lapses in attention. *Nature Neuroscience, 9* (7), 971–978.

Wiese, W. (2016). Action is enabled by systematic misrepresentations. *Erkenntnis*. https://dx.doi.org/10.1007/s10670-016-9867-x.

Wilson, G. & Shpall, S. (2016). Action. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy* Metaphysics Research Lab, Stanford University.

Windt, J. M. (2015). *Dreaming: A conceptual framework for philosophy of mind and empirical research*. Cambridge, MA: MIT Press.

# Tracing the Roots of Cognition in Predictive Processing

Giovanni Pezzulo

Can PP (Predictive Processing) help us understand "the roots of cognition", and how we may have acquired (during evolution and/or development) our sophisticated cognitive abilities from the relatively simpler adaptive control mechanisms of our early evolutionary ancestors? Here I make the case that some cognitive operations may be constructed as *detached actions* — where the detachment process rests on the construction of generative PP models, which permit one to *internalize* action-environment dynamics. I provide three examples. The first example focuses on the role of internally-generated sequences of (hippocampal) neuronal activity across goal-directed navigation and detached tasks such as planning. This example illustrates how neuronal sequences (putatively forming an internal model for spatial navigation) may have a "dual use": they may support both overt navigation and covert cognitive operations while running, respectively, in stimulus-based and internally-generated modes. Furthermore, the latter (internally-generated) mode may be considered a form of internalization of the former (stimulus-based) mode. The second example focuses on actions to resolve epistemic uncertainty, and the formal similarity in PP between *epistemic actions* that are executed overtly (e.g., exploration) and those executed covertly (e.g., episodic retrieval). This example illustrates the possibility of defining *mental actions*, such as reducing one's uncertainty before making a choice, as internalized information foraging acts that have the same intentionality as externally-directed actions. The third example focuses on the detachment of cognitive goals such as "eating in a fancy restaurant" from homeostatic drives such as "being satiated". This example illustrates that by internalizing regulatory loops within hierarchical PP models, one can build cognitive goals that can in turn enjoy some form of detachment — for example, one can go to a restaurant or buy food even when one is not hungry. I discuss these examples in relation to alternative theories of how higher cognition originates from (or is independent of) action-perception loops, including various versions of action-oriented, embodied and enactivist views.

**Keywords**

Active inference | Embodied cognition | Internalization | Predictive processing | Reuse

**Acknowledgements**

I would like to thank Thomas Metzinger, Wanja Wiese, Lucy Mayne and the anonymous reviewers for useful suggestions and editorial help.

## 1 Introduction

Predictive processing (PP) — especially in its most comprehensive version, Karl Friston's *free-energy principle* (Friston 2010) — has recently become an influential framework for understanding brain activity and cognition across many disciplines, including systems neuroscience, cognitive science, philosophy, psychology and psychiatry. The most important constructs of PP — predictions and prediction errors, generative models, and precision — are increasingly mentioned in all these disciplines. This is sometimes in descriptive ways (i.e. to describe behavioural or neuronal regularities without committing to the ontological validity of these constructs) but is more often done with an implicit or explicit assumption that brains implement these mechanisms neurally.

PP suggests that the brain is a "prediction machine". However, PP is not a unitary theory but rather refers to a variety of approaches. These approaches need to specify, for example, which brain functions are predictive and which are not, what exactly the brain predicts (e.g., the unfolding of a visual scene or action-perception contingencies) and at which timescales, and which computational mechanisms

Pezzulo, G. (2017). Tracing the Roots of Cognition in Predictive Processing.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 20. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573215

(e.g., forward models, predictive coding) implement prediction and what their neuronal underpinnings are. One domain where PP was initially applied was (visual) perception. In this domain, *predictive coding* (Rao and Ballard 1999) emphasized the importance of a hierarchical (Bayesian) scheme, in which higher levels convey predictions to lower levels, and lower levels convey prediction errors to higher levels, with this process being iterated until prediction error is minimized, thereby disambiguating the most supported perceptual hypothesis. In this article, I will focus on *active inference* (under the Free Energy principle, (Friston 2010)), which starts from a predictive coding scheme but extends it to cover the domain of action control. Active inference uses an approximate Bayesian inference scheme and assumes that action control consists in producing proprioceptive predictions and successively fulfilling them by acting, rather than specifying motor commands (as is more commonly assumed in computational neuroscience and optimal control theory). In turn, proprioceptive (and other) predictions stem from priors encoded at high hierarchical levels, which thus essentially play the role of goal representations rather than the perceptual hypotheses of predictive coding.

PP is widely recognized in the two aforementioned domains — perception (e.g., *predictive coding*) and action (e.g., *active inference*). However, the PP framework is also increasingly used to explain a wide variety of cognitive phenomena of varying complexity, which go beyond action-perception loops and target abilities that have been traditionally considered to be the province of "higher cognition" (including for example planning, mindreading, foresight and cognitive control) as well as other domains including interoception, awareness and consciousness (Clark 2015; Clark 2016; Donnarumma et al. 2017; Hohwy 2013; Friston et al. 2013; Friston and Frith 2015a; Friston and Frith 2015b; Friston et al. 2016a; Pezzulo and Rigoli 2011, Seth 2013; Stoianov et al. 2016). This is appealing as, in principle, one can use the language of PP across multiple domains of cognition and even across different disciplines. However, a gap remains between the domains of (relatively simpler) action-perception loops and (relatively more complex) higher cognitive abilities. The former have been characterized in formal and quantitative terms using PP, whereas explanations of the latter tend to appeal to the same PP concepts but often lack a comprehensive quantitative and computational characterization. Thus, it remains to be seen if PP really "scales up" to higher cognition domains.

A second, related question is how exactly we should construct a PP theory of higher cognition. The mere fact that one can apply the principles of PP to action-perception loops and higher cognitive abilities leaves open the question of whether and how these domains are interconnected. A first logical possibility is that both action-perception loops and higher cognitive abilities comply with PP principles at an abstract level but use distinct sets of (neuro-computational) mechanisms, with higher cognition therefore being independent from action-perception loops. This "modular" perspective (or "a theory of two brains") is compatible with more traditional cognitive theories that segregate perception, action and cognition (and their neuronal underpinnings). For example, one may assume that children possess innate modules for language or "intuitive theories" of physics and psychology, and although these abilities may be described using PP principles, these are fully distinct from the PP mechanisms involved in action-perception loops[1]. A second logical possibility is that higher cognitive abilities are elaborations of action-perception loops which have never become (fully) segregated from them, and hence higher cognition remains functionally dependent on action-perception loops both during development and (at least in some cases) afterwards. This second, more "gradualist" perspective is compatible with various (stronger or weaker) forms of embodied or enactive cognition. Within this view, the existence of "cognitive mediators" — or sets of mechanisms that are shared across

1 Modularism need not necessarily postulate innatism. One possibility is that the (evolutionarily more primitive) mechanisms supporting action perception loops scaffolded (the evolutionarily more recent) higher cognitive abilities during evolution and/or development, with the latter later becoming segregated and functionally independent from the former. This "scaffolding" view implies PP mechanisms in the construction of abstract (and possibly amodal) cognitive domains, which at some point became independent from their scaffold and hence became functionally modular. For example, one may assume that language acquisition initially leveraged (e.g., speech) action-perception loops but this process resulted in amodal representations that fully support language processing. This view would be partially in agreement with some versions of embodied theories but not with others that reject amodal symbols (Barsalou 1999).

action-perception cycles and higher cognitive abilities — has often been postulated. One example I will discuss below is the idea that internal forward modeling is used on-line for action prediction and off-line for action simulation[2].

In principle, PP can be used to construct both modularist and gradualist theories. However, PP, and in particular active inference, has often been conceptualized in embodied or enactive terms that invite a gradualist view. There is however a problem that any gradualist PP theory has to face.

## 1.1 The Problem of "Detachment"

Traditional cognitive theories have been attacked for their inability to deal with the "symbol grounding" problem, that is, how abstract knowledge and internally manipulated symbols acquire their semantics, and how abstract cognitive operations link to the action-perception loops that realize them (Harnad 1990). Embodied and enactivist theories of cognition are better placed to solve the grounding problem because semantics can be directly grounded in the predictive mechanisms underlying action-perception loops. Ironically, however, these theories face the opposite problem: detachment. Because they assume functional and/or causal relations between action-perception loops and higher cognitive abilities, embodied and enactivist accounts need to explain 1) how the latter originated from the former during ontogenesis and/or phylogenesis; 2) whether and how the latter become functionally autonomous (or "detached") from the former, as exemplified by the fact that one can imagine an action without executing it; and 3) what detachment implies at the mechanistic level, i.e., whether imagining (or observing) an action engages action-perception loops covertly, recruits other mechanisms, or engages a combination of the two. All these problems are widely debated in cognitive and computational neuroscience, psychology and philosophy. How can PP help to shed light on these questions?

In this article I will discuss how PP can help us understand how living organisms could develop higher cognitive abilities from the mechanisms supporting adaptive action control. Central to this proposal are two interconnected ideas: 1) generative models that support PP in action-perception loops can progressively *internalize* aspects of agent-environment interactions; and 2) these generative models can be used in a "dual mode", with one stimulus-tied mode supporting overt action control, and another internally-generated or spontaneous mode supporting covert and detached forms of cognition.

The internalization process plausibly operates at an evolutionary timescale as it requires building sophisticated internal models, but it can sometimes also operate (or be completed) during development. Either way, the result of the internalization process is a form of cognition that that retains embodied and even enactivist aspects within the usual inferential scheme that PP uses to explain action and perception — and it is in this sense that one can trace back the roots of cognition in PP. From this perspective, the distinction between overt and covert processes — using the same generative model — is sufficient to explain the differences between simpler forms of action-perception loops and some forms of higher cognition.

Similar ideas regarding the internalization and reuse of predictive mechanisms have been advanced multiple times in cognitive science, at least since Piaget's proposals on the construction of intelligence from sensorimotor experience (Piaget and Cook 1952); but also in many more recent variants (Clark and Grush 1999; Cotterill 1998; Hesslow 2002; Grush 2004) and especially under the theoretical umbrella of action-oriented representation (AOR). In the rest of this article, I will firstly summarize some ideas on the reuse of predictive mechanisms across motor control and higher cognition from the perspective of AOR, pointing out some limitations of these proposals. Then, I will provide

2 For brevity, I cannot focus on the many other versions of this idea. For example, a "radical" version of embodied and enactivist approaches is that there is no real distinction between action-perception loops and higher cognition because the latter reduces to the former, i.e., higher cognition is fully supported by action-perception loops that comply to PP principles, or is possibly off-loaded to the external environment when it is necessary to implement forms of "symbolic processing".

three concrete examples of how generative models may support the internalization and reuse of PP dynamics across action-perception loops and detached cognitive abilities. The first example discusses rodent goal-directed navigation and highlights the importance of *internally generated sequences* of hippocampal neurons (place cells) that support both online navigation and the off-line "replay" of experience, along with the possibility of understanding this latter phenomenon as the internalization of a generative model. The second example discusses the formal similarities in PP between epistemic behaviour or information foraging in the external world (e.g., overt exploration) and the internal milieu (e.g., a mental action that lowers the uncertainty of a belief state), discussing how the latter may be an internalization of the former. The third example demonstrates how the internalization of homeostatic mechanisms (that, e.g., satisfy hunger) may lead to the formation of cognitive goals (e.g., buying food) at a higher level of a PP hierarchy, and how the latter may support detached cognition — for example, buying food even when one is not currently hungry. These three examples will clarify that the PP perspective does not simply recapitulate previous proposals but brings new and significant insights, extending the notion of internalization over and above the off-line engagement of internal forward models implied in action performance (a key tenet of AOR). Finally, I will briefly discuss the implications of this proposal, in particular in relation to embodied cognition and enactivist theories.

## 2 The Reuse of Predictive Dynamics for Higher Cognition in the Action Oriented Representation (AOR) Framework

There have been many attempts to discuss the origins of cognition as elaborations of a predictive control system, in particular appealing to the idea of *action-oriented representation* (AOR) or — often with a similar meaning — of "motor cognition" (Clark and Grush 1999; Cruse and Schilling 2015; Grush 2004; Jeannerod 2006; Pezzulo et al. 2011). AOR theories propose that the motor prediction and control architecture of our early ancestors was gradually improved to afford higher cognitive functions such as cognitive control, executive function, imagery, planning and declarative knowledge — and in parallel, joint action and communication in the social domain — but that these higher cognitive abilities retain important "signatures" of their situated origins, thus making even higher cognition embodied to some extent. A core mechanism for extending primitive architectures to more complex, higher cognitive domains is the reuse of motor predictions in an off-line mode, to support (for example) "what if" simulations in decision-making or the covert simulation of another's actions to understand her intentions. The basic idea is that, while engaged in an action-perception loop, agents also run another loop in parallel — a predictive loop (using a "forward model") to aid action control (e.g., to compensate for delays) which mimics an action-perception loop. However, under certain circumstances, such as when external inputs and external outputs are inhibited, the forward model can also operate in isolation from the action-perception loop. It is in such cases that agents perform covert (cognitive) operations such as action simulation or imagination. Unlike enactivist theories (Gallagher 2005; Varela et al. 1992), AOR theories emphasize the importance of internal models in supporting covert cognitive operations while the agent is disengaged from online interactions with the environment (including other agents). In summary, AOR theories constrain the space of cognitive operations to those that can effectively use forward models that were originally developed for online interaction. For this, according to AOR, higher cognition *retains essential features of* online interactions (i.e., forward models) although *it does not consist in* online interaction.

These and other proposals within AOR (or related frameworks) have highlighted the importance of prediction in the development of higher cognition from sensorimotor control. However, several aspects remained underspecified. It is unclear whether internalization exclusively regards forward models supporting action control, or whether it is a broader phenomenon. It is also unclear which aspects of action-perception loops can be internalized. Furthermore, AOR theories have been constructed on top of a process model of sensorimotor action that stems from optimal control theory (or its variants),

and it is unclear whether this is the right foundation for understanding cognitive operations. This question is pressing as optimal control theory does not easily include some aspects of active inference that are appealing from embodied or enactivist perspectives. These include *epistemic* aspects of behaviour (e.g., epistemic foraging for information (Pirolli and Card 1999) or hypothesis testing) which may be important for explaining a range of actions (including mental actions) that change an agent's informational or belief state as opposed to a state of the external world (Friston et al. 2015). Finally, in AOR, the relations between action (sensorimotor) control and the adaptive processes of homeostatic regulation (and associated sensorimotor loops) have rarely been investigated, but they may be important for linking actions and motivations and for constructing notions of cognitive goals that go beyond the execution of simple responses (Pezzulo et al. 2015; Pezzulo and Cisek 2016).

Can PP help understand the "roots of cognition"? Is the framework of active inference generalizable to higher cognitive abilities, and how can the relations of these cognitive abilities to action-perception loops be conceptualized? Is the notion of internal generative model useful for understanding how animals may "detach" from the here-and-now and engage in sophisticated forms of (retrospective or prospective) cognition? Below I will address these and other questions by discussing three examples of how generative models may support the internalization and reuse of PP dynamics across action-perception loops and detached cognitive abilities. Each of these examples is supported by a convergence of empirical studies and modeling studies using PP. They are: 1) goal-directed navigation and the role of hippocampal internally generated sequences (IGSs) within it; 2) information foraging or epistemic actions in the external world and in mental space; and 3) the detachment of goal states from homeostatic drives.

## 3 Internally Generated Sequences (IGSs) in Goal-Directed Navigation

The first example concerns the role of hippocampal dynamics in rodent goal-directed navigation. This example is relevant because the rodent hippocampus can process sequences of neuronal activity in two modes: a *stimulus-tied* mode while the animal is actually foraging in the environment, and an *internally-generated* (or spontaneous) mode in the partial or even total absence of external stimuli (e.g., while the animal sleeps).

The stimulus-tied mode of activity is evident when a rodent is engaged in a navigation task (e.g., when it actively explores its environment). During navigation, the animal's spatial position can be decoded by considering the so-called "place cells" in the hippocampus, which fire preferentially in specific portions of the environment (i.e., have localized place fields). Place cells are sequentially activated as the animal visits successive spatial positions corresponding to the cells' place fields. Therefore, at the population level, place cells form sequences that code for the animal's current spatio-temporal trajectory. At the behavioural timescale of rodent navigation, sequences occur in the presence of external cues or landmarks (O'Keefe and Dostrovsky 1971).

However, sequential neuronal activity can also arise in the hippocampus due to a distinct, internally-generated mode that is self-organized in the sense that it operates in the (complete or partial) absence of changing cues or feedback. These *internally generated sequences* (IGSs) of place cells correspond to temporally compressed representations of particular spatio-temporal trajectories that the animal has taken (recently or remotely), or might take (Diba and Buzsáki 2007; Foster and Wilson 2006; Pezzulo et al. 2014). Recent evidence suggests that IGSs play pivotal roles across a variety of cognitive tasks such as memory function (e.g., consolidation) and future-oriented cognition (e.g., route planning) (Pfeiffer and Foster 2013).

There are at least two important forms of IGSs. The first form of IGSs is the "replay" of spatial trajectories during sleep or when the animal is in the delay period of a memory task. Replay implies that the same sequence of neurons that coded for spatial locations during the actual rodent navigation (place cells) can be reactivated endogenously in the absence of triggering stimuli, in a time-compressed way:

within Sharp Wave Ripple (SWR) complexes (sub-second bursts of high frequency oscillation of up to 220Hz, see Buzsáki 2006). SWR sequences can proceed in both a forward and a backward direction, the latter more prominently after the animal collects a reward (Ambrose et al. 2016). Replays were initially linked to memory consolidation, following the influential hypothesis that the hippocampus may be specialized for the fast learning of episodic memories and may replay experiences off-line to train and consolidate cortical semantic memories (McClelland et al. 1995). More recent findings support the hypothesis that replays are also involved in prospective forms of cognition. For example, when animals rest between goal-directed spatial navigation episodes, replays are preferentially directed toward known goal sites and are predictive of future choices, suggesting a role in planning (Pfeiffer and Foster 2013). Furthermore, replays are not limited to the verbatim recollection of spatio-temporal trajectories that the animal has experienced, but can also generalize to novel trajectories or novel combinations of already experienced trajectories (Gupta et al. 2010), as well as to unexplored spaces in which reward delivery has been observed (Olafsdottir et al. 2015) or novel environments before they are visited, i.e., preplay (Dragoi and Tonegawa 2011).

The second form of IGSs is "theta sequences" — time-compressed trajectories that can be decoded in the hippocampal theta rhythm of rodents engaged in behavioural tasks (Foster and Wilson 2007). Within each theta cycle (7-12Hz), short sequences of place cells (four to six on average) fire with very precise temporal dynamics: each cell fires at a specific phase of the theta rhythm, which changes cycle after cycle (i.e., phase precession) while preserving the sequential order at the population level and the forward direction. Theta sequences are formed very rapidly (Feng et al. 2015) and often act as a "moving window", coding for a (forward) sequence of spatial positions loosely centred on the moving animal. The fact that theta sequences (often) include place cells that correspond to (have their "true" place field in) a future position of the animal's trajectory has motivated the influential proposal that theta sequences afford prospective coding and the prediction of upcoming locations (Lisman and Redish 2009). Notably, during difficult decisions, theta sequences can support the prospective coding of behavioural plans (e.g., trajectories that lead to preferred goal sites, see Wikenheiser and Redish 2015) and choice alternatives (e.g., branches of a T-maze, see Johnson and Redish 2007), possibly implementing a serial deliberation between them. This latter example refers to the vicarious trial and error (VTE) behaviour of rodents at decision points in T-mazes: in early trials before they accumulate sufficient knowledge about the reward location, rodents stop and repeatedly look to the left and right as if they are deliberating between the alternatives (Tolman 1938). During VTE behavior, hippocampal theta sequences "sweep forward" serially in the two branches of the maze (while the animal remains at the decision point). This suggests that the animal is performing a "search through mental information space" (Redish 2016).

In summary, sequential neuronal activity in the hippocampus is observed both at a behavioral timescale (while the animal visits successive locations during navigation and receives external stimuli), and at faster timescales (when theta and SWR sequences run in an internally-generated mode). To explain this finding, it has been proposed that support for a broader range of detached cognitive functions stems from the *internalization* of the stimulus-tied hippocampal sequences (and associated phenomena such as theta rhythms) that initially supported overt spatial navigation. Thus, after internalization, hippocampal sequences have a "dual use" and can operate in both a stimulus-tied mode and an internally-generated mode, the latter possibly supporting a wide range of cognitive operations (Buzsáki et al. 2014; Pezzulo et al. 2014). The possible functions of IGS (internally generated sequence) are various and still under investigation, and include memory consolidation (e.g., forming declarative memories, training an internal model), prediction and planning (e.g., preparing a route to a goal location), and the covert "what if" evaluation of possible action sequences (possibly in combination with other brain structures such as the ventral striatum).

The "dual use" may be conceptualized in terms of internal generative models for PP, which act as "sequence generators" for sequences of spatial locations (or more generally for sequences of events

which may not be navigational) and can form different functional networks with various brain areas (e.g., the enthorhynal cortex, the prefrontal cortex and the ventral striatum), depending on task demands, see Pezzulo et al. 2017). These internal models are learned while an animal navigates an environment (though they may be preconfigured to some extent, see Section 6). During navigation, the models are engaged by external stimuli (conveyed to the hippocampus mainly through the enthorhinal cortex) and can support the estimation of the animal's spatial position and produce short-range predictions. However, learning the internal models amounts to partially or fully internalizing the agent-environment dynamics, such that the same models can also be spontaneously reactivated (by tapping the self-sustaining internal dynamics of the model) in the partial or almost total absence of external stimuli (e.g., during sleep). This "tapping" can be either intentional, as in the below example of epistemic actions (see Section 4), or non-intentional, as in the case of replay of experience (e.g., for rodents, spatial trajectories and other events) during sleep.

How can this internally-generated mode be useful? An authoritative view is that the replay of spatial trajectories in rodents is useful for aggregating a series of *episodic* memories (temporarily stored in the hippocampus) into a *semantic* internal model in the cortex. This is supported by recent machine learning advancements which suggest that off-line experience replay significantly improves learning (e.g., by removing undesired correlations, Kumaran et al. 2016). There may be additional benefits if one thinks about the hippocampus in terms of an internal model rather than merely as a "storage" of episodic memories. Theoretical considerations suggest that when an internal model is spontaneously engaged in the absence of external stimuli, it can produce "unbiased" resamples of its content (in the case of IGSs, one might say that it would produce samples of trajectories based on the model's prior probability distribution (Buesing et al. 2011)). However, in practice, there will often be some external input or bias to this process. For example, the representation of a desired goal location such as the "home" location of the animal (possibly stemming from the prefrontal cortex) can influence this "resampling" process and bias the resampled sequences towards the goal location, possibly supporting planning function. This can be explained using the mechanisms of active inference (or the related framework of *planning-as-probabilistic-inference* (Botvinick and Toussaint 2012)), if one considers that goal-representation acts as a sort of constraint (or allegorically, a sort of attractor state) that funnels the resampling process. The same process can be used to repeatedly resample past experience for memory consolidation or for cognitive map formation, with the possibility to "bias" the sampling in a way that (for example) over-represents rewarded experiences (Kumaran et al. 2016). Importantly, the generative processes described here do not consist in the verbatim recollection of past episodes (as suggested by the term "experience replay") but have constructive elements. They therefore permit (for example) recombination or interpolation from past experience (Gupta et al. 2010).

This example has illustrated that (hippocampal) neuronal dynamics can operate in a dual mode: one stimulus-tied and one internally-generated. This finding suggests the presence of an internal model that internalized agent-environment dynamics and is able to reproduce them spontaneously, in the absence of stimulus. Although this hypothesis remains to be fully empirically tested, it is exemplificative of a possible pathway from action-perception to detached cognition via internal modeling of the kind used in PP.

Although the evidence I reviewed comes from animal studies and touches only a limited set of detached operations — spatial memory and planning — the scope of IGS and related mechanisms may extend well beyond this. It has been suggested that mechanisms analogous to IGS may support more advanced human abilities including imagination, prospection and "mental time travel" to the past and the future, since these "detached" activities also recruit shared brain structures including the hippocampus and the medial temporal lobe (Buckner and Carroll 2007; Schacter and Addis 2007; Suddendorf 2006). One theoretical proposal bridging these seemingly disconnected fields is that "mechanisms of memory and planning have evolved from mechanisms of navigation in the physical world" and "the neuronal algorithms underlying navigation in real and mental space are fundamen-

tally the same" (Buzsáki and Moser 2013, p. 130). This would suggest that navigation and reasoning in arbitrary domains ("mental spaces") may be based on the same mechanisms that support overt spatial navigation.

## 4 Epistemic Actions Can Be Executed both Externally and Internally

The second example concerns *epistemic actions*, which can be executed both externally (e.g., through overt exploration of the environment) and internally (e.g., by using a generative model to simulate the outcome of a series of actions and "gather evidence" in favour of a select few before a choice is made). Recent theoretical and empirical studies have shown formal similarities between these two forms of "information foraging" (Hills et al. 2015; Pezzulo et al. 2013), both of which may be invoked in the face of exploration-exploitation dilemmas, or when collecting information (and thereby reducing uncertainty) prior to a choice is more cost-effective than taking action based on current knowledge.

Both overt exploration (exemplified by searching for external cues before making a choice (Friston et al. 2015) and covert mental exploration (exemplified by rodent *vicarious trial and error* behaviour at decision points (Pezzulo et al. 2016a) have been recently modelled using PP. Importantly, both appeal to the same concept of *epistemic value*, which is an integral part of the active inference scheme. In this scheme, an agent's plans must balance extrinsic value (e.g., reaching goals) and epistemic value (e.g., reducing uncertainty about the goal location); the latter can gain prominence over the former in circumstances where uncertainty is too high.

It is tempting to speculate, given the formal analogy between overt and covert forms of exploration and epistemic action, that some covert mental operations result from the internalization of mechanisms that balance overt exploration (i.e., exploring novel action possibilities) and greedy exploitation (i.e., selecting the most rewarding action found thus far) in conditions of uncertainty or risk. One might therefore use the internal model to consider and evaluate hypotheses in one's mind (or to "collect more evidence" for and against each hypothesis) until one is either confident about one's decision or decides that it is not worth investing further cognitive effort in that task. This captures the trade-off between exploring novel action possibilities and selecting the most rewarding action found thus far. Cognitive neuroscience is starting to scrutinize some of the brain mechanisms underlying exploration-exploitation and cost-benefits computations, including the balance between deliberative and habitual forms behavior (Daw et al. 2005; Redish 2016; Pezzulo et al. 2013) and the trade-offs between the costs of increasing attention demands (or exerting cognitive control over a task) versus the benefits in terms of increased reward (Shenhav et al. 2013). These trade-offs can be conceptualized using hierarchical PP architectures, in which (for example) deliberative mechanisms can supersede and contextualize habitual forms of behavior. Under PP, habitual forms of behavior would be selected when engaging the full deliberative system is not cost-effective, for example, when the animal is sufficiently confident that the environment has not changed, and so repeating a previously successful action is likely to result in a higher pay-off than exploring new opportunities (Pezzulo et al. 2015).

I have thus far focused on an intentional kind of epistemic action that consists in "tapping" or "interrogating" a generative model in order to probe hypotheses or collect evidence. However, this may be one instance of a more general cognitive mechanism that permits one to exert control over one's own mental processes (as opposed to control over the external world); in other words, a mechanism that sees "thinking as the control of imagination" (Pezzulo and Castelfranchi 2009). This perspective implies that the concepts of "actions" and "skills" are extended beyond those that require the expression of overt behaviour to also include mental operations that have no immediate external referent. In a similar vein, Metzinger 2017 discusses how mental operations can have *epistemic goal-states* (e.g., "Knowing what the sum of 2 + 3 is"). One can also imagine other kinds of mental operations that are controlled towards some desired end-state. For example, an interior designer can move or change furniture pieces in her mind until she reaches a configuration that fits the style of the house; or an

animal can mentally resolve a competition between affordances, or plan to create new affordances, before acting (Pezzulo and Cisek 2016). Here, again, PP permits the identification of a crucial feature of these mental activities: the fact that they are actively controlled *towards* a desired goal state — where achieving the goal state has epistemic value.

As briefly mentioned above, the functional organization of action in PP (specifically, in active inference), as opposed to other schemes such as optimal control theory, revolves around achieving goal states using prediction and error correction mechanisms. In active inference, goals control action and perception engenders a cascade of predictions that are hierarchically decomposed down to set points for peripheral reflex arcs, which steer bodily movements. The same scheme can be adopted in a more internalized way without arc reflexes, if one allows an active inference agent to express goal states that concern his own mental states or "beliefs" (where the term "belief" is used in the technical sense of probability theory, not in the sense of classical propositional attitudes, and may denote for example a Gaussian probability distribution, defined by the two parameters of *expectation* or mean and *precision* or inverse variance). One example of an epistemic internal goal state is "having highly precise beliefs about the value of choice offers" when one needs to gather new evidence before making a decision about an investment. Another example is "having highly precise beliefs about the best placement of furniture pieces in this house" if one wants to design a fancy house layout. Yet another example is "having highly precise beliefs about the best way home" during route planning. In all these examples, an agent can execute a mental action to change his or her belief state[3], and to make some of his or her beliefs very precise before making a choice.

It is worth noting that, in active inference, the precision of all the relevant beliefs (e.g., about the agent's current and goal locations) is always optimized before a choice. This optimization is considered to be a standard aspect of active inference (or free energy minimization), not a form of meta- or cognitive control. However, there may exist (mental) operations that override or finesse the default optimization mechanisms of active inference, which would provide "mental actions" a truly causal role in the architecture of PP (see Metzinger 2017 for a comprehensive discussion). These mental actions may be cast within a Bayesian learning or active inference scheme, too. For example, one can use priors about the precision that a belief or a set of beliefs needs to have before a choice, or one can monitor precision levels, until one has a sufficient "sense of confidence" (Meyniel et al. 2015), i.e., a sufficiently high likelihood that one's inferences are correct. In other words, the selection of a mental action — and the solution of "decide now vs. collect more evidence" (or optimal stopping) problems — may rest on the precision-modulation of internal epistemic states (e.g., raising priors on expected precision or confidence before a choice). This is analogous to the way raising the precision of an internal belief makes it a strong goal representation that generates a cascade of predictions, which in turn enslave overt action.

From this perspective, the achievement of epistemic goal states (and the resolution of epistemic uncertainty) may be seen as a form of cognitive control over one's own mental activity by using monitoring, error correction and precision modulation mechanisms that are analogous to overt action control (Pezzulo 2012) in order to control epistemic behavior and attain sufficient confidence in one's choices. Some examples are: continuing to mentally compare the pros and cons of various invest-

3 While both mental action and perceptual inference change the agent's belief state, there are significant differences between the two. Perceptual inference is implemented under a predictive coding scheme; it uses prediction errors to changes the agent's belief state as a function of new evidence and prediction error but does not include action selection or the active search for new evidence. Conversely, a mental action can be conceptualized as an epistemic action under the active inference scheme, which generalizes predictive coding to also include action selection and planning. Like other (overt) epistemic actions, mental actions result from the deliberate choice to search for new information (or reconsider old information) for epistemic purposes, e.g., to intentionally reduce uncertainty before a difficult choice. For example, if a conference has two parallel symposia and one is uncertain about which one to attend, he can execute an epistemic action externally (e.g., spend some time reading the talk abstracts) or internally, as a mental action (e.g., explicitly recall past examples of talks by the same speakers). Epistemic (or mental) actions are selected as part of a plan whose benefits (e.g., reading all task abstracts until one is very confident about the best symposium) and costs (e.g., the cognitive costs of abstract reading) are considered and compared with those of alternative action plans (e.g., drinking another coffee and then going directly to a random symposium). The selection of a mental action thus complies with the same formal principles that regulate the balance of intrinsic (epistemic) and extrinsic (economic) value in active Inference (Friston et al. 2015).

ments (or reading business webpages) until one is confident enough, continuing to imagine moving furniture (or actually moving it) until one is happy with final configuration, or striving to remember past travels (or consulting a GPS navigator) until one is certain about a travel plan. The computational efficiency and empirical validity of these or alternative schemes, and their relations to meta-cognition and cognitive control, remain to be assessed. Furthermore, it remains to be studied whether mental action selection obeys cost-benefit considerations, permitting one to trade off the benefits of extra information and extra confidence against the cognitive and temporal costs of achieving these epistemic goals (Shenhav et al. 2013).

## 5 From Homeostatic Drives to More Abstract Goal States

My third example concerns the detachment of goal states from homeostatic drives. In active inference, one can describe adaptive control loops by starting from cybernetic error-correction mechanisms (Butz 2016; Pezzulo et al. 2015; Seth 2013). To illustrate the concept, one can start with a homeostatic drive — such as a felt need for glucose — that produces (interoceptive) prediction errors. These errors, in turn, engender autonomic responses but also a sophisticated (crossmodal) generative model that produces a cascade of (exteroceptive and proprioceptive) prediction errors. The latter engage an action pattern — such as locating and consuming an apple — that suppresses all the prediction errors, including the initial interoceptive prediction error (by restoring homeostasis), thus terminating the process.

However, not all adaptive actions are initiated (and controlled) by current needs and interoceptive prediction errors. The fact that one can buy food even when one is not hungry exemplifies the human ability to set and achieve goals in open-ended ways. In other words, there is often a strong *functional* dependence between homeostatic imperatives (e.g., to be satiated) and goal states that drive adaptive action (e.g., finding and then consuming food), but the *causal* (or proximal) coupling between the two can be sometimes loosened. In other words, an interoceptive prediction error (signaling e.g., low glucose levels) is not always required to initiate active inference and control loops for food consumption.

To understand how this may be possible, one needs to understand the aforementioned cybernetic scheme in more detail, in particular, its anticipatory aspects. In PP, adaptive action is realized by a generative model that encodes contingencies across interoceptive, exteroceptive and proprioceptive modalities, e.g., between glucose levels, the visual appearance of apples, and the actions required to secure them. The predictive capabilities of the internal model permit going beyond feedback-based error correction, to steering a series of anticipatory regulatory (or *allostatic* (Sterling 2012)) loops. For example, one can stop eating an apple predictively (i.e., by using the internal model to predict that eating a certain amount will restore glucose levels) rather than reactively (i.e., only after receiving a signal that the glucose level is actually restored). This is more adaptive given that generating the latter signal may take too much time. Moreover, one can use predictions rather than just feedback to select and regulate action. For example, one can decide on an action in anticipation of a predictable need (be satiated) rather than waiting for an interoceptive error signal (for hunger). Another example of (implicitly) anticipatory process during regulatory eating loops is salivation, which prepares resources to digest a to-be-eaten food (Pavlov and Thompson 1902). The theory of *allostasis* (Sterling 2012) encompasses many more examples of anticipatory regulatory mechanisms, which involve (for example) hormonal processes that mobilize resources in anticipation of a need, and which up- or down-regulate the whole system rather than using fixed set points as would be suggested by the idea of homeostasis. All these examples illustrate that even the (relatively) simple regulation of drive states can be largely anticipatory rather than just reactive. The neuronal PP architecture supporting the highly integrative functions required for allostasis is necessarily hierarchical and includes important hubs that combine interoceptive, exteroceptive and proprioceptive modalities (e.g., the insula, see Craig 2015)

During learning and development, the internal model can increase its scope and internalize drive-based regulatory loops to generate (for example) goal-representations and plans for "eating". It can then initiate a plan for "searching for a restaurant" or "buying food" in an internally-generated mode (i.e., based on goals) rather than in a stimulus-driven mode (i.e., only after feeling hunger). This can be done using a hierarchical PP model that progressively learns regularities at increasingly deeper levels and longer temporal timescales — such as the relations between the act of ordering food in a restaurant and the integrity of the internal milieu, as measured by low prediction error of interoceptive signals (Pezzulo et al. 2015). These internal models permit an organism to anticipate needs rather than merely reacting to them, and to prepare to satisfy a drive (e.g., hunger) before there is an interoceptive error signal. From this perspective, the role of a higher-level cognitive goal like "buying food" would be to produce a sort of anticipatory error-signal, which triggers error-correction actions (for food consumption) before a lower-level drive system produces an interoceptive error-signal (e.g., loss of glucose), which would be more dangerous. The proximal mechanisms for producing goal-related error signals and for monitoring goal achievement error signals may be borrowed from more primitive mechanisms that monitor reward achievement (Montague 2006).

This example illustrates that a goal-based mechanism for action selection provides some detachment from immediate needs and homeostatic drives; in other words, goal-directed (intentional) action can distally relate to basic drives but also acquire autonomy from them. The act of finding a restaurant before one is hungry retains the full intentional and adaptive character of eating an apple when one is hungry; while the latter is driven by interoceptive stimuli, the former is internally-generated (goal-driven) but still adaptive, because the model has internalized a basic homeostatic loop (and the subsequent corrective actions). The power of this hierarchical PP scheme rests on the fact that it allows animals to control and produce effects in the external world and invent "cognitive goals" in open-ended ways, which go well beyond the satisfaction of the homeostatic drives that (often) originated them.

The question of how much cognitive goals can "diverge" from simpler physiological imperatives is still open. In principle, the fact that goal states are learned by internalizing (and predicting) allostatic loops should prevent a radical divergence between the two. Furthermore, in the PP hierarchy envisaged here, the "higher" layers that encode more cognitive goals (like finding a restaurant) remain to some extent linked to the "lower" layers that implement more basic allostatic loops. Prediction errors need to be minimized at all levels, and so even if the "cognitive goal" of eating at a good restaurant has been achieved, one can change restaurant if (after a while) the simpler drive of "being satiated" is not achieved — at least if one has enough time and money — which would also change the "model" of the restaurant. Finally, it is not necessary that higher layers supersede all the operations performed by lower layers. For example, in evaluating how good a food is, one can rely on (relatively higher) cognitive representations of the quality of a restaurant — for example, whether it was positively reviewed in a gourmet magazine — but also engage a (relatively lower) interoceptive simulation that provides anticipated feelings of taste.

However, there may be cases where cognitive goals truly diverge from physiological imperatives. My examples regarded very simple cases of goal states (like finding a restaurant) for which one can reconstruct a plausible causal history back to physiological states (hunger). However, even these (apparently) simple goal states have aspects that originate from cultural dynamics and may not be easily reducible to homeostatic imperatives — and sometimes may run against them. This is even more evident in sophisticated goals such as pursuing an ascetic ideal. Although one may link social and cultural practices to the usual imperatives of survival and reproduction, the ways proximal mechanisms (goal achievement) link these domains are not always easy to reconstruct. Finally, it is important to consider that there is a strong habitual component in human behavior (e.g., buy food at my usual supermarket, or go to a restaurant every Friday); while habits may originate from the routinization of goal-directed control (Pezzulo et al. 2015; Friston et al. 2016b), they do not retain its flexibility (e.g., they can be insensitive to changes in interoceptive state) and thus may become maladaptive. These ex-

amples illustrate that in humans, the relations between sophisticated goals and simpler physiological imperatives may be multifarious — but at least hierarchical PP modeling offers some guiding principles for studying them.

## 6 Open Questions and Interrogatives

I started with the problem of 'scaling-up' action-oriented theories of cognition to account for 'higher' cognitive phenomena (such as imagery, navigation, and so on). I provided three examples of such 'higher' cognition (in spatial navigation, mental actions, and the creation and attainment of cognitive goals) and discussed them in terms of detached actions — where the detachment process rests on the construction of generative PP models, which permit the internalization of action-environment dynamics. My proposed view of detached cognition as internalized PP has several implications, but also raises numerous interrogatives, which I summarize below schematically:

- It follows from this framework that, although detached cognition is free from some of the constraints of action-perception cycles (e.g., the necessity to support real time action), it may preserve key situated and embodied aspects of overt action control, such as the "natural statistics" of control tasks (e.g., navigation or grasping). One example is the fact that hippocampal theta and SWR sequences run at a faster timescale (thus without the temporal constraints of action-perception cycles), yet their content (in terms of place cells and their relative temporal arrangement) can be the same or very similar. Visual and motor imagery studies similarly suggest important similarities at both functional and neuronal levels between perceptual (or motor) experiences and their covert counterparts (imagery), despite the fact that the latter are free from some constraints (e.g., one can imagine legendary animals "in the mind's eye" and in some cases one can imagine faster than one can see Jeannerod 2006; Pylyshyn 2001). Which constraints are shared and which are not (and why) across overt and covert cognitive operations remains to be fully assessed.
- In active inference, bodily processes are part of the inference and contribute to prediction error (or free energy) minimization. One example is active sensing: one can use hand or eye movements to infer (or reduce surprise about) the position and shape of an object (Friston et al. 2012a). Under certain conditions, other bodily movements may be part of covert cognitive operations, too. For example, eye movements upwards may contribute to our understanding of a story about a person who climbs a building (Spivey and Geng 2001). Alternatively, some aspects of bodily movements may be internalized, too, and become part of generative models that can run covertly in an internally-generated way; this may be the case, for example, in head direction cells in rodents (Buzsáki et al. 2014).
- From a neurophysiological viewpoint, an intriguing question is how the same neuronal population can operate in two modalities: stimulus-tied and internally-generated[4]. From the PP perspective, this question is also interesting as it may help shed light on how the brain implements internal generative models. One domain where this question can be addressed is hippocampal processing. In the hippocampus, learning is extremely rapid. This has led to the proposal that it may rest on preconfigured neuronal dynamics: pre-coded sequences of neuronal activities that may serve as "dynamical templates", permitting the learning of new experiences "in one shot"

4 Usually, "stimulus-tied" refers to a process that is initiated by a stimulus form the outside world (e.g., a visual stimulus). In the case of homeostatic regulation and interoceptive loops, the same idea maps to the availability of bodily sensations and feelings (e.g., feeling hungry) rather than truly "external" stimuli. However, the very idea of a process that is initiated by stimuli (either perceptual stimuli or bodily sensations and feelings) is too reactive for PP theories. Active inference is guided (proactively) by exteroceptive, proprioceptive and proprioceptive predictions that obey high-level beliefs and homeostatic imperatives, rather than being governed (reactively) by stimuli — except in special cases such as habits (Friston et al. 2016b). From this perspective, active inference has always "internally generated" components — especially at high hierarchical levels, which can embed self-sustained dynamics that continuously engender predictions, see e.g., (Friston et al. 2012c). The distinction between "stimulus-tied" and "internally generated" still makes sense in this context if one considers that in the latter case external stimuli are partially or entirely missing.

with successive refinement of what was learned on the basis of additional experience (Dragoi and Tonegawa 2011). A more comprehensive proposal is that (unlike most current machine learning approaches) this form of learning may consist of aligning internal spontaneous brain dynamics (and rhythms) to external environmental dynamics, and once this is done, brain rhythms spontaneously "replay" agent-environment interactions. This would mean that internal models do not become detachable only after learning but are able to run in a spontaneous mode from the beginning (Buzsáki 2006). The idea that it is possible to *model* agent-environment dynamics by *aligning* internal and external brain dynamics is also appealing from an enactivist viewpoint (see Section 7).

- The second example, on "epistemic action", suggests that we should expand the concept of an "action" beyond bodily movements that achieve an externally defined goal state, to also encompass (covert) mental actions that have no immediate external correlates but maintain action intentionality. This extended notion of action is fully compatible with PP. Indeed, in PP an action is directed towards a goal state, but the notion of a goal state does not have to be restricted to the exteroceptive or proprioceptive domains (as for physical actions such as grasping), but can encompass an informational state of the system (e.g., the precision of a belief state). This is most evident if one considers that, in PP, actions have both pragmatic (or extrinsic) and epistemic (or intrinsic) components, with the latter having the goal of changing informational state (e.g., searching for a cue). This notion of an epistemic action in PP adequately covers those "mental" actions whose goal is changing the agent's belief state (and not, technically speaking, the state of the external environment or the agent's physical internal environment, such as being hungry).
- One interesting implication is that mental actions that resolve uncertainty would have the same intentionality as external exploratory actions. In other words, information foraging is the same when performed in the mind or in the external world. A second interesting implication is that "mental actions" can be conceptualized as actions that change the cognitive (belief or goal) state of an agent that is involved in (for example) action planning or control, and need not to operate on a distinct (amodal) cognitive substrate. Thus, for example, it would not be necessary to postulate two distinct cognitive representation of state uncertainty, one that is controlled using "mental actions" and one that influences action selection.
- The third example, on drives and goals, provides an example of detachment that rests on the construction of a hierarchical generative model rather than a strict dual use of the same parts of the model, as in the case of hippocampal sequences. In such cases, the internalization of a homeostatic loop and the ensuing regulatory plans requires building additional (hierarchically higher) components of the model. While the fact that cognitive goals can be engaged even when there is no sensed interoceptive loop suggests that the hierarchically higher goals enjoy some degree of detachment, it is unclear whether and in which cases this detachment of the "higher part" from the "lower part" can be complete, especially since prediction error has to be minimized across the whole hierarchy, not just in a part of it.

## 7 Relations with other Approaches Including Embodied and Enactivist Views

This perspective has some similarities with, but also important differences to other proposals, which I briefly summarize below.

### 7.1 Action-Oriented (AOR) Framework

Compared to most proposals advanced within the AOR or motor cognition frameworks, here the emphasis is not just on the reuse of motor predictions outside motor control loops, but the engagement

of generative models in an internally-generated mode, which may be a broader phenomenon. In other words, the off-line reuse of the motor system's predictive abilities may be just one of the mechanisms permitting a biological organism to temporarily disengage from the perceptual-motor loop and engage in detached forms of cognition.

## 7.2 Cortical Recycling and Neural Reuse

The ideas of internalization and "dual use" have some relation to theories of "cortical recycling" (Dehaene and Cohen 2007) and of "neural reuse" (Anderson 2010), which focus on the exaptation or recycling of neuronal resources that then acquire novel functions. For example, a brain area adapted for perception might be exapted to also recognize letters. However, implicit in the idea of "dual use" is the assumption that covert cognitive abilities (e.g., planning) remain connected to the overt processes (e.g., spatial navigation) that scaffolded them, because they use a common generative model. In other words, these abilities are not just connected by their ontogenetic or phylogenetic history (e.g., recycling), but continue to share a generative model, which can operate in two dynamic modalities (stimulus-tied vs. internally-generated). The possibility of operating in two modalities is intrinsic to the notion of a generative model, in which the internally-generated mode corresponds to the generative process of "imagining" or "hallucinating" patterns such as images, faces, video frames, etc. (Hinton 2007).

## 7.3 Dual-Process Theories

The notion of "dual use" is not the same as dual-systems theories in cognitive science (such as the idea of two separate systems of thought, one reflexive and one deliberative (Kahneman 2011)). Nor does the distinction between stimulus-tied and internally-generated processes map to the distinction between habitual and goal-directed control in dual-process theories of reinforcement learning (Daw et al. 2005). Stimulus-tied modalities reflect a process occurring at the same timescale as the action-perception cycle. This process can fully incorporate external stimuli independently of whether goal-directed action planning or stimulus-response is implemented (the latter, in active inference, applies only in rare cases such as in the presence of habits). Rather, internally-generated refers to processes that are outside the action-perception cycle (e.g., the covert replay of spatial trajectories while the animal sleeps and is deprived of external sensations).

Dual-process theories in reinforcement learning assume that goal-directed action and habits depend on segregated neuronal and computational processes, and that they compete to control behaviour (although the possibility that they may also cooperate has been sometimes recognized). Within active inference, however, goal-directed actions and habits are better conceptualized within a hierarchical scheme in which the higher layers that implement goals can contextualize lower levels that implement less flexible responses (Pezzulo et al. 2015). The result is a continuum between goal-directed and habitual action that depends on the relative weight assigned to the different layers. Habits can arise in this scheme too, when the lower layers acquire sufficient precision to become essentially impermeable to the influences of the higher layers. As a further development of this view in the context of policy selection, one can consider that habitual policies can arise from the self-observation of goal-directed action planning when there is no (residual) ambiguity (Friston et al. 2016b). In this scheme, one would initially select policies in a goal-directed manner, and successively (when there is no ambiguity over time) develop a habitual policy: a "copy" of the most-often selected goal-directed policy, which can be selected in a stimulus-based manner rather than using deliberation and expected free energy

minimization. It is worth noting that in this scheme, habitual policies are not learned in parallel with goal-directed action (as assumed typically in dual theories) but only afterwards[5].

## 7.4 Perceptual Symbol System Theory

The PP-inspired view of detached cognition sketched here connects quite well with the most developed conceptual framework for embodied cognition: perceptual symbol system theory (PSS) (Barsalou 1999). In PSS, experiences are internalized to form embodied concepts (or "perceptual symbols"), whose re-enactment produces a "simulator" that steers "situated simulations". Here, one can consider that "perceptual symbols" link to specific (unimodal) elements of a generative model and "simulators" link to multiple interconnected elements that form a multimodal concept (e.g., the concept of a dog generates multimodal predictions regarding what a dog looks like, how the bark sounds and the anticipated softness of touching a dog). A "simulation" refers to the generative process of the generative model, which produces (or "hallucinates") observations that are compatible with a given simulator or a combination of multiple simulators, much like deep (generative) neural networks are used to generate exemplars in machine learning (Hinton 2007). In PPS, however, a simulation is always "situated": the (prior) information encoded in the simulator is combined with various contextual elements that are present at the moment a person instantiates a simulation. This implies that a person would produce different "situated simulations" of an airplane if he is flying or at home, if he is happy or worried, or if he is engaged in a memory task (e.g., recalling names of parts of an airplane) or imagining a future flight. This situatedness (or context-sensitivity) is a hallmark of human cognition and is currently beyond what current machine learning techniques can do; perhaps it would require embodying a PSS into an agent that dwells in realistic environments and has a rich set of personal experiences. Despite these limitations, some key ideas of PSS may be explained (or implemented) using the usual constructs of PP; a situated simulation might construct perception by generating and predicting exteroceptive observations (predictive coding), guide action by generating proprioceptive predictions (active inference), and scaffold emotional experience by generating and regulating interoceptive states (interoceptive inference or embodied predictive coding; Barrett and Simmons 2015; Pezzulo 2013; Seth 2013).

## 7.5 Enactivism and the Relations between Internal Modeling and Representation

Most of the ideas I discussed in this article would also lend themselves quite naturally to an enactivist perspective. This is consistent with previous observations that PP (and in particular active inference) has enactivist elements (Allen and Friston 2016; Friston et al. 2012b; Bruineberg et al. 2016). This seems prima facie surprising, given that active inference includes the (cybernetic) notion that adaptive control requires an *internal model* of the environment, and the idea of an internal model is closely related to the idea of internal representation, which is antithetic to enactivism. However, in active inference, internal modeling is instrumental to accurate goal-directed action control, over and above representation (which is not the case in all theories of PP and perceptual predictive coding). Priors play the dual role of hypotheses (in perceptual processing) and goals (in action control). However, in most practical cases, the latter goal-oriented role is more fundamental, because there are some prediction errors, such as those generated by homeostatic processes, which cannot be minimized by "changing one's mind" but require taking action (e.g., eating or drinking). Accordingly, the brain develops internal models and generates predictions to satisfy the agent's goals (or to maintain allostasis) rather than to maintain an accurate internal representation of the external environment per se. In other words, the

5 The rationale for this is that the full power of active inference is only necessary under risk (e.g., many-to-one relations between outcomes and internal states) and ambiguity (i.e., many-to-one relations between internal states and outcomes). When there is no ambiguity, then the external stimuli are usually sufficient to afford a successful state-action policy of the kind commonly used in fields like reinforcement learning. Therefore, developing a habitual policy would amount to a sort of "simplification" of the internal model, which is just another way to reduce free energy, see Friston et al. 2016b.

success criteria for internal models of agent-environment dynamics are accurate prediction and goal achievement, not accurate mirroring of an external reality. In most practical applications, a model can afford good prediction even if it is sketchy and does not capture the full complexity of environmental dynamics. This is evident if one looks at published studies and compares the agent's generative model with the "true" generative process (aka the "real" environmental dynamics). In summary, active inference can be conceived of as the synthesis of two ideas: that "the brain is for prediction" and "the brain is for action". The focus on the latter, action-based and embodied aspects of brain function (which is not mandatory in other PP approaches) relaxes representational aspects of internal modeling.

An even more nuanced view of the relations between internal modeling and representation emerges if one assumes that developing an internal model boils down to aligning (or synchronizing) pre-existent brain dynamics and rhythms to environmental dynamics, as discussed above in relation to hippocampal processing. This may be not entirely satisfactory from an enactivist viewpoint, though, as it still requires postulating that internal models are *within* the brain. Alternatively, one can consider that even if an internal model is required for control, the internal model is not *within* the brain; rather, *the brain-body-environment system as a whole* implements an internal model[6]. For example, a robot may produce efficient locomotion by aligning internal dynamics (e.g., rhythmic behaviour produced by central pattern generators for locomotion) and external dynamics (a treadmill moving at a certain speed), while also exploiting some aspects of its embodiment (e.g., the design of its legs which may afford correct posture) to simplify control (Pfeifer and Bongard 2006). This echoes the claim that "the system" that operates locomotion is not reducible to a brain controller, and hence one need not postulate that internal models (or representations) are *within* the brain. While this latter argument is credible for tasks that require on-line engagement with the external environment (including other agents), as in the walking robot example, it is less clear whether it is sufficient for implementing higher cognitive skills that may require the *detachment* of internal generative models from the rest of the system (e.g., from online environmental dynamics). For example, it is unclear how exactly the walking robot described above may form (or select among) locomotion plans. In other words, if internal models are formed by brain-body-environment systems, they may not be detachable from on-line interactions, or they may require additional elements to be detachable such as the internal emulation of environmental dynamics. It thus remains an open question whether treating the brain-body-environment system as an internal model would be sufficient to explain the kind of phenomena I have discussed here: for example, hippocampal internally generated sequences and their roles in memory and planning; mental actions; and detached goal processing.

## 8 Conclusions

The key constructs of PP (e.g., prediction and prediction error, generative model, and precision) are increasingly used to explain cognitive phenomena of various complexity, ranging from action-perception loops to interoception and emotion, decision-making, planning, and beyond. However, the mere application of the same principles to several domains leaves room for different interpretations. Embodied theories of cognition tend to assume interdependence between action-perception loops and higher cognitive domains, yet it is unclear how the latter may have originated from the former.

I have discussed three examples that illustrate a general principle: cognitive (or covert) mental activities may result from an internalization process which engages brain circuits and internal generative models originally used for overt behaviour (e.g., goal-directed spatial navigation, epistemic foraging in the external environment, or acquiring food to satisfy a currently felt hunger). Generative models implied in

6 Technically speaking, internal states of a dynamic system can be distinguished from external states by appealing to the statistical concept of the "Markov blanket" that separates them. In this scheme, internal states model and act on external state to preserve the system's integrity (Friston 2013). However, this formulation does not prescribe where the boundary between internal and external states should be located, and whether the "Markov blanket" separates (for example) the brain from the rest of the system, or the brain-body from the rest of the system, etc. One can also appeal to the idea that there are multiple, nested "Markov blankets"; see Allen and Friston 2016.

PP may permit us to internalize (or to use fancy words, *em-brain* or *cognitivize*) key aspects of agent-environment interactions, including interoceptive loops as in the case of allostasis. This would permit the use of internal models in a "dual mode", stimulus-tied vs. internally-generated (or spontaneous). The former mode is associated with (overt) action-perception cycles and the latter with (covert) cognitive processing which is detached from the here-and-now and can thus support, for example, future-oriented (prospective) and past-oriented (retrospective) forms of higher cognition. These examples illustrate a gradualist PP perspective in which higher cognitive abilities are distinct from action-perception loops because they run (covert or overt) PP processes, not because they are implemented in distinct modules.

Interestingly, the three examples illustrate that internalization and dual use may be implemented in various ways. In the spatial navigation example, internalization has a clear neurophysiological connotation: the same neuronal circuit (involving the hippocampus but also other brain areas) can operate in two modes, which correspond to two distinct dynamic regimes or brain rhythms (Buzsáki 2006; Buzsáki et al. 2014). In the epistemic action example, internalization refers to a functional principle — mental actions can achieve the same epistemic goals as external exploration actions — but it is currently unclear whether overt and covert forms of information foraging use shared neuronal circuits. Finally, in the example of drives and goals, internalization rests on the construction of a PP hierarchy (e.g., a hierarchy of drives and goals) that can be successively engaged in a stimulus-tied mode (search for food when hungry) and an internally-generated mode (search for food when not hungry but anticipate hunger), with the latter resting on cognitive goals that enjoy some detachment from drive states. This latter example suggests that some cognitive operations (e.g., the processing of abstract goals) can rest on hierarchical elaborations of action-perception (or interoceptive) loops, but also that the latter can be engaged when necessary during abstract goal processing (e.g., for anticipating the taste of a food).

The idea of internalization is not novel, but PP offers a conceptual framework that facilitates discussions and empirical validation, and generalizes and extends several distinct proposals within a mechanistic and biologically-grounded scheme. I focused on three examples that have been recently characterized in PP terms. However, I consider them to be illustrations of a phenomenon that may be much more general. For example, as discussed above, theories of AOR and motor cognition have provided other useful examples of the off-line reuse of motor predictions in action understanding or simulation (Jeannerod 2006). Furthermore, one can construct the formation of self-models as an internalization of the "self" as the center of predictions and experience (Hohwy and Michael forthcoming), or the construction of an epistemic agent model (Metzinger 2015). Or one can think of *somatic markers* in terms of an internalization of evaluative processes, which permits the running of "what if" loops (Damasio 1994). Finally, possible extensions of the same framework to social contexts, cultural and linguistic practices remain to be investigated (Clark 2016; Pezzulo et al. 2016b).

The extent to which one can describe higher forms of cognition in terms of the internalization of the process described here remains to be assessed, but using PP as a process model may help in charting this territory. It is also important to clarify that internalization does not exclude other ways to implement higher cognition; for example, this model is not necessarily antithetic to the idea that part of cognition is "externalized" or off-loaded to the environment, as in the case of using a computer (or even an abacus) to do maths, or rotating Tetris pieces to aid in deciding where to place them (the latter was indeed an early example of "epistemic action" in the literature, cf. Kirsh and Maglio 1994). A complete theory should encompass these and other possible roots to higher cognition.

## References


Allen, M. & Friston, K. J. (2016). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, 1–24.

Ambrose, R. E., Pfeiffer, B. E. & Foster, D. J. (2016). Reverse replay of hippocampal place cells is uniquely modulated by changing reward. *Neuron, 91* (5), 1124–1136.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences, 33* (04), 245-266. https://dx.doi.org/10.1017/S0140525X10000853.

Barrett, L. F. & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience,* 16, 419–429.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-600.

Botvinick, M. & Toussaint, M. (2012). Planning as inference. *Trends Cogn Sci, 16* (10), 485–488. http://dx.doi.org/10.1016/j.tics.2012.08.006.

Bruineberg, J., Kiverstein, J. & Rietveld, E. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*, 1–28.

Buckner, L. B. & Carroll, D. C. (2007). Self-projection and the brain. *Trends Cogn Sci, 11* (2), 49–57. https://dx.doi.org/10.1016/j.tics.2006.11.004.

Buesing, L., Bill, J., Nessler, B. & Maass, W. (2011). Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comput Biol, 7* (11), e1002211.

Butz, M. V. (2016). Towards a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology, 7*, 925.

Buzsáki, G. (2006). *Rhythms of the brain*. New York: Oxford University Press.

Buzsáki, G. & Moser, E. I. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature Neuroscience, 16* (2), 130–138.

Buzsáki, G., Peyrache, A. & Kubie, J. (2014). *Cold Spring Harbor symposia on quantitative biolog*. Emergence of cognition from action (pp. 41–50). Cold Spring Harbor Laboratory Press.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York: Oxford University Press.

——— (2015). Embodied prediction. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND:* 7(T). Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570115.

Clark, A. & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior, 7* (1), 5–16.

Cotterill, R. (1998). *Enchanted looms: Conscious networks in brains and computers*. Cambridge University Press.

Craig, A. D. (2015). *How do you feel? An interoceptive moment with your neurobiological self*. Princeton University Press.

Cruse, H. & Schilling, M. (2015). The bottom-up approach: Benefits and limits. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND:* 9(R). Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570931. http://open-mind.net/papers/the-bottom-up-approach-benefits-and-limits2014a-reply-to-aaron-gutknecht.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason and the human brain*. New York: Grosset/Putnam.

Daw, N. D., Niv, Y. & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8* (12), 1704–1711. https://dx.doi.org/10.1038/nn1560.

Dehaene, S. & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron, 56* (2), 384-98.

Diba, K. & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience, 10* (10), 1241–1242.

Donnarumma, F., Costantini, M., Ambrosini, E., Friston, K. & Pezzulo, G. (2017). Action perception as hypothesis testing. *Cortex*. http://dx.doi.org/10.1016/j.cortex.2017.01.016

Dragoi, G. & Tonegawa, S. (2011). Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature, 469* (7330), 397–401.

Feng, T., Silva, D. & Foster, D. J. (2015). Dissociation between the experience-dependent development of hippocampal theta sequences and single-trial phase precession. *The Journal of Neuroscience, 35* (12), 4890–4902.

Foster, D. & Wilson, M. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature, 440*, 680–683.

Foster, D. J. & Wilson, M. A. (2007). Hippocampal theta sequences. *Hippocampus, 17* (11), 1093–1099.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nat Rev Neurosci, 11* (2), 127–138. https://dx.doi.org/10.1038/nrn2787.

Friston, K. J. (2013). Life as we know it. *J R Soc Interface, 10* (86), 20130475. https://dx.doi.org/10.1098/rsif.2013.0475.

Friston, K. J. & Frith, C. (2015a). A duet for one. *Consciousness and Cognition*.

Friston, K. J. & Frith, C. D. (2015b). Active inference, communication and hermeneutics. *Cortex, 68*, 129–143. http://dx.doi.org/10.1016/j.cortex.2015.03.025.

Friston, K. J., Adams, R. A., Perrinet, L. & Breakspear, M. (2012a). Perceptions as hypotheses: Saccades as experiments. *Front Psychol, 3*, 151. https://dx.doi.org/10.3389/fpsyg.2012.00151.

Friston, K., Samothrakis, S. & Montague, R. (2012b). Active inference and agency: Optimal control without cost functions. *Biol Cybern, 106* (8-9), 523–541. https://dx.doi.org/10.1007/s00422-012-0512-8.

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. E. & Bestmann, S. (2012c). Dopamine, affordance and active inference. *PLoS Comput Biol, 8* (1), e1002327. https://dx-.doi.org/10.1371/journal.pcbi.1002327.

Friston, K. J., Schwartenbeck, P., Fitzgerald, T. A., Behrens, T. & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Front Hum Neurosci, 7*, 598. https://dx.doi.org/10.3389/fnhum.2013.00598.

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., FitzGerald, T. & Pezzulo, G. (2015). Active inference and epistemic value. *Cogn Neurosci, 6*, 187–214. https://dx.doi.org/10.1080/17588928.2015.1020053.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. (2016a). Active inference: A process theory. *Neural Computation*.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J. & Pezzulo, G. (2016b). Active inference and learning. *Neuroscience & Biobehavioral Reviews, 68*, 862–879.

Gallagher, S. (2005). *How the body shapes the mind.* Oxford.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences, 27* (03), 377–396.

Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S. & Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron, 65* (5), 695–705. https://dx.doi.org/10.1016/j.neuron.2010.01.034.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42* (1-3), 335–346.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences, 6*, 242–247.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D. & Couzin, I. D. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences, 19* (1), 46–54.

Hinton, G. E. (2007). To recognize shapes, first learn to generate images. *Progress in Brain Research, 165*, 535–547.

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

Hohwy, J. & Michael, J. (forthcoming). Why should any body have a self? In F. de Vignemont & A. Alsmith (Eds.) *The body and the self, revisited*. Cambridge, MA: MIT Press.

Jeannerod, M. (2006). *Motor cognition.* Oxford University Press.

Johnson, A. & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J Neurosci, 27* (45), 12176–12189. https://dx.doi.org/10.1523/JNEUROSCI.3761-07.2007.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

Kirsh, D. & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science, 18*, 513-549.

Kumaran, D., Hassabis, D. & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences, 20* (7), 512–534.

Lisman, J. & Redish, A. D. (2009). Prediction, sequences and the hippocampus. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 364* (1521), 1193–1201.

McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev, 102* (3), 419–457.

Metzinger, T. (2015). M-autonomy. *Journal of Consciousness Studies, 22* (11-12), 270–302.

———(2017). The problem of mental action. Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Meyniel, F., Schlunegger, D. & Dehaene, S. (2015). The sense of confidence during probabilistic learning: A normative account. *PLoS Comput Biol, 11* (6), e1004305.

Montague, R. (2006). *Why choose this book? How we make decisions.* EP Dutton.

O'Keefe, J. & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research Volume, 34*, 171–175.

Olafsdottir, H. F., Barry, C., Saleem, A. B., Hassabis, D. & Spiers, H. J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. *Elife, 4*, e06063.

Pavlov, I. P. & Thompson, W. H. (1902). *The work of the digestive glands.* Charles Griffin.

Pezzulo, G. (2012). An active inference view of cognitive control. *Frontiers in Theoretical and Philosophical Psychology, 3*, 478.

——— (2013). Why do you fear the bogeyman? An embodied predictive coding model of perceptual inference. *Cognitive, Affective, and Behavioral Neuroscience,* 14 (3), 902–11.

Pezzulo, G. & Castelfranchi, C. (2009). Thinking as the control of imagination: A conceptual framework for goal-directed systems. *Psychological Research, 73* (4), 559–577.

Pezzulo, G. & Cisek, P. (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends Cogn Sci, 20* (6), 414–424. https://dx.doi.org/10.1016/j.tics.2016.03.013.

Pezzulo, G. & Rigoli, F. (2011). The value of foresight: How prospection affects decision-making. *Front. Neurosci, 5* (79), 79.

Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K. & Spivey, M. (2011). The mechanics of embodiment: A dialogue on embodiment and computational modeling. *Frontiers in Psychology, 2* (5), 1-21.

Pezzulo, G., Rigoli, F. & Chersi, F. (2013). The mixed instrumental controller: Using value of information to combine habitual choice and mental simulation. *Front Psychol, 4*, 92. https://dx.doi.org/10.3389/fpsyg.2013.00092.

Pezzulo, G., van der Meer, M. A. & Lansink, C. S. and P. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in Cognitive Sciences, 18* (12), 647–657. https://dx.doi.org/10.1016/j.tics.2014.06.011.

Pezzulo, G., Rigoli, F. & Friston, K. J. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Prog Neurobiol, 134*, 17–35. https://dx.doi.org/10.1016/j.pneurobio.2015.09.001.

Pezzulo, G., Cartoni, E., Rigoli, F., Pio-Lopez, L. & Friston, K. J. (2016a). Active inference, epistemic value, and vicarious trial and error. *Learn Mem, 23* (7), 322–338. https://dx.doi.org/10.1101/lm.041780.116.

Pezzulo, G., Iodice, P., Donnarumma, F., Dindo, H. & Knoblich, G. (2016b). Avoiding accidents at the champagne reception: A study of joint lifting and balancing. *Psychological Science*.

Pezzulo, G., Kemere, C. & Van der Meer, M. (2017). Internally generated hippocampal sequences as a vantage point to probe future-oriented cognition. *Annals of the New York Academy of Sciences*. https://dx.doi.org/10.1111/nyas.13329.

Pfeifer, R. & Bongard, J. (2006). *How the body shapes the way we think: A new view of intelligence.* Cambridge, MA: MIT Press.

Pfeiffer, B. E. & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature, 497*, 74–79.

Piaget, J. & Cook, M. (1952). *The origins of intelligence in children*. International Universities Press New York.

Pirolli, P. & Card, S. (1999). Information foraging. *Psychological Review, 106* (4), 643.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80* (1-2), 127–158.

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

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience, 17* (3), 147–159.

Schacter, D. L. & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philos Trans R Soc Lond B Biol Sci, 362* (1481), 773–786. https://dx.doi.org/10.1098/rstb.2007.2087.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17* (11), 565–573.

Shenhav, A., Botvinick, M. M. & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron, 79* (2), 217–240.

Spivey, M. J. & Geng, J. J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychol Res, 65* (4), 235–241.

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology & Behavior, 106* (1), 5–15.

Stoianov, I., Genovesio, A. & Pezzulo, G. (2016). Prefrontal goal-codes emerge as latent states in probabilistic value learning. *Journal of Cognitive Neuroscience, 28* (1), 140–157.

Suddendorf, T. (2006). Foresight and evolution of the human mind. *Science, 312*, 1006–1007.

Tolman, E. C. (1938). The determiners of behavior at a choice point. *Psychological Review, 45* (1), 1.

Varela, F. J., Thompson, E. T. & Rosch, E. (1992). *The embodied mind: Cognitive science and human experience.* Cambridge, MA: MIT Press.

Wikenheiser, A. M. & Redish, A. D. (2015). Hippocampal theta sequences reflect current goals. *Nature Neuroscience, 18* (2), 289–294.

# The Overtone Model of Self-Deception

Iuliia Pliushch

In this paper I will argue for what I call an 'overtone model of self-deception'. The analogy to overtones (higher-order frequencies of a tone) is as follows: a self-deceiver's optimal degree of instability (the term is borrowed from Friston et al. 2012a, and applied to self-deception) is elevated so that constant exploration (of a certain number of hypotheses) is pursued instead of disambiguation in favor of a certain hypothesis. These hypotheses are explored in parallel (for similar ideas with respect to higher-order cognition in general see Pezzulo and Cisek 2016, and Metzinger 2017) and are like overtones of the currently active self-deceptive hypothesis (the base frequency) so that what we as self-deceivers, as well as observers, perceive as one tone (self-deception) is actually a fusion of different frequencies. The term 'fusion' is relevant because the phenomenology of the self-deceiver is co-determined by overtones.

**Keywords**

Binocular rivalry | Counterfactuals | Interoception | Overtones | Predictive coding | Predictive processing | Self-deception

## 1 Introduction

Recently, I was sitting in a train when I heard the following snippet of a conversation "Well, you know how the saying goes: I have gone through many terrible things and some of them really *were* terrible". In other words, what one perceives as terrible may, on reflection, turn out not to be so. Thus, humans often misrepresent things, be it for the better (optimists) or for the worse (pessimists). One may say that misrepresentation already is self-deception, or one may require additional criteria to label something ´self-deception´. In the following, I will elaborate on the kinds of behavioral and phenomenological criteria those may be. In any case, the above example shows that our perception of the world and of ourselves is always filtered. This is the basis for several phenomena that may occur, for example delusions or illusions, and also self-deception. Though more often than not self-deception is described as over-optimism or denial, such that the aspects filtered out are negative, this is not necessarily the case (see, e.g., Mele 2001 for "twisted" self-deception). The basis for both is a certain kind of distortion, which, though elusive when philosophically analyzed (there is still no consensus on what self-deception is, how it is brought about, or whether one should relinquish — abandon — it), is a frequent visitor of everyday conversations. To err on the side of caution, I do not claim that the above conversation demonstrates a case of self-deception — there is not enough information for such a judgement — but rather that, if enriched with details (see the template below), it can become one.

On the one hand, we are able to successfully interact in the world we live in, which means that there has to be at least a kernel of truth in the way we perceive it. On the other hand, there is quite some divergence between the way the world is and how we perceive it, else there would be no research on self-deception. As I will argue later, self-deception is a cluster concept. Many different phenomena have been subsumed under it. For the reader to get a better feeling for this concept, I will first give a folk-psychological template for self-deception. This is how we get acquainted with it in our everyday life. A word of caution though: the template involves the simplistic folk-psychological assumption that self-deception is a personal-level process, which is to say that self-deceivers possess an epistemic agent model (EAM, see Metzinger 2017) — a conscious model of them as directing their attention and guiding their reasoning process in a certain manner. I will question this assumption in the next section.

Pliushch, I. (2017). The Overtone Model of Self-Deception.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 21. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573222

*Self-Deception Template*
*A* is motivated to believe X. She starts the hypothesis testing process with the aim of finding the truth. Somehow, despite the unbiased *evidence* (or at least unbiased information, because acknowledging something as evidence is already a step further in information processing[1]) pointing into one direction, she ends up believing the conclusion that is not supported by the evidence because of *motivation Y* [the goal representation to believe this false conclusion/believe otherwise/relieve anxiety etc]. While being self-deceived, she experiences (at least sometimes) *tension* because of the inconsistency between the acquired conclusion and the evidence. Upon relinquishing self-deception, *A* experiences *insight*: she has the feeling of having known the truth all along. When questioned about believing *X* by observers, *A* would *justify* her belief, but at different time slices, her behavior would be *inconsistent* from the point of view of the observer with regard to her belief that *X*. The reader is encouraged to fill out *X*, *Y* and the concrete arguments with cases from her personal life.

Examples covered in the philosophical literature include denial of having a terminal illness such as cancer (Rorty 1988), denial of the unfaithfulness of one's spouse (Funkhouser 2009), and denial of one's child's criminal inclinations (Mele 2001). Among the psychological examples not only can one list unrealistic optimism and self-enhancement (Taylor 1989; Von Hippel and Trivers 2011), but also anosognosia, the denial of one's own paralysis (Levy 2009).

In this paper, inconsistency as characteristic of self-deception will take center stage. I employ the term 'inconsistency' here in a broad way as either first- or third-person inconsistency. First-person inconsistency is the possession of contradictory representations. Third-person inconsistency is behaving in a manner contradictory to one's verbal assertions, or to one's own behavioural dispositions. I call it 'third-person inconsistency,' because this kind of inconsistency is noticed and pointed out to us by others (friends, relatives, observers) and leads to another characteristic of self-deception, namely that self-deceivers *justify* themselves rather than relinquishing the self-deception. I employ this broad formulation of inconsistency because I do not want to commit myself to either an intentional or a deflationary account of self-deception. One point of disagreement between these two accounts is whether self-deception involves contradictory mental states. According to Alfred Mele's (Mele 2012) deflationary account, self-deception results from motivationally biased belief acquisition. In this vein, Erik Helzer and David Dunning (Helzer and Dunning 2012) hold that "motivated reasoning emerges as a paradigmatic case of self-deception". Yet this account is not without its problems, one of which is the 'slippery slope' problem that has been brought by Neil Van Leeuwen (Van Leeuwen 2013) against Robert Trivers (Trivers 2011) examples of self-deception. Van Leeuwen's (Van Leeuwen 2013) worry is that if the category of self-deception encompasses too large a range of phenomena, then the concept of self-deception would lose its scientific value. This criticism can be applied to the deflationary definition of self-deception as motivationally biased beliefs as well. If motivated biases are defined as those that "may be triggered and sustained by desires in the production of *motivationally* biased beliefs" (Mele 2012, p. 7), and if every cold (not motivated) bias can be triggered by motivation, then every bias is — potentially — self-deceptive. In this paper, to make the most philosophically compelling case, I sketch how one could analyze the strongest kind of inconsistency possible in self-deception. Weaker cases of inconsistency would be easier to analyze in a similar manner.

The main goal of this paper is to present a predictive processing inspired account of how to explain inconsistency in self-deception. First, I will present my own account of self-deception as a cluster concept with a certain behavioral and phenomenological profile. I will describe its profile, situate inconsistency as one of the behavioral characteristics of self-deception, and elaborate on how inconsistency has been previously incorporated into philosophical theories of self-deception. Thereafter, I will briefly introduce the predictive processing tools that I will need in order to, lastly, propose my

1 For discussion on acknowledging evidence see, for example, Michel and Newen 2010, and Bagnoli 2012.

'overtone theory of self-deception' for how inconsistency can be analyzed in cases of self-deception. I will argue that overtones enrich our phenomenal experience not only by generating tension, but also in enabling the experience of more nuanced affective consequences such as being glad that something is the case or anticipating that it is the case.

## 2 The Philosophical Problem: Self-Deception

In this section I will present my own account of self-deception and explain how inconsistency fits into the picture. Existing accounts of self-deception struggle in satisfying two important constraints for an adequate theory of self-deception, namely the *parsimony* and the *demarcation* constraints. The first constraint can be formulated as the requirement to keep the analysis of self-deception as simple as possible, e.g. not to postulate unnecessary internal states. The second requires identifying the criteria for distinguishing self-deception from other phenomena.

The necessity of the first constraint arises from the debate about the *nature of self-deceptive representations*, i.e. which kind of attitude self-deception is. Although the nature of a self-deceptive representation is standardly thought to be belief, alternatives exist. One is *pretense*, which is an attitude akin to imagining that fulfills a belief-like role (Gendler 2007). *Avowal* is another example of a belief-like attitude that has been argued to be produced by the process of self-deception (Audi 1997). A criticism brought against *avowal* and in favor of *belief* as a self-deceptive attitude is that there has to be independent motivation, apart from the wish to solve the paradox of self-deception, in order to postulate the attitude of avowal (Van Leeuwen 2007, p. 429-431). I think the parsimony constraint has to be applied not only to the nature of self-deceptive representations, but also to the question of the kind of self-deceptive motivation and the nature of the self-deceptive process.

The second constraint has its roots in the *intentionalist-deflationary debate*: it has been used as a point of critique for accounts of self-deception. The intentionalist-deflationary debate is about the nature of the motivation for self-deception. Intentionalists argue that self-deceivers are motivated by intentions, while proponents of deflationary theory argue for specific kinds of desires. The so-called selectivity problem is an example of the failure of the demarcation constraint. Curiously, both intentionalist and deflationary accounts have been argued to suffer from it. For example, it has been argued by Mele (Mele 2001) that it is possible to imagine cases in which one would not be able to self-deceive despite the intention to do so (p. 66). This is the so-called dynamic paradox of self-deception: if one has the intention to self-deceive, then in order to carry out this intention, one would need make oneself believe something one does not currently believe. On the other hand, José Luis Bermúdez (Bermúdez 2000) argues that it is not always the case that non-intentional motivation, such as desire, leads to the acceptance of certain self-deceptive hypotheses (p. 317). Another example is the criticism voiced by Neil Van Leeuwen against Robert Trivers that I mentioned in the previous section.

My solution to the parsimony and demarcation constraints is that self-deception is a cluster concept with a certain phenomenological and behavioral profile. Self-deception is a cluster concept in virtue of the slippery slope that has been enabled by the understanding of self-deception as a motivationally biased belief. The more behavioral and phenomenological properties characteristic of self-deception a phenomenon possesses, the stronger a case of self-deception it is. For example, inconsistency and justification belong to the behavioral properties. In other words, self-deceivers are prone to behave inconsistently while justifying their behavior. The self-deceiver's phenomenology is characterized by tension (feelings of uneasiness or distress), and insight when self-deception is relinquished. I will leave open the question of which and how many properties are required for a minimal case of self-deception. I would, however, like to frame self-deceptive motivation in terms of goal representations, because they are folk-psychologically neutral, unlike to intentions or desires.

From the fact that the self-deceiver's behavior is inconsistent, it might be deduced that there must be an inconsistency in their belief set. This is an inference to the best explanation given the following

three assumptions. First, beliefs are stored entities.[2] Second, beliefs (more often than other attitudes) determine our actions (Van Leeuwen 2007). Third, self-deceivers behave inconsistently (Funkhouser 2005; Funkhouser 2009). One possibility for framing this inconsistency is to argue that self-deceivers possess inconsistent *beliefs* (Davidson 1986).

Since, on the assumption that self-deceivers are generally rational beings, it is difficult to explain how such an inconsistency is possible, the inconsistency requirement has been weakened in many recent accounts to involve, for example, a belief and a suspicion that the belief is false (Mele 2001; Mele 2012), or attitudes other than beliefs are argued to result from self-deception, such as avowal (Audi 1997) or pretense (Gendler 2007). The implicit assumption in these accounts is that consistency is only required among attitudes of the same kind, whereas different kinds of attitudes can contradict each other. Psychological experiments testing self-deceptionsuppose the presence of inconsistency not among propositional attitudes, but between a propositional attitude and skin conductance response (Gur and Sackeim 1979) or between different types of processing: conscious/unconscious, implicit/explicit, automatic/controlled (Von Hippel and Trivers 2011).

In the remainder of this section I will address the process of self-deception and the properties of self-deceptive representations, like transparency or its affective component. In doing this, I will use the terminology of Thomas Metzinger's self-model theory of subjectivity. According to this theory, phenomenal mental models are those that incorporate consciously experienced content (Metzinger 2003). What an agent thinks of as real, is part of the transparent *world-model*, like an apple on a table. Transparency means that earlier processing stages are inaccessible. Thus, if something is transparent, it is experienced as real and not as a representation. This is true with respect to our experience of the world and of ourselves. We experience the world as real, not as the result of the transparent phenomenal world-model that has been constructed. The same applies to the phenomenal *self-model*. It is not the case, though, that the world- and self-models are the only models that an agent possesses. From time to time, humans use their cognitive capacities such that epistemic agent models (EAMs, see Metzinger 2017) are constructed. Epistemic agent models are transparent, conscious self-representations of the agent as possessing the capacity for epistemic agency (control of attention and control of goal-directed thoughts) and/or actually executing epistemic actions (Metzinger 2013; Pliushch and Metzinger 2015).

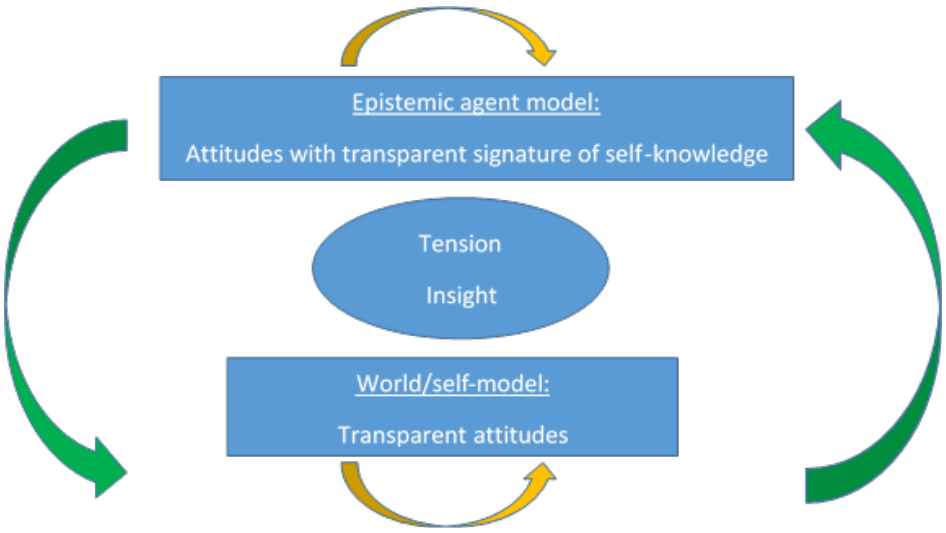


**Figure 1.** Phenomenology of the self-deceiver. Tension and insight are phenomenological characteristics of self-deception. The two kinds of model that can accommodate self-deceptive representations are the epistemic agent model and the world/self-model. The arrows denote changes in and between models due to changes in their various aspects, like transparency of certain attitudes.

2 For a criticism of this view and an argument in favour of a constructivist view which holds that beliefs are reevaluated and constructed at different time slices instead of being retrieved, see Michel 2014.

The application of these concepts to self-deception is as follows: the template I presented in the introduction suggests that a self-deceptive process involves what is folk-psychologically called 'conscious thoughts'. Mind wandering may also contain conscious thoughts, but since it lacks such characteristics as veto control (it cannot always be terminated at will) and epistemic agency, it is just a subpersonal process that has become conscious (Metzinger 2015, Metzinger 2017). Conscious thinking is a subpersonal process in that it is the result of specific patterns of neural activity that may be integrated into the epistemic agent model (Metzinger 2015). An epistemic agent model is the conscious representation of the system as an agent executing epistemic actions, e.g., directing attention or controlling strands of thought. The unfolding of the cognitive landscape is an alternation between mind wandering episodes and episodes of epistemic agent model construction. The switches between the two often go unnoticed and have, thus, been called a 'self-representational blink' (Metzinger 2013). What about the switches in directing an agent's attention from one object to another? Or the control of the succession of thoughts? I think that human reasoning is at best full of gaps: thoughts popping out, skipping (e.g., think about trying to understand a proof of a theorem given by a skillful mathematician — she will, without noticing it , do three steps at once that need to be explained to you in detail) and merging. As an example of merging, consider the argumentative fallacy of equivocation: using a certain concept in one way for the first part of an argument and then switching to another of its connotations in the second part. Here, the concept's ambiguity allowed its misuse. This gapful nature of reasoning may be described according to the dolphin model of cognition (Pliushch and Metzinger 2015) such that our argumentations are dolphin-like trajectories in the argumentative space that at certain points can merge and mutate into one another. In predictive processing terms, the reason for changes and fusions can be argued to consist in the ever changing precisions of pursued policies.

If self-deception is achieved by means of epistemic agent model construction, or reasoning in folk-psychological terms, then the result would be a certain attitude that is not experienced as real, but as a representation: as something that can be true or false. Another possibility is that a self-deceptive representation becomes transparent and thus becomes part of the agent's world- or self-model. Anosognosia, the denial of a disability, may be one example, such as when someone's misrepresentation that their arm is not paralyzed becomes transparent (Pliushch and Metzinger 2015). Another example is denial of pregnancy. Denial of pregnancy does not exhibit such characteristics of self-deception as inconsistency from the third-person perspective (in this case nobody would notice that the self-deceiver is self-deceiving) and, hence, there is no need for justification. No tension is present either, but there is something, below awareness, at the level of the autonomic system, which can be described as "knowledge" of the pregnancy (Sandoz 2011, p. 784). The reason for theassumption of such knowledge is a silhouette effect: the figure of someone denying pregnancy does not change; instead the fetus expands in a vertical direction (that of diaphragm; p. 783). When, upon medical examination, insight into their body condition comes, there are cases where almost instantly the position of the fetus changes and the pregnant woman's silhouette changes in front of the doctor. As Sandoz 2011 puts it, there is a "complete physical metamorphosis that had taken place so quickly because of her psyche" (p. 783). An explanation given by Sandoz 2011 (p. 784) for the silhouette effect is *reactive homeostasis:* reflexive immediate escape from the emergent situation. An emergent situation is one of denial, and is described as a case of the "persistence of paradoxical realities" (p. 784). The author wonders which kind of informational pathway between the semantic knowledge of pregnancy provided by the doctor and the autonomic system might have led to the immediate appearance of pregnancy signs. The tentative answer would be that the cognitive and interoceptive domain is more tightly connected than has been assumed thus far (see Pezzulo and Cisek 2016).

One should not overlook that a more thorough investigation of the phenomenology of the self-deceived is needed in order to establish whether in cases when the self-deceptive attitude is not itself transparent, it nevertheless contains another transparent element, namely the phenomenal signature of knowing. This may be a feature that self-deceptive attitudes share with intuitions. For example,

Thomas Metzinger and Jennifer Windt (Metzinger and Windt 2014) argue that in intuitions of certainty, the *phenomenal signature of knowing* has become transparent. The phenomenal signature of knowing is characterized by the phenomenology of direct accessibility of knowledge (which may be preceded by the initial phase of the phenomenology of ambiguity). Self-deceivers may then also be susceptible to the epistemological fallacy (E-fallacy), which consists in ascribing epistemic status to phenomenal states due to the presence of the phenomenal signature of knowing.

It is possible that the phenomenal signature of knowing may become transparent through repeated explanation giving. Sangeet Khemlani and Philip Johnson-Laird (Khemlani and Johnson-Laird 2012) conducted experiments in which participants had to detect logical inconsistencies. Their results suggest that when participants actively construct explanations of inconsistencies, this makes it harder for participants to detect them afterwards. In an attempt to frame self-deception within predictive processing, consider the following findings. Ecstatic seizures evoke a transparent feeling of knowing (a subjective feeling of certainty) and have been explained as a case when an interoceptive mismatch is not computed, such that prediction errors remain unexplained and as a result certainty in one's prediction arises (Picard 2013). When a feeling of knowing is transparent, one would be certain in knowing something, without knowing why.

Apart from transparency, another feature that attitudes may vary in are affective consequences. The affective dimension possesses many roles in self-deception:

- Cause/motivation (as desires) (e.g., Mele 2000)
- Phenomenological accompaniment or tension (e.g., Noordhof 2003)
- Functional role, e.g. to reduce anxiety (e.g., Johnston 1988; Barnes 1997)
- Content about which one self-deceives (e.g., Borge 2003; Damm 2011)
- Mechanism of self-deception: *emotional coherence satisfaction* according to which there are two sets of constraints — cognitive (activations) and emotional (valences) such that which attitude is accepted depends not only on cognitive, but also emotional constraints (Sahdra and Thagard 2003)

If Ray Jackendoff (Jackendoff 2012) is correct in arguing that every thought possesses an affective component, then, similarly to transparency, it is a feature of self-deceptive attitudes (and not only them, but of all attitudes in general) that they can vary in affective properties.

Summing up, how many attitudes self-deception involves; and which propositional content or other characteristics, such as transparency or affective consequences, self-deceptive attitudes possess, may vary. As a consequence, it is not only that propositional content of asserted self-deceptive attitudes may vary over time (the yellow arrows in figure 1), but also that the model to which the self-deceptive attitude belongs may change too (the green arrows in figure 1). An example of the change in the epistemic agent model over time is the claim that self-deceptive attitudes vary in their degree of conviction/certainty (see, e.g., Lynch 2012; Porcher 2012). An example of the change of the model itself over time would be a case in which self-deceptive attitudes become transparent.

## 3 New Conceptual Tools: The Predictive Processing Approach

In the previous section I suggested that self-deception can concern both the construction of epistemic agent models and world/self-models. Epistemic agent models — 'conscious thinking' — can be understood as personal level hypothesis testing, particularly given the assumption that human beings are mostly rational: a hypothesis is chosen, evidence is sought out, and conclusions are drawn depending on whether the hypothesis is supported or not. If predictive processing is true, then both the epistemic agent model and the world/self-model construction are cases of hypothesis testing, but of a subpersonal kind. This is the case because, first, predictive processing is a theory about subpersonal

level of information processing, according to which perception, action and cognition are the results of hierarchical prediction error minimization. Second, it is an inference to the best explanation that the "selection of mental representations into consciousness" can be accommodated by the predictive processing framework (Hohwy 2015, p. 321).

An obvious benefit for analyses of self-deception using predictive processing is that rationality is a personal-level property: hypotheses at the subpersonal level do not need to be consistent with each other. In this section I will, first, briefly introduce some useful aspects of predictive processing, then I will argue that self-deception is a case where ambiguity is preserved such that a range of hypotheses is present. The following section will be dedicated to describing these hypotheses and their relationships. For that I will employ the overtone metaphor.

Predictive processing is a subpersonal kind of hypothesis testing insofar as a causal hypothesis about the structure of the world is inferred or acted upon (Friston 2009). An agent possesses a *generative* model of the causal structure of its environment. It is generative in that it generates predictions of sensory input that is actually sampled. When sensory input deviates from predictions, there is a prediction error that leads to changes of predictions (or actions — to change the sensory input). The former process is called *perceptual inference*; the latter is called *active inference*. Predictions about the causes of sensory input can also be characterized as hypotheses, analogously to scientific hypotheses — both need to be tested (e.g. Gregory 1980, Seth 2015a).

Generative models are part of the computational level of description. They are to be distinguished from phenomenal mental models or "those mental models which are, functionally speaking, globally available for cognition, attention, and the immediate control of behavior" (Metzinger 2003, p. 210). It is possible that those two are formally equivalent (Hobson et al. 2014), yet as such they belong to different levels of description. Nevertheless, if predictive coding can cast light onto consciousness (Hohwy 2015), then phenomenal models, which describe our experiences, and computational models, which describe brain processes, relate to each other in some way. It is not an aim of this paper to establish this relation (which would require much more empirical testing), but if predictive processing is to be a theory of cognition (Clark 2013), particularly higher-order cognition such as conscious thought, then, as a first step, a hypothesis about self-deception (a higher-order cognitive phenomenon) can be voiced that employs the tools of predictive processing. I will now turn my attention to the predictive processing explanation of binocular rivalry that shares at least two characteristics with self-deception: there is contradictory input (1) that leads models to alternate (2). The idea that self-deception involves alternating models (*alternation*, tool #1) will be extended by the suggestion that alternation is brought about by the optimal degree of instability (*instability*, tool #2), which then leads to the proposal that in cases of self-deception there is no sequential alternation, but rather a *parallel* exploration of hypotheses (*counterfactual processing*, tool #3).

*Binocular rivalry* is characterized as rivalry between competing, alternating perceptual models, only one of which is usually experienced at a time. The procedure by which binocular rivalry is brought about is such that one image is shown to one eye and a different image is shown to the second eye. This procedure can evoke a condition in which it appears to the agent that both objects occur at the same spatiotemporal location (Hohwy et al. 2008). The viewer experiences an alternation between models about the causes of visual input, for example, that there is a face or a house in the vicinity (Hohwy et al. 2008). This is the case because of the prior that there is only *one* object per spatiotemporal location (1), as well as the presence of prediction errors when either model is chosen, in this case, that one sees a house or a face (2).

The application of the binocular rivalry analogy to self-deception is as follows: both are characterized by an alternating pattern of hypotheses that are provided by the system as an explanation of the sensory input. Both hypotheses cannot be true at the same time (for logical reasons in the case of self-deception and due to empirical or evolutionary priors in the case of binocular rivalry). In the case of binocular rivalry, different models explain different sensory prediction errors. In the case of self-de-

ception, it is argued that, at least sometimes, there is contradictory evidence (see previous section), analogously to the two different images presented to the eyes. Yet, and here the analogy to binocular rivalry breaks, self-deception extends over relatively long periods of time. To my best knowledge, study participants have been exposed to binocular rivalry only for short periods of time. For the description of longer-term contradiction, another predictive processing tool (#2) will be employed, namely the *optimal degree of instability*:

> In brief, if neuronal activity represents the causes of sensory input, then it should represent *uncertainty* about those causes in a way that precludes overly confident representations. This means that neuronal responses to stimuli should retain an *optimal degree of instability* that allows them to explore *alternative hypotheses* about the causes of those stimuli. (Friston et al. 2012a, p. 3; my italics)

In other words, in order to not miss something important (not to become prematurely fixed on a certain hypothesis), uncertainty has to be preserved to a certain degree. The term 'uncertainty' can be used at both the personal (an agent being uncertain over something) and subpersonal (representations in the model having low degree of precision) levels. Further, regarding the subpersonal usage of this term, different kinds of representations can possess differing precision, i.e. be more or less uncertain. In the case of binocular rivalry, there is uncertainty because of contradictory (given higher-level assumptions) sensory prediction *errors* (bottom-up). Yet in the case of exploration (see the quotation above), there seems to be uncertainty of *predictions* (top-down). Simplistically, one might consider that the first is outer and the second is inner uncertainty. In the case of self-deception, outer uncertainty (ambiguity of the situation that enables its alternative interpretation) has been postulated (Sloman et al. 2010).[3] I will argue for the case of inner case of uncertainty. Robin Carhart-Harris et al. (Carhart-Harris et al. 2014) hold that wishful, imaginative and creative thinking is characterized by more entropy (more states are visited by the system) than rational or goal-directed thinking. The reasoning is that rational and/or goal-directed reasoning impose constraints on the system such that it is driven towards a more limited set of states than, for example, in the case of wishful inferences, which are biased by emotions. Though I agree that there is enhanced entropy in cases of self-deception, I think that constraints in the form of goal-representations and affective states constrain it so as to enable intricate patterns of oscillation with respect to different properties of self-deceptive representations.

Optimism can be seen as occupying one end of the *optimal degree of instability* scale by which self-deceivers may be characterized, because instead of alternating between different judgments on a certain topic, in the case of optimism the judgment is fixed — it is overly positive. Optimism (see Friston et al. 2013) has been explained in predictive processing terms as a case in which control states (beliefs about future actions) are overly precise and influence the transitions between hidden states (assumed states of the world). In other words, states of the world which are beneficial to the agent's

3 Steven Sloman et al. (Sloman et al. 2010) follow George Quattrone's and Amos Tversky's (Quattrone and Tversky 1984) pain endurance paradigm. In Quattrone and Tversky 1984, study participants endured cold water for longer when told that their ability to do so was indicative of their health. Similarly, Sloman et al. (Sloman et al. 2010), argue that if subjects are quicker to reach a dot after being told that this correlates with their intelligence, then this is an indication of self-deception, or that the subject knew of the desirable correlation which led to speeding up, but denied doing this deliberately. Philip Fernbach et al. (Fernbach et al. 2014), similarly, led participants to believe that the way in which they search for objects in a picture (detailed or holistic) is indicative of self-control. If told that detailed search is indicative of self-control, participants search longer for objects, but deny the effort of doing so.
These are, without a doubt, interesting results, but they demonstrate a fairly weak kind of self-deception, because it is devoid of such properties as first-person inconsistency, tension and need for justification. Weak kinds of self-deception can be explained in predictive processing terms in the same way as optimism (see Friston's idea referred to in the main text). First, there are two kinds of theories about how motivation influences cognition: *qualitative (e.g.,* Kunda 1990) and *quantitative (e.g.,* Ditto et al. 1998). According to the qualitative theory, bias influences the way information is processed. Mele's theory of self-deception employs this understanding of motivation. According to the quantitative theory, motivation changes only the time that we spend theorizing. The pain endurance paradigm is of this kind. Fernbach et al. (Fernbach et al. 2014) argue that "self-deception is enabled by people's tendency to adopt a mental representation of their own behavior that yields the most beneficial inference" (p. 6). The most parsimonious way of explaining participant behavior, though, is not by means of reasoning or conscious inference, but by the circular dependency between goal representations and perception (see main text). Goals representations are, on par with intentions and desires, one way of describing what motivates agents.

goal representations are assumed. Thus, in the case of optimism there is *high* precision (low uncertainty) of control states. Control states represent the goals of the system. Optimism occurs because there is a circular dependency between perception and goal representations (beliefs about future actions). Since control states influence the transitions between hidden states, and vice versa, in the case where goal representations are certain to be fulfilled, they influence perception.

In the case of alternation between competing hypotheses, there is more instability than in the case of optimism. A certain degree of instability may mean that instead of hypotheses being *disambiguated*, ambiguity is *preserved*, enabling the switching that is characteristic of at least some cases of self-deception. Instability, per se, says only that there will be changes in explored hypotheses, but in the case of binocular rivalry and self-deception there is also *constancy*: exploration does not cover too much ground but rather keeps returning to a select few models. Thus to explain self-deception, one needs to not only explain what leads to an enhanced degree of instability, but also why that instability is resolved in the same way over time. A change in estimates of uncertainty over time is *volatility* (Mathys et al. 2011). Volatility can be applied to different objects, including priors, policies, and models themselves. Further research needs to be done to establish how volatility can account for the constancy of model switches, but I think that the more complex the self-deception, the more kinds of volatility will be involved.

So far I have compared self-deception to binocular rivalry — where models also may alternate — and argued that a certain degree of instability may lead to the alternation observed in self-deception. What I want to argue for now is that, contrary to binocular rivalry where models alternate sequentially, in the case of self-deception two or more hypotheses are explored *concurrently*. In order show this, I first need to describe the role of counterfactuals in predictive processing. This role is fundamental in that action — accomplished by active inference — involves counterfactuals that encode "what we would infer about the world, if we sample it in a particular way" (Friston et al. 2012b, p. 2). Then, when sampling takes place, these counterfactuals are realized by active inference. There is a set of counterfactuals because the form of representation is probabilistic: several bets with different probabilities are encoded,[4] and replaced when another set of bets minimizes errors better.[5] A benefit of such probabilistic encoding is that perception and action are intertwined to a greater degree, enabling quicker action, for example, estimating the right trajectory of a flying ball and acting to catch it (Clark 2016), because multiple affordances are taken care of at the same time:

> It [pro-active readiness] must allow many possible responses to be simultaneously *partially prepared,* to degrees dependent upon the current balance of evidence — including estimations of our own uncertainty — as more and more information is acquired. (Clark 2016, p. 179; my italics)

Affordances (potential actions) have been argued to be computed during embodied decisions, for example, the affordance of reaching a berry or an apple (Cisek 2012, Cisek and Pastor-Bernier 2014). Recently, this claim has been extended to higher-order cognition: the same neural circuits and resources are used during a cognition process, as well as overt action (Pezzulo and Cisek 2016). Metzinger (Metzinger 2017) draws an analogy been drawn between affordances in the 'external' world (actually affordances in the internal, transparent world-model) and "affordances for cognitive agency", e.g. what to think or to calculate next: mind wandering creates a constant stream of the latter. Competing affordances are computed in parallel if both can still be realized. This is like fleeing from the predator along a road that forks at some future point; until the fork is reached, the two paths for fleeing can be

4 As Andy Clark (Clark 2016, p. 181) points out: "At every level, then, the underlying form of representation remains thoroughly probabilistic, encoding a series of deeply intertwined bets concerning what is 'out there' and (our current focus) how best to act".

5 For example, in case of speech processing it is argued that "we may rely upon stored knowledge to guide a set of guesses about the shape and content of the present sound stream: guesses that are constantly compared to the incoming signal, allowing residual errors to decide between competing guesses and (when necessary) reject one set of guesses and replace it with another" (Clark 2016, p. 194).

traversed together (Clark 2016).6 Parallel hypothesis exploration in self-deception is in this context an extreme case in which the parallel computation does not stop when the fork has been reached; options which, given the evidence, should have been abandoned, are instead kept available.

Interestingly, for action (executing movements) to take place instead of perception (changing the current hypothesis about the state of the world), precision expectations have to be down-played, and insofar as this occurs, they have been argued to be self-deceptive:

> In sum, action (under active inference) requires a kind of targeted dis-attention in which current sensory input is attenuated so as to allow predicted sensory (proprioceptive) states to entrain movement. At first sight, this is a rather baroque […] mechanism […] involving an implausible kind of self-deception. According to this story, it is only by downplaying genuine sensory information specifying how our bodily parts are *actually* currently arrayed in space that the brain can 'take seriously' the *predicted* proprioceptive information that determines movement, allowing those predictions to act […] directly as motor commands. (Clark 2016, p. 217)

This quotation is interesting for two reasons. First, it demonstrates the wide scope in which the term 'self-deception' is used. Second, it leads into a discussion of which phenomena are appropriately labelled 'self-deception' and which should be termed 'misrepresentation' in the predictive processing framework. I think that the above example is one of misrepresentation, not self-deception, because none of the behavioral or phenomenological characteristics of self-deception are present. Thus, the fact that the predictive processing framework is able to incorporate misrepresentations does not mean that each case of misrepresentation is a case of self-deception, or that if misrepresentations are produced in particular ways they are necessarily also kinds of self-deception. The fulfilment of behavioral and phenomenological characteristics is important to self-deception. This can be illustrated by two examples. First, Jean Daunizeau et al. (Daunizeau et al. 2010) argue that "categorization errors are optimal decisions if the risk of committing an error quickly is smaller than responding correctly after a longer period" (p. 6). In such cases of a speed-accuracy conflict, a speedy decision and not an accurate one, is optimal.[7] Second, Rosalyn Moran et al. (Moran et al. 2014) argue that with age the *complexity* of the Bayesian model (of the causes of sensory input) decreases over time. Here, the idea is the following: to infer the causes of our sensations, a simpler or more complex model (with a higher number of parameters) can be constructed. Complexity can be roughly understood as the number of parameters needed to model those causes. The accuracy of the model (how well the model predicts the data) is not the only quantity that is minimized during prediction error minimization, rather, accuracy minus complexity is minimized[8] (for an exact definition of accuracy and complexity see Friston 2010). Thus complexity is a penalty term. It is a penalty because a model is more useful if it is generalizable to different pieces of data, which are not necessarily consistent with each other. Imagine that you have a pool of different data instances and a model can predict every instance perfectly. When it encounters a new piece of data and cannot predict it, the necessity of a perfect prediction would enforce a change to the model introducing new parameters, thereby making the model more complex. Thus, according to Moran et al. (Moran et al. 2014), Occam's razor, which is utilized here to avoid overfitting, leads to attenuated learning with age. This means that complexity reduction leads to the reduction of *short-term* Bayesian updating (according to which each time one encounters an instance the model cannot predict, one would be prone to change the model) and a shift to enhanced top-down processing (Moran et al. 2014, p. 6). The authors voice the optimistic conclusion that "as we age, we converge on an

6 Interestingly, competing affordances are resolved only when action is needed: "Instead, to minimize prediction error is to minimize failures to identify the affordances for action that the world presents. Here, a good strategy is to deliver (at every moment) a partial grip upon a number of competing affordances: an 'affordance competition' that is plausibly resolved only *as and when action requires*." (Clark 2016, p. 202; my italics)

7 Specifically, it is Bayes-optimal.

8 The reason for this is that the quantity that is to be minimized is actually free energy, which can be approximated by prediction error (Friston 2010).

accurate and parsimonious model of our particular world […] whose constancy we actively strive to maintain" (Moran et al. 2014, p. 1). Thus, the second example is one in which, given a complexity-accuracy conflict, the less complex model, not the more accurate one, wins. I doubt that these two ways *necessarily* lead to self-deception, nor are they the only ways in which self-deception in the predictive processing framework occurs. This is because it is not true that every misrepresentation is self-deceptive, or, for that matter, that it is every misrepresentation which has been acquired in a certain way. Rather, if some phenomenon fits the phenomenological and behavioral characteristics, then it is more or less self-deceptive, depending on how many characteristics it fulfils.

The main insight into self-deception thus far from the combination of the predictive processing approach and applying the tools of alternation (#1), optimal degree of instability (#2), and counterfactual processing (#3), is that there is no need for external evidence for the alternation of explanatory models, but that a high level of exploration can be kept in order to ensure that ambiguity prevails.[9] My main claim is that this high level of exploration, for the self-deceiver, is determined by the *optimal degree of instability*. This is congruent with psychological findings which show that self-deception supposedly correlates with openness to new experiences (Kurt and Paulhus 2008, p. 843; Paulhus and John 1998, p. 1030). To pin down the results so far, I have argued that in self-deception cases, the optimal degree of instability is such that there is more exploration than disambiguation. Self-deceivers, thus, seem to be open (on the personal level), but not open enough — they are not open to acknowledge their self-deception, and experience insight upon relinquishing their self-deception.

Not being open enough leads to a discussion of how intuitions contribute to self-deception, and how self-deception can be relinquished such that insight can kick in. In the previous section I hypothesized that self-deception and intuitions have a common phenomenological characteristic — a transparent signature of self-knowledge (Metzinger and Windt 2014). There is another feature that they might have in common, namely that one has to distract oneself in order to actually abandon both intuition and self-deception. Regarding intuitions, it has been argued that to change someone's intuitions, one needs to distract attention from the context about which intuitions are to be changed (Weatherson 2014, p. 526). If in the case of self-deception there is not enough openness, then the same may be true regarding that phenomenon. Mere distraction is not enough, though. Changes in attention allocation have been argued to be a mechanism to *uphold* self-deception, not *relinquish* it (e.g., Noordhof 2009; for critique of the application of thought-suppression to self-deception see Lynch 2014). Thus, the effects of distraction can vary; an additional element is needed to make the acceptance of self-deceptive content possible.

My hypothesis is that this element is not of conceptual, but of interoceptive nature. Compelling arguments for this claim can be found when considering the following empirical evidence. *Anosognosia* (Turnbull et al. 2014) and *unrealistic optimism* (McKay et al. 2013) are (temporarily) attenuated by caloric vestibular stimulation (CVS). CVS consists in applying cold water to the ear canal of the patient, eliciting a nystagmus, i.e. rapid eye movements. In other words, for a short period of time after CVS, anosognosics acknowledge their paralysis, and optimistic people become less overly optimistic. Thus, something non-cognitive — the vestibular apparatus — influences higher-order cognition. Interestingly, CVS is also argued to "modulate the alternation rate in binocular rivalry" (Mast et al. 2014, p. 8). This, on the one hand, strengthens my analogy between binocular rivalry and self-deception. On the other hand, the fact that CVS influences binocular rivalry also makes the matter more complex: what is the underlying mechanism for such an inference? Oliver Turnbull et al. (Turnbull et al. 2014), on the premise that the vestibular sense influences both affective and spatial processing of information, view

9 My idea bears a certain resemblance to the idea that tension between goal representations can be creative and upheld, instead of resolved: "Often, one does not (and perhaps cannot) seamlessly meld the two original conflicting goals into a unified higher-level goal, but rather holds them in creative tension with each other — both goals remain, and continue to pull in opposite directions in many instances, and one is simply required to decide one way or the other in any particular circumstance. Our view of cognitive coherence as a *defeasible* rational requirement (McIntyre 1990) enables us to permit this approach, and we argue further that in some cases this creative tension presents a preferable solution to integration." (Saunders and Over 2009, p. 328)

anosognosia as a "dynamic, emotional by-product of a cognitive deficit" (p. 24) in "veridical spatial cognition" (p. 21). In other words, they argue that anosognosics perceive the world (spatial cognition aspect) in an egocentric fashion, i.e. how they want the world to be (emotional by-product aspect). Given that both anosognosia and unrealistic optimism are reduced by CVS and the explanation of the former as an emotional by-product, this depiction might be applied to self-deception as well. The more cautious claim is, then, that emotions, by means of CVS as an intermediary, influence the process of self-deception. In this vein, Bigna Lenggenhager and Christophe Lopez (Lenggenhager and Lopez 2015) voice a hypothesis that vestibular stimulation might influence interoception (p. 14). On balance, the point, in predictive processing terms, is that perceptual, interoceptive and possibly other kinds of inferences, are connected (Clark 2016), and in the case of self-deception, this connection plays a more important role than has been acknowledged so far.

To sum up, while being self-deceived, there is exploration, which is bounded such that insight is precluded. In predictive processing, the sequences of beliefs about future actions are termed 'policy' and the value of a policy can be decomposed into *extrinsic* (utility) and *intrinsic* (exploration) reward (Friston et al. 2013). When the expected utility of a policy drops, exploration is engaged, such that alternatives are explored which might contradict the policy pursued before, allowing insight to occur. The central question now is why exactly such a policy is explored which leads to insight, given that previously, during the time of the self-deceptive episodue, uncertainty of the policy (which underlies exploration) was used to uphold the self-deceptive cycle (analogously to binocular rivalry), undermining insight. This is where motivation comes into the picture. Motivation determines when we start to self-deceive and when we relinquish self-deception. In the predictive processing framework, what motivation is and how it determines an agent's actions can be described in a very general manner. A standard view, because of its proximity to the folk-psychological level of description, is that motivation encompasses both goal representations and the affective consequences of our actions. A more general view would be one in terms of the reduction of free energy (of which prediction error is an approximation). Free energy is an upper bound on surprise regarding the sensory states of the agent. Minimization of surprise is crucial for survival of the agent (Friston 2010). There is no need for a value function in free energy minimization, as free energy itself is actually the only ultimate value that exists, and whatever other descriptions of purportedly valuable states one might give, if they do not minimize free energy, they would not actually be valuable. A cost function can be defined as the "rate of change of value" (Friston 2010, p. 8). This just means that the bigger the positive difference in value between two states (the current one and the state to be visited), the better. For example, if money were the value in question, then for a beggar, winning a lottery would result in a huge change of value.[10] If value is substituted by free energy reduction, then one would get Mateus Joffily and Giorgio Coricelli's (Joffily and Coricelli 2013) explanation of emotional valence as the rate of change of free energy, such that, for example, an agent who could reduce free energy the most would be very happy. The consequence for self-deception of such a general account of value is that motivation should not be seen as an independent element that can be switched on and off. Rather, the sheer fact that agents reduce free energy, or pursue one policy and not another, is a demonstration of the ubiquity of motivation in the agent's cognition and action. In the following section, I will turn my attention to the set of hypotheses that are explored in self-deception.

## 4 The Overtone Metaphor

In this section, I will apply the overtone metaphor to self-deception. Recall that my own philosophical thesis is that during self-deception, uncertainty is preserved and several hypotheses are explored, but

10 The *evolutionary* value of a phenotype has been argued to depend on the amount of time that is spent by the phenotype in valuable states (Friston 2010, p. 7).

that they are at the same time bounded in order to preclude insight, and that affective consequences may play a role in overcoming the boundary. I will now describe how I think they relate to each other.

I borrowed the term 'overtone' from music theory where it denotes an additional frequency of a tone beyond the base frequency. Each tone that you hear, e.g. a musical instrument playing or a cup breaking, has an intensity, as well as one or several frequencies (the lowest being called fundamental). Those frequencies determine how the tone sounds (its timbre) and can be modelled as sine or cosine waves. Such wave functions (several added sine and cosine waves) describe an *oscillation* (how much it oscillates in each direction is the amplitude and how long a wave is — i.e., how long one has to wait until the same value will be repeated — is the frequency). For example, a tone played on a violin will sound different to the same tone being sung. This is due to the difference in the expression of overtones. Imagine, for example, that you have computer software with which you can produce sound waves. Each wave will sound a certain way, and if you click to add more and more waves, then the sound produced will differ.

First and foremost, I want to mention that some authors in the self-deception literature (in virtue of the difficulty of ascribing a clear-cut belief to self-deceivers) have favored the view that self-deceptive attitudes oscillate in the degree of certainty (for a short summary see Pliushch and Metzinger 2015). Christoph Michel (Michel 2014) has even argued that the constancy of our beliefs is an accidental property thereof: at each moment in time attitudes are constructed from the evidence available at that time and, luckily, they often stay constant. Thus, what has been offered is the oscillation of one attitude over time.

What I want to propose instead is that, in analogy to physical action (see the previous section), there is a set of hypotheses that changes over time. A self-deceiver's optimal degree of instability (which determines how much exploration is pursued, in order to preclude overfitting a model) is heightened so that constant exploration (of a certain number of hypotheses) is pursued at the cost of disambiguation in favor of any particular hypothesis. These hypotheses are like overtones of the currently active self-deceptive hypothesis (the base frequency) so that what self-deceivers or observers perceive as one tone (self-deception) is actually a fusion of different frequencies. The hypotheses are those in which certain attributes — propositional content, transparency, and affective consequences — vary. It is a *fusion* insofar as the phenomenology is determined not only by the proposition that is currently verbally affirmed, but *by the entire set*. One of the reasons I favor the overtone metaphor is that overtones are different depending on the *kind* of instrument being played, hence the richness and shrillness of the sounds also changes from instrument to instrument (and from human voice to human voice). Analogously, self-deception is idiosyncratic such that whether and how it develops depends on several factors, such as the phenotype of the agent and his personality traits. Depending on the amount of tension one experiences during self-deception, it can be compared to a consonant (pleasant, absent tension) or a dissonant (unpleasant) sound.[11]

Let me consider two examples that concern sets of hypotheses in which the transparency attribute varies. First, consider the following statement: "I know that we broke up but I don't want to call him, because then it will really be over (I would know for sure)." The tools that a rational analysis would provide us with would not suffice to analyze such an example, because of the logical impossibility of believing a contradiction (knowing and not knowing the same proposition). Yet an agent might reflect on several models of reality that she is possessing at the same time — they are somehow connected with each other, overlaid upon each other so that in both cases only the agent remains constant, while thoughts and feelings change depending on the accepted reality model. In this example, I think that the base frequency is the knowledge about the break up, yet the other hypothesis — that it is not over — contributes to relieving the negative affective consequences of the first. The base frequency is transparent (or maybe also unconscious) to a certain degree.

11 I am grateful to Wanja Wiese for pointing this out to me.

Second, consider the following study. Laura Aymerich-Franch (Aymerich-Franch et al. 2014) tested the relationship between virtual self-similarity (the similarity between a person's virtual avatar and their real-world appearance) and social anxiety. Virtual self-similarity was varied in that participants were assigned avatars that were more or less similar to their appearance. They were then required to give a talk in front of a virtual audience, a task expected to produce social anxiety. The results were not statistically significant, but the trend indicated that embodying dissimilar avatars reduces public speaking-induced anxiety. One interpretation of the results is that embodying dissimilar avatars was associated with a weaker sense of presence. Thus, if the avatar is not experienced as being oneself, then embodying it would not lead to strong reactions. Alternatively, the results can be interpreted as showing that the world/self-model that one possesses affects one's cognitive and affective processes. Further, if one asked which self-model — one's own or that of the avatar — functioned as the subject's unit of identification during the public speaking task, my hypothetical answer would be: both, but each only to a certain degree.

These two examples are very simplistic. Both involve only two representations, which means there is only one overtone. Richer examples of self-deception with several overtones are also possible. Among other things, an agent's fears, hopes, and wishes, all of which are affective attitudes, can serve as overtones too.

My claim that overtones influence phenomenology rests on the hypothesis that counterfactuals possess such influence. In the previous section I used the claim that counterfactuals implement physical action to argue that several hypotheses might be explored at once. Notably, the role of counterfactuals in predictive processing is not restricted to that of implementing physical action (see also Metzinger 2017). They have also been hypothesized to lead to certain kinds of phenomenology. In this vein, the concept of 'counterfactual richness' has been introduced by Anil Seth (Seth 2014) as a range of counterfactual sensorimotor contingencies[12] and it is argued to influence objecthood and/or the phenomenology of how real[13] an object appears (Seth 2015b). The idea can be most intuitively described on the personal level as follows. We could use any given object in several ways. This means we can imagine how an object would change given our actions, which leads to a sense of objecthood. For example, if there is a potato in front of me, I see it as something I could peel, throw, or turn around its axis, instead of simply seeing a picture of a potato. What is invariant in all those counterfactual scenarios is the object — the potato. Counterfactual sensorimotor contingencies are the subpersonal equivalent of this idea. The implication of this idea is that insofar as our counterfactuals are not violated when we use objects, they appear real.

Generalizing this line of thinking, this means that counterfactuals that represent sensorimotor contingencies influence certain aspects of our phenomenology. In transferring this claim from perception to cognition, a problem arises: in cognition there are no sensorimotor contingencies. Luckily, *exteroceptive* (directed at the outer world) counterfactuals do not exhaust our possibilities, because given that there is autonomic control too, there may be interoceptive counterfactuals that also influence experience in certain ways. For example, while in allostatic control, proprioceptive and kinematic consequences of actions are represented and then fulfilled via active inference, in autonomic control, interoceptive counterfactuals may represent the affective consequences of actions. Interestingly, the idea that the affective dimension depends on the presence of counterfactuals of a certain kind does not depend on predictive processing. For example, Jérôme Dokic (Dokic 2012) hypothesizes that one

12 But note that Karl Friston (Friston 2014) argues that there are also "purely sensory expectations that do not inform future or counterfactual outcomes — and have no sensorimotor contingency" (p. 120).

13 Referring to the need to distinguish the personal and subpersonal use of counterfactuals, the determination of the degree of realness by counterfactuals leads to the possibly counter-intuitive conclusion that subpersonal counterfactual selection cannot determine personal counterfactual selection. Personal counterfactual selection is selection among the representations of possible *phenomenal* world- and self-models one could possess. If transparency correlates with counterfactual richness (Seth 2014; Metzinger 2014), then counterfactual richness would play a different role than the phenomenal possible worlds we might represent — it might aid in our experiencing those phenomenal possible worlds as real in the first place due a change in the degree of transparency. This conclusion came up in our discussion of mental agency with Wanja Wiese for the poster at KogWis14.

could describe the *degree* of a metacognitive feeling modally as depending on the number of possible worlds in which a certain mental action would be successful. Thus, for example, a strong feeling of knowing indicates that one's competence is robust and will not easily fail in nearby possible worlds. Drawing on these findings, I wish to extend the view found in the self-deception literature that inconsistent representations might produce tension (characterized by a feeling of uneasiness, see previous sections) and argue for a broader claim. Overtones enrich experience not only in degrees of transparency, but also through more nuanced affective consequences such as being glad that something is the case or anticipating that it is the case.

Thus far in this section, I have focused on and elaborated one possibility for the application of the overtone metaphor: that it may be applied to self-deceptive hypotheses which, if phenomenally available, are often characterized as beliefs that something is the case, such as the belief that this paper is insightful. There is an alternative way in which the metaphor might be applied, namely with respect to the different ways in which self-deceptive beliefs are *justified*. Justification of a beliefs requires an epistemic agent model and, so self-deception about justification is a special case. For example, consider a lazy person who is content not learning something new (such as the Italian language) in her spare time. Upon being questioned about why this is the case, several (possibly inconsistent) lines of argumentation may be generated, for example, too much to do at work, too expensive, or no conversation partners available to practice with. If one line of reasoning no longer fits, such as a pay rise making paying for lessons no longer too expensive, another justification is chosen. The actual underlying reason in this case, however, is just laziness, but this is a negative self-characteristic which not everyone is ready to accept.

In summary, I have shown that a predictive processing account of subpersonal hypothesis testing indicates that sets of hypotheses are tested in perception, cognition and action, but we also experience a more or less unified sequential phenomenology, even in cases of self-deception. To reconcile both these observations and offer a description of the self-deceivers' inconsistency, I propose applying the overtone metaphor to self-deception: in cases of self-deception, several hypotheses are explored in parallel such that there is a basic phenomenally available hypothesis and one or more overtones that to different degrees also are reflected in the phenomenal experience.

## 5 Conclusion

In this paper, I introduced the *overtone* model of self-deception. After introducing the topic, I argued in the second section that in classical philosophical discussion, "self-deception" is a cluster concept that is characterized by behavioral and phenomenological profiles. I focused on the behavioral characteristic of inconsistency. In the third section, I compared self-deception to binocular rivalry and came to the conclusion that self-deception is a kind of continued exploration in which disambiguation is precluded. My main claim in this paper was that several (subpersonal) hypotheses might be explored at once. In the final section, then, I applied the overtone metaphor to the hypotheses that are continually explored in self-deception. I argued that the phenomenology of the self-deceiver is determined by the fusion of the overtones that the self-deceiver possesses. What those overtones look like (or how they develop) depends on the idiosyncratic characteristics of the self-deceiver.[14]

14 I want to thank Thomas Metzinger, Wanja Wiese, Lisa Quadt and Paweł Gładziejewski for extremely helpful comments, as well as the Barbara-Wengeler foundation (BWS) for funding my PhD thesis, ideas from which can be found in this paper. I am also very grateful to Lucy Mayne, who has proofread the English language.

## References


Audi, R. (1997). Self-deception vs. self-caused deception: A comment on professor Mele. *Behavioral and Brain Sciences, 20* (1), 104.

Aymerich-Franch, L., Kizilcec, R. F. & Bailenson, J. N. (2014). The relationship between virtual self similarity and social anxiety. *Frontiers in Human Neuroscience, 8*, 944. https://dx.doi.org/10.3389/fnhum.2014.00944.

Bagnoli, C. (2012). Self-deception and agential authority. A constitutivist account. *Humana.Mente Journal of Philosophical Studies, 20*, 99–116.

Barnes, A. (1997). *Seeing through self-deception.* Cambridge, New York: Cambridge University Press.

Bermúdez, J. L. (2000). Self-deception, intentions, and contradictory beliefs. *Analysis, 60* (4), 309–319.

Borge, S. (2003). The myth of self-deception. *The Southern Journal of Philosophy, 41* (1), 1–28. https://dx.doi.org/10.1111/j.2041-6962.2003.tb00939.x.

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D. R. & Nutt, D. (2014). The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience, 8*, 20. https://dx.doi.org/10.3389/fnhum.2014.00020.

Cisek, P. (2012). Making decisions through a distributed consensus. *Current Opinion in Neurobiology, 22* (6), 927–936. https://dx.doi.org/10.1016/j.conb.2012.05.007.

Cisek, P. & Pastor-Bernier, A. (2014). On the challenges and mechanisms of embodied decisions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 369* (1655). https://dx.doi.org/10.1098/rstb.2013.0479.

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

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

Damm, L. (2011). Self-deception about emotion. *The Southern Journal of Philosophy, 49* (3), 254–270.

Daunizeau, J., den Ouden, H. E., Pessiglione, M., Kiebel, S. J., Friston, K. J., Stephan, K. E. & Sporns, O. (2010). Observing the observer (ii): Deciding when to decide. *PLoS ONE, 5* (12), e15555. https://dx.doi.org/10.1371/journal.pone.0015555.

Davidson, D. (1986). Deception and division. In J. Elster (Ed.) *The multiple self* (pp. 79–92). Cambridge: Cambridge University Press.

Ditto, P. H., Scepansky, J. A., Munro, G. D., Apanovitch, A. M. & Lockhart, L. K. (1998). Motivated sensitivity to preference-inconsistent information. *Journal of Personality and Social Psychology, 75* (1), 53–69. https://dx.doi.org/10.1037/0022-3514.75.1.53.

Dokic, J. (2012). Seeds of self-knowledge: Noetic feelings and metacognition. In M. J. Beran, J. Brandl, J. Perner & J. Proust (Eds.) *Foundations of metacognition* (pp. 302–321). Oxford: Oxford University Press.

Fernbach, P. M., Hagmayer, Y. & Sloman, S. A. (2014). Effort denial in self-deception. *Organizational Behavior and Human Decision Processes, 123* (1), 1–8. https://dx.doi.org/10.1016/j.obhdp.2013.10.013.

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

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

———(2014). Active inference and agency. *Cognitive Neuroscience, 5* (2), 119–121. https://dx.doi.org/10.1080/17588928.2014.905517.

Friston, K. J., Breakspear, M. & Deco, G. (2012a). Perception and self-organized instability. *Frontiers in Computational Neuroscience, 6*. https://dx.doi.org/10.3389/fncom.2012.00044.

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

Friston, K. J., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T. & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Frontiers in Human Neuroscience, 7*, 598. https://dx.doi.org/10.3389/fnhum.2013.00598.

Funkhouser, E. (2005). Do the self-deceived get whay they want? *Pacific Philosophical Quarterly, 86* (3), 295–312. https://dx.doi.org/10.1111/j.1468-0114.2005.00228.x.

———(2009). Self-deception and limits of folk psychology. *Social Theory and Praxis, 35* (1), 1–16.

Gendler, T. S. (2007). Self-deception as pretense. *Philosophical Perspectives, 21* (1), 231–258. https://dx.doi.org/10.1111/j.1520-8583.2007.00127.x.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 290* (1038), 181–197.

Gur, R. C. & Sackeim, H. A. (1979). Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology, 37* (2), 147–169. https://dx.doi.org/10.1037/0022-3514.37.2.147.

Helzer, E. & Dunning, D. (2012). On motivated reasoning and self-belief. In S. Vazire & T. D. Wilson (Eds.) *Handbook of self-knowledge* (pp. 379–396). New York: Guilford Publications.

Hobson, J. A., Hong, C. C.-H. & Friston, K. J. (2014). Virtual reality and consciousness inference in dreaming. *Frontiers in Psychology, 5*. https://dx.doi.org/10.3389/fpsyg.2014.01133.

Hohwy, J. (2015). Prediction error minimization, mental and developmental disorder, and statistical theories of consciousness. In R. Gennaro (Ed.) *Disturbed consciousness* (pp. 293–324). Cambridge, MA: MIT Press.

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

Jackendoff, R. (2012). *A user's guide to thought and meaning.* New York: Oxford University Press.

Joffily, M. & Coricelli, G. (2013). Emotional valence and the free-energy principle. *PLoS Computational Biology, 9* (6), e1003094. https://dx.doi.org/10.1371/journal.pcbi.1003094.

Johnston, M. (1988). Self-deception and the nature of the mind. In B. P. McLaughlin & A. O. Rorty (Eds.) *Perspectives on self-deception* (pp. 63–91). Berkeley C.: University of California Press.

Khemlani, S. S. & Johnson-Laird, P. N. (2012). Hidden conflict: Explanations make inconsistencies harder to detect. *Acta Psychologica, 139*, 486–491.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108* (3), 480–498. https://dx.doi.org/10.1037/0033-2909.108.3.480.

Kurt, A. & Paulhus, D. L. (2008). Moderators of the adaptiveness of self-enhancement: Operationalization, motivational domain, adjustment facet, and evaluator. *Journal of Research in Personality, 42* (4), 839–853. https://dx.doi.org/10.1016/j.jrp.2007.11.005.

Lenggenhager, B. & Lopez, C. (2015). Vestibular contributions to the sense of body, self, and others. In T. Metzinger & J. M. Windt (Eds.) *Open MIND:* 23(T). Frankfurt am Main: MIND Group.

Levy, N. (2009). Self-deception without thought experiments. In T. Bayne & J. Fernández (Eds.) *Delusion and self-deception* (pp. 227–242). New York: Psychology Press.

Lynch, K. (2012). On the ‚tension" inherent in self-deception. *Philosophical Psychology, 25* (3), 433–450. https://dx.doi.org/10.1080/09515089.2011.622364.

——— (2014). Self-deception and shifts of attention. *Philosophical Explorations, 17* (1), 63–75. https://dx.doi.org/10.1080/13869795.2013.824109.

Mast, F. W., Preuss, N., Hartmann, M. & Grabherr, L. (2014). Spatial cognition, body representation and affective processes: The role of vestibular information beyond ocular reflexes and control of posture. *Frontiers in Integrative Neuroscience, 8*, 44. https://dx.doi.org/10.3389/fnint.2014.00044.

Mathys, C., Daunizeau, J., Friston, K. J. & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience, 5* (39). https://dx.doi.org/10.3389/fnhum.2011.00039.

McKay, R., Tamagni, C., Palla, A., Krummenacher, P., Hegemann, S. C., Straumann, D. & Brugger, P. (2013). Vestibular stimulation attenuates unrealistic optimism. *Cortex, 49* (8), 2272–2275. https://dx.doi.org/10.1016/j.cortex.2013.04.005.

Mele, A.R. (2000). Self-deception and emotion. *Consciousness & Emotion, 1* (1), 115–137. https://dx.doi.org/10.1075/ce.1.1.07mel.

Mele, A. R. (2001). *Self-deception unmasked.* Princeton, NJ: Princeton University Press.

——— (2012). When are we self-deceived? *Humana.Mente Journal of Philosophical Studies, 20*, 1–15.

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

——— (2013). The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology, 4*, 931. https://dx.doi.org/10.3389/fpsyg.2013.00931.

———(2014). How does the brain encode epistemic reliability? Perceptual presence, phenomenal transparency, and counterfactual richness. *Cognitive Neuroscience, 5* (2), 122–124. https://dx.doi.org/10.1080/17588928.2014.905519.

———(2015). M-Autonomy. *Journal of Consciousness Studies, 22* (11-12).

———(2017). The problem of mental action. Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing.* Frankfurt am Main: MIND Group.

Metzinger, T. & Windt, J. (2014). Die phänomenale Signatur des Wissens: Experimentelle Philosophie des Geistes mit oder ohne Intuitionen? In T. Grundmann, J. Hor-

vath & J. Kipper (Eds.) *Die experimentelle Philosophie in der Diskussion* (pp. 279–321). Berlin: Suhrkamp.

Michel, C. (2014). *Self-knowledge and self-deception: The role of transparency in first personal knowledge*. Münster: mentis.

Michel, C. & Newen, A. (2010). Self-deception as pseudo-rational regulation of belief. *Consciousness and Cognition, 19* (3), 731–744. https://dx.doi.org/10.1016/j.concog.2010.06.019.

Moran, R. J., Symmonds, M., Dolan, R. J., Friston, K. J. & Sporns, O. (2014). The brain ages optimally to model its environment: Evidence from sensory learning over the adult lifespan. *PLoS Computational Biology, 10* (1), e1003422. https://dx.doi.org/10.1371/journal.pcbi.1003422.

Noordhof, P. (2003). Self-deception, interpretation and consciousness. *Philosophy and Phenomenological Research, 67* (1), 75–100. http://www.jstor.org/stable/20140582.

———(2009). The essential instability of self-deception. *Social Theory and Practice, 35* (1), 45–71.

Paulhus, D. L. & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-decceptive styles with basic traits and motives. *Journal of Personality and Social Psychology, 66* (6).

Pezzulo, G. & Cisek, P. (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in Cognitive Sciences, 20* (6), 414–424. https://dx.doi.org/10.1016/j.tics.2016.03.013.

Picard, F. (2013). State of belief, subjective certainty and bliss as a product of cortical dysfunction. *Cortex, 49* (9), 2494–2500. https://dx.doi.org/10.1016/j.cortex.2013.01.006.

Pliushch, I. & Metzinger, T. (2015). Self-deception and the dolphin model of cognition. In R. Gennaro (Ed.) *Disturbed consciousness* (pp. 167–207). Cambridge: MA, MIT Press.

Porcher, J. E. (2012). Against the deflationary account of self-deception. *Humana.Mente Journal of Philosophical Studies, 20*, 67–84.

Quattrone, G. A. & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology, 46* (2), 237–248. https://dx.doi.org/10.1037/0022-3514.46.2.237.

Rorty, A. O. (1988). The deceptive self: Liars, layers, and lairs. In B. P. McLaughlin & A. O. Rorty (Eds.) *Perspectives on self-deception* (pp. 11–28). Berkeley: University of California Press.

Sahdra, B. & Thagard, P. (2003). Self-deception and emotional coherence. *Minds and Machines, 13* (2), 213–231.

Sandoz, P. (2011). Reactive-homeostasis as a cybernetic model of the silhouette effect of denial of pregnancy. *Medical Hypotheses, 77* (5), 782–785. https://dx.doi.org/10.1016/j.mehy.2011.07.036.

Saunders, C. & Over, D. E. (2009). In two minds about rationality? In J. Evans & K. Frankish (Eds.) *In two minds* (pp. 317–334). Oxford: 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. https://dx.doi.org/10.1080/17588928.2013.877880.

——— (2015a). The cybernetic Bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. M. Windt (Eds.) *Open MIND:* 35(T). Frankfurt am Main: MIND Group.

———(2015b). Presence, objecthood, and the phenomenology of predictive perception. *Cognitive Neuroscience, 6* (2-3), 111–117. https://dx.doi.org/10.1080/17588928.2015.1026888.

Sloman, S. A., Fernbach, P. M. & Hagmayer, Y. (2010). Self-deception requires vagueness. *Cognition, 115*, 268–281. https://dx.doi.org/10.1016/j.cognition.2009.12.017.

Taylor, S. E. (1989). *Positive illusions: Creative self-deception and the healthy mind.* New York: Basic Books.

Trivers, R. (2011). *Deceit and self-deception: Fooling yourself the better to fool others.* London: Allen Lane.

Turnbull, O. H., Fotopoulou, A. & Solms, M. (2014). Anosognosia as motivated unawareness: The ‚defence' hypothesis revisited. *Cortex, 61*, 18–29. https://dx.doi.org/10.1016/j.cortex.2014.10.008.

Van Leeuwen, N. D. (2007). The product of self-deception. *Erkenntnis, 67* (3), 419–437.

——— (2013). The folly of fools: The logic of deceit and self-deception in human life. *Cognitive Neuropsychiatry, 18* (1-2), 146–151. https://dx.doi.org/10.1080/13546805.2012.753201.

Von Hippel, W. & Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences, 34* (01), 1–16. https://dx.doi.org/10.1017/S0140525X10001354.

Weatherson, B. (2014). Centrality and marginalisation. *Philosophical Studies, 171* (3), 517–533. https://dx.doi.org/10.1007/s11098-014-0289-9.

# Action-Oriented Predictive Processing and Social Cognition

Lisa Quadt

The research field on social cognition currently finds itself confronted with two conflicting theoretical camps, cognitivism and enactivism. In their most extreme formulations, the former claims that mindreading skills exhaust our social cognitive capacities, while the latter stresses the sufficiency of interaction and embodiment. My aim is to find a middle position that provides the basis for discussing social cognition as interactive and embodied, while remaining in non-radical territory.

This can be achieved by situating social cognition within the framework of action-oriented predictive processing (Clark 2013). Specifically, I propose three conceptual tools, namely (1) embodied social inference (EmSI), (2) action-oriented predictions (a-o predictions) (Clark 2016), and (3) interactive inference (InI).

The first concept of EmSI refers to the more general term of "embodied inference" (Friston 2012), which means that an organism's morphology incorporates the demands of its environment. This idea can be applied to the social realm, in the sense that the kind of body an individual has constrains the kind of social interaction they can engage in. While humans, for example, can exploit their speech apparatus for communication, ants instead rely on their pheromone system. The body of an individual thus also constrains social cognitive skills and can be said to play a crucial role for interactions. This becomes obvious when considering the second concept of "action-oriented predictions". The basic idea is that the job of a predictive model is to distribute the cognitive workload and recruit embodied action whenever possible. Here too the body plays an indispensable role in that it realizes prediction error minimization by engaging with the external world via active inference. Related to this idea is the last concept of "interactive inference". I claim that interaction plays the same role for social cognition as action does for general cognition — namely gathering information about the *social environment* and thus actively sculpting not only one's external, but also internal environment. InI can be described as the minimization of prediction error while navigating the social environment. It serves to actively sample proof for predictions or to disambiguate competing models about the other.

In what I call replicative interactive inference (RInI), the bodily state (e.g., posture, movements) of another person is mimicked in order to supplement exteroceptive information about them with interoceptive and proprioceptive information. Mimicry, synchronization and automatic imitation are instances of RInI that function to make predictions about the other more precise by increasing the number of signal sources that yield relevant information.

Secondly, complementary interactive inference (CInI) refers to changing one's internal or external environment in response to the other person. It serves to either regulate the other's current state (e.g., mothers lowering their body temperature to cool down their infant's feverish body; Nyqvist et al. 2010), or to evoke further behavioral responses that then serve as additional exteroceptive input (e.g., using gestures to express one's uncertainty).

These conceptual tools can serve to alleviate the tension between enactivist and cognitivist theories. The present proposal thereby enables a dialogue about social cognition as an interactive and embodied process.

**Keywords**

Active inference | Action-oriented predictive processing | Action-understanding | Embodied inference | Interaction | Social cognition

Quadt, L. (2017). Action-Oriented Predictive Processing and Social Cognition.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 22. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573239

## 1 Introduction

The topic I will pursue in the current paper concerns the implications that predictive processing (PP; Hohwy 2013) has for research on social cognition. More specifically, I will discuss the possibilities that action-oriented PP (Clark 2013) holds for beginning to build a comprehensive theoretical framework for social cognition.

The paper is divided into two parts. In the first part I argue that a fresh view on the phenomenon is needed because the research field of social cognition currently finds itself confronted with two conflicting theories, viz., cognitivism and phenomenology/enactivism (henceforth *phenactivism*[1]). In their radical[2] formulations, the former claims that so-called mindreading skills — i.e., simulation (Gallese and Goldman 1998) and theoretical inference (Gopnik and Wellman 1992) — exhaust our social cognitive skills. The latter, on the other hand, emphasizes that social cognition entails embodied interaction and even claims that interaction patterns may constitute social cognitive processes (De Jaegher and Di Paolo 2007). This situation is problematic because both theoretical camps have their problems, which leave them unfit for serving as the basis for a comprehensive theory of social cognition. Phenactivism, it will be claimed, has neither a sound conceptual nor empirical basis, and therefore is unable to provide the means for a theoretical framework in which social cognition can be embedded. Cognitivism, on the other hand, neglects the issues of interaction and embodiment almost entirely and thus draws an incomplete picture of the manifold phenomenon of social cognition. At the same time, however, both theories make valuable positive proposals that should be considered in a theory of social cognition.

In the second part of the paper I argue that what is required to alleviate this tension is a new view on social cognition that integrates insights from both sides of the theoretical spectrum while remaining in non-radical territory. Action-oriented PP, as will be described in section 3, provides the conceptual tools to do just that. The term has been introduced by Clark (Clark 2013) to capture the idea that PP unifies action, perception and cognition in one theoretical framework. Perception and action are thought to follow the same computational principles and to crucially depend on each other in their joint mission to minimize prediction error. Where perception generates prior expectations about the unfolding of sensory consequences, action functions to fulfill these expectations by sampling the world (cf. Friston 2009, p.12). In this scheme, perception cannot do without action, and *vice versa*. Three aspects of action-oriented PP will be discussed in order to later embed them into the context of social cognition, viz., embodied inference (Friston 2010), action-oriented predictions (Clark 2016), and active inference (Clark 2015b).

In section 4, I aim to exploit this picture of the mind as drawing both on internal modeling and engagement of the environment by embodied agents in order to make it fruitful for research on social cognition. I will introduce three conceptual tools that shall provide a conceptual basis for further research. First, the notion of embodied social inference (EmSI) is presented. EmSI is meant to capture the idea that the very physiology of an agent constrains their range of social interactions. Secondly, the concept of action-oriented predictions is applied to and made fruitful for social cognition. Lastly, I introduce the term interactive inference (InI) in order to be able to assign an important role to interactive processes for social cognition.

1 Enactivism puts much emphasis not only on the body, but especially on interaction as a potentially constitutive element of social cognition. The difference between enactive and phenomenological theories seems to boil down to the explanatory scope. While enactivism explicitly claims to offer a radically different alternative to cognitivism and thus to build a proper account of cognition (Varela et al. 1993), phenomenology is mostly seen as a description of experiential phenomena (Gallagher 2008). I use the word 'phenactivism' to describe views that merge phenomenology and enactivism. Since they share fundamental premises (Quadt 2015) they can be subsumed under this concept.

2 I use the term radical in the sense of 'extreme', not in the sense of "anti-representationalist". It is important to notice that I will here describe only one and a rather radical version of each theoretical strand. Of course, either theory has been presented in various ways and with differing assumptions, some more and some less radical. Presenting the multitude of versions of each theory is neither necessary nor within the scope of this paper.

## 2 Phenactivism vs. Cognitivism

### 2.1 Conceptual Clarification

The goal of this section is to lay out the basic claims of phenactivism and cognitivism and to discuss their different approaches to general and social cognition. It is claimed that the two accounts can be seen as marking the endpoints of a theoretical spectrum, both of them providing valuable insights and assumptions about the nature of the human mind.

I will start with clarifying the concepts of "phenactivism" and "cognitivism". Both terms refer to specific accounts of cognition that hold a distinct set of metaphysical, methodological, and epistemological background assumptions. Most readers will be more familiar with cognitivist views on the mind, since these have not only been the prevalent accounts since the rise of cognitive science but still continue to form the theoretical background of most researchers in the field. My description of cognitivism will thus be rather short and my focus will be on disentangling the more obscure and less famous notion of what I call phenactivism. Useful definitions of each term are provided by De Bruin and Kästner (de Bruin et al. 2012, p. 542-543) and serve to give a first idea of what they amount to:

> Classic Cognitivism (COG): The mind is basically an intracranial information processing system manipulating (sub-)symbolic representations; cognition essentially is this computational process.
>
> Enactive Cognition (ENAC): Rather than a representational process, cognition is a process of sense-making that emerges from the dynamic online interaction or 'coupling' between autonomous agents and the environment in which they are embedded.

In other words, while cognitivism describes the human mind as a computational device that can be found exclusively inside the skull of an individual and operates on representations, (ph)enactivism claims quite the opposite; the mind is neither inside nor outside the individual but instead emerges within the *relation* of agent and environment. In the following, I will unpack each term further.

### 2.2 Cognitivism

Classic or radical cognitivism has been described above as viewing the mind as an entirely internal device that operates on representations in a specific, symbolic format. The notion of representations is as central as the claim that cognition is skull-bound and computational. This theory is the metaphysical and methodological background for so-called mindreading theories of social cognition. These theories are traditionally simulation-theory and theory-theory and in principle describe internal processes of a certain kind that underlie the inference of mental states of others. In their most extreme formulations, these theories state that social cognitive skills are mindreading skills and thereby draw a fundamentally individualistic, internalist, and representationalist picture of the phenomenon.

The role of social interaction and embodiment in radical cognitivist views is quickly explained. Mindreading theories have paid little attention to social interaction and embodiment and how these factors could influence, change, or even constitute social cognition. However, it should be noted that they are not obliged to deny that both are important factors for social understanding (Overgaard and Michael 2013). This is especially obvious from the experimental paradigms that are used to investigate mindreading skills. Typically, social stimuli consist of the picture of another person or a video of this person executing a specific action (e.g., Iacoboni et al. 2005; Wicker et al. 2003). While this kind of experimental design is well controlled, it lacks ecological validity since it situates participants in rather unrealistic situations.

Taken together, radical cognitivist theories foster a rather inflexible view on the mind as an input-output device. This view is then transferred to the social realm, drawing a picture of social cognition that ignores the fact that social encounters involve embodied agents that engage in interactions with each other.

## 2.3 Phenactivism

The shortcomings of cognitivism discussed above motivated phenactivists to find an alternative perspective that considers not just the brain in the skull but also the organism in the environment. Phenactivism describes the mind as *relational*, as emerging in the interaction of agent and environment. In the early 1990s, Varela , Thompson and Rosch (Varela et al. 1993) published their book *The Embodied Mind* in which they aimed to provide a non-cognitivist, alternative model of the mind. Their motivation was to criticize the view that describes mental processes as computations and the manipulation of representations. Such a model is said to be unsatisfactory, since it lacks a pragmatic approach to cognition and fails to integrate an inherent connection between mind and life (cf. Thompson 2010, p. 12).

A radical cognitivist picture of the mind depicts mental processing as fully internal and will thus not attribute any decisive role to the body. Phenactivists, however, adopt a rejection of the distinction between inner and outer, claiming that the first mistake to make in thinking of cognition is to assume that it has a location which is found either inside or outside the skull (Arnau et al. 2014). This point is most crucial for understanding the difference between phenactivism and cognitivism. Even tamer versions of cognitivism, which state that the mind can be extended to brain-external structures, are still fundamentally different from phenactive views.

While cognitivism places epistemic mechanisms within the skull and attributes a mere input-role to the external world and an output-role to action, phenactivism ties in both of these elements into the epistemic process. This is captured by the central notion of *sense-making;* within their embodied activity, agents not only actively regulate their coupling with the environment, they thereby establish a perspective onto the world (cf. De Jaegher and Di Paolo 2007 p. 488). Agents thus *create* meaning, there is no passive reception of information which is processed into or in virtue of internal representations which then (potentially) bear meaningful content.

The centrality of interaction is the core assumption of phenactive accounts and builds the starting point for further claims. Social interactions are seen as providing enabling conditions and forming constitutive elements for both the development and maintenance of social skills (De Jaegher and Di Paolo 2007; De Jaegher et al. 2010; Di Paolo and De Jaegher 2012). In order to expound this view, the claim is couched in theoretical terms of general phenactivism. Empirical set-ups, such as the perceptual crossing paradigm (Auvray et al. 2009) are assumed to corroborate these theoretical aims.

At this point it will be helpful to look at how proponents of the theory conceive of interaction. Here is a definition that is now generally accepted:

> Social interaction is the regulated coupling between at least two autonomous agents, where the regulation is aimed at aspects of the coupling itself so that it constitutes an emergent autonomous organization in the domain of relational dynamics, without destroying in the process the autonomy of the agents involved (though the latter's scope can be augmented or reduced). (De Jaegher and Di Paolo 2007, p. 493)

In other words, interactions are viewed as building autonomous systems which then are irreducible to local mechanisms physically realized within the individuals involved. Two systems are furthermore said to be coupled when their behavior and mental states depend on each other.

The concept of 'participatory sense-making' was introduced to capture these ideas. De Jaegher and colleagues (De Jaegher and Di Paolo 2007, p. 497) define the term as "the coordination of intentional activity in interaction, whereby individual sense-making processes are affected and new domains of social sense-making can be generated that were not available to each individual on her own." Together with the definition of interaction given above, this means that individuals 'merge' into one interactive, autonomous system. Since sense-making can be seen as the phenactive term for cognition (Thompson 2010), these claims boil down to the statement that interacting individuals, mutually and in virtue of

the emergent interaction dynamics, constitute (at least part of) their social cognitive processes. Social cognition as participatory sense-making then exhibits a *relational* kind of cognition. It is not to be located in either individual's head, brain or even body, but *in between* interacting individuals.

In sum, phenactive views on (social) cognition draw a radically different picture than cognitivist theories and come with radically different premises. This is problematic for several reasons, which I will detail in what follows.

## 2.4 Problems with Phenactivism and Cognitivism

Both phenactivism and cognitivism — in their radical formulations — are ill-suited for providing a comprehensive account on social cognition. The problems that come with a radical cognitivist view on social cognition are rather obvious and mostly refer to the fact that they exclude the importance of interaction and embodiment. While they are good at accounting for high-level phenomena such as explicitly thinking about the causes of another person's behavior, it is mostly ignored that interactions form a context that could change and influence social cognitive processing quite profoundly. If the goal is to find a comprehensive theory of social cognition, a theory that excludes the role of embodied interaction thus is undesirable.

What about the alternative at the other end of the theoretical spectrum? Obviously, phenactivism attributes quite some weight to interaction and embodiment. There is, though, the question of how well their claims are backed up, both conceptually and empirically. In what follows, I will discuss the conceptual and empirical validity of phenactive accounts and conclude that there are many incoherences and uncertainties which leave them unfit to offer a sound theoretical background for social cognition.

To begin with, it appears that phenactivism confuses enabling and constitutive conditions, leaving the phenactivist's claims unclear. A first hint of confusion is found when looking at the taxonomy of possible roles of interaction for a social cognitive process X that De Jaegher and colleagues (De Jaegher et al. 2010, p. 443) have worked out:

> Accordingly, given X, and a particular situation in which X occurs: F is a contextual factor if variations in F produce variations in X, C is an enabling condition if the absence of C prevents X from occurring and P is a constitutive element if P is part of the processes that produce X.

As Herschbach (Herschbach 2012) points out, however, it is rather unclear what exactly De Jaegher and colleagues judge to be a constitutive element. For additionally to the characterization given above, they also refer to it as a *part of the phenomenon itself*:

> A constitutive element is part of the phenomenon (it must be present in the same time frame as the phenomenon). The set of all the constitutive elements is the phenomenon itself. The presence of these elements is necessary, and therefore also enabling. (De Jaegher et al. 2010, p. 443)

This ambiguity leaves us with two possibilities in which interaction can constitute social cognition: (1) it can either be among those processes that *produce* the phenomenon, but does not have to be a *part of the phenomenon* (e.g., through interacting with her mother, the child learns to 'read' emotions and can later use this skill outside of interactions when she merely thinks about her mother), or (2) interaction constitutes social cognition in the sense that it must be present at the same time as the phenomenon and is a *necessary part* of it (e.g., only when the child interacts with her mother she can 'read' her emotions).

Claim (1) describes a condition that should count as causally enabling, not constitutive. The idea seems to be that interaction *enables* a particular mechanism to *arise* in that it was present as a necessary part of the development of that skill, and therefore it should be called constitutive. This confuses

the concepts profoundly and boils down to the assertion that interaction is an enabling condition and not that it constitutes a phenomenon in the sense that, metaphysically, it is a necessary part of it without which it would not exist. Moreover, the view that being immersed in social interactions — especially from a developmental perspective — enables particular social cognitive skills can in principle be accounted for by any non-phenactive theory that assigns a sufficiently strong role to extra-individual and situational contexts. To see this, consider that human newborns are completely helpless without a caregiver for an extraordinarily long time. Additionally, given some rather anecdotal evidence of children that lacked interactive and emotional engagement in early development and had severe mental as well as bodily impairments (e.g., Zimmer 1989; Bick et al. 2015; Fox et al. 2011), the fact that these contexts play a necessary role for social cognition seems almost trivial. It is questionable whether any theory would reject the assumption that interactive contexts play an enabling role for social cognition.

Further it should be noted that just because something is present in the same time frame as the phenomenon under scrutiny it obviously does not mean that it is part of the phenomenon. However, this is how one could read the quotation above. We can thus draw a first conclusion, stating that phenactive views lack a solid conceptual taxonomy to back up their strong claims in that it is unclear how they identify and separate sets of enabling and constitutive conditions for social cognition. The consequence is that they are left with statements that non-phenactive theories can account for as well.

What is the state of empirical evidence for the claim that interaction constitutes social cognition? Auvray et al's (Auvray et al. 2009) perceptual crossing paradigm is taken as providing an empirical ground for the phenactive position on interaction and as such picks up the idea that there might be something inherent in the interaction dynamics that is irreducible to individual mechanisms. In the experiment, two individuals were blindfolded and had to move their mouse cursor along a line. There were three objects that they could encounter on this line; a fixed object, the avatar of the other person, and the shadow of the other's avatar (Figure 1) Whenever they encountered an object, they would receive tactile feedback. Their task was then to click whenever they thought to have encountered the other's avatar. The results of the study show two things. First, participants were clearly able to distinguish between a fixed and moving object. Secondly, they appeared to favor avatar-avatar encounters, which was obvious from the higher number of these meetings.

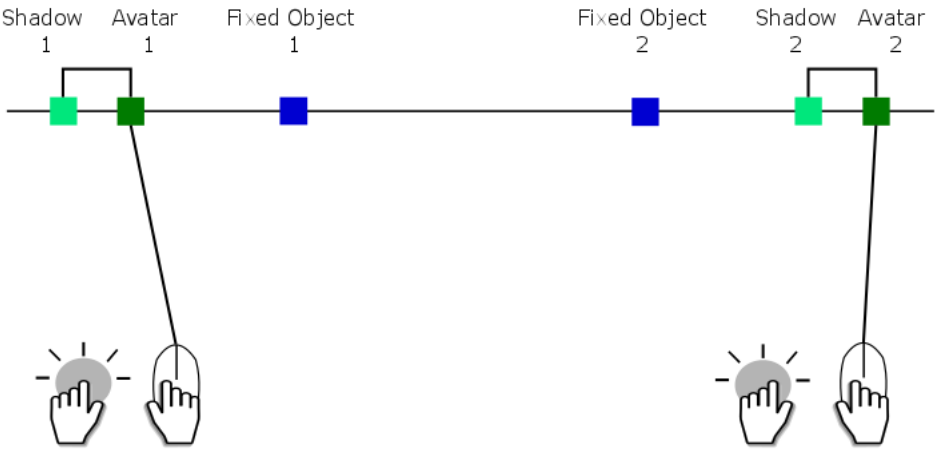


**Figure 1:** The perceptual crossing paradigm: In the perceptual crossing paradigm, two participants control an avatar (dark green box) that they can move with a computer mouse along a one-dimensional line with their right hand. The left hand rests on a buzzer, which provides the participants with tactile feedback when their avatar encounters an object in this one-dimensional space. Attached to their avatar is a mobile lure, or "shadow" (light green box), which follows the avatar at a constant, fixed distance. Additionally, there is a fixed, immobile object (blue box) on the line. Participants receive tactile feedback when their avatars encounter the other participant's avatar, their shadow, or the fixed object. The participant's task is to determine when avatars meet; that is, to tell when one participant's avatar encounters the other participant's avatar.

The second finding is said not to be explainable in individual terms and thus to require a non-reductive explanation at the level of collective dynamics. It was found that participants reversed their direction of movement after encountering any object, but only when both avatars meet, both receive a tactile feedback. The result is that, according to the authors, "this co-dependence of the two perceptual activities forms a relatively stable dynamic configuration." (Auvray and Rohde 2012, p. 3) The fact that an avatar-shadow encounter elicits feedback in only one subject is seen as not allowing the emergence of a stable interaction pattern.

There are, however, ways to interpret the results without referring to interaction dynamics as an emergent macro-structure whose properties substitute a part of individual mechanisms. There are basically three distinct conditions either individual can be in during the task; they can encounter the other person's avatar, this avatar's shadow, or the fixed object. Each situation differs with respect to the type of encounter. Importantly, individuals exhibit different behavioral patterns following each encounter. Thus each situation is indeed different, but *in virtue of* the behavior of either individual. It is therefore possible that individuals simply pick up subtle cues in the change of behavior of the other avatar, particularly because the situation in which both participants receive a tactile feedback elicits a different kind of reaction than the other situations.

Further, Froese and colleagues (cf. Froese et al. 2014, p. 8) claim that the results of the paradigm speak for an extendable mind that outsources parts of the cognitive work into the environment. This interpretation is, however, compatible with an extended, yet non-phenactive theory. The same holds for the claim that interaction dynamics influence the cognitive process. The ability to discriminate moving from fixed objects can easily be explained by perceptual learning, the ability to pick up statistical regularities from the environment. Taken together, it appears that the perceptual crossing paradigm does not yield evidence that unequivocally speaks for the hypothesis that interaction dynamics constitute part of the social cognitive processes that are needed to solve the task.

The conceptual and empirical uncertainties presented above should therefore leave us reluctant to adopt a radical phenactive view and strive to find a less controversial theoretical framework.

## 2.5 If Radicalism Is the Problem, Action-Oriented Predictive Processing Is the Solution

The main problem with radical cognitivism and phenactivism is that they exclude important aspects of social cognition, leaving their depiction of the phenomenon incomplete. While cognitivism does not take into account the importance of interaction and embodiment, it is questionable how phenactive views account for 'representation-hungry' elements of social understanding, such as explicit 'offline' reasoning about another person. In what follows, I argue that neither radical view can yield a comprehensive view on (social) cognition and that a middle-way is needed.

I previously presented De Bruin and Kästner's (de Bruin et al. 2012) definitions of cognitivism and (ph)enactivism. They examined which of these theories provide the most comprehensive view of cognition and come to the following conclusion:

> To conclude our diagnosis: neither COG [classic cognitivism] nor ENAC [enactive cognition] has been successful in providing a convincing account of both online and offline forms of cognitive processing. It hence seems fruitful to aim at a unified theoretical framework that solves the stalemate between ENAC and COG and integrates online and offline processes into a coherent story of how cognition can best be understood. (de Bruin et al. 2012, p. 547)

What the authors express in this quotation is twofold. First, it shows that there is a *spectrum* of theoretical claims, whose ends appear to be classic cognitivism and (ph)enactivism. Secondly, they rightfully gather that either account is yet to come up with a comprehensive and coherent account of cognition.

The same can be argued for the more specific case of social cognition. Phenactive views have indeed brought to awareness important aspects of the phenomenon that were previously ignored. This mainly concerns the aspects of embodiment, interaction, and the experiential quality of social encounters. Although they have been brought up by traditional phenomenology, they indeed got lost when the philosophical debate focused on cognitivist mind reading schemes. I thus agree with proponents of the phenactive view that a narrow view on the observational inference of mental states does not reflect the manifold nature of social cognition. On the other hand, it is questionable whether phenactive theories are able to capture the whole picture of social cognition. Although they might yield ways to grasp interaction, embodiment, and phenomenology, it is unclear how they would account for other aspects of the phenomenon, such as offline construction of reasons for another person's behavior.[3]

If the goal is to provide a *comprehensive* theoretical framework for social understanding that includes — among others — interaction, it is undesirable to adopt any radical position. As matters stand now, it seems that both cognitivist and phenactive theories have contributed valuable insights to the debate. It could be that some social processes need a rather non-representational, non-computational view, while others require a more cognitivist picture. I therefore argue that we should preserve a middle course and try to prevent any extreme, radical position that potentially excludes important aspects of the phenomenon. It would be advisable to attempt to find a theoretical framework that is able to integrate the full spectrum of social mechanisms.[4] In what follows I suggest that PP yields just the right ideas to do so.[5] In doing so, I will draw on three notions that are central to PP and, or so I shall argue, open up the possibility to combine cognitivist and phenactivist theoretical elements. These three notions are embodied inference (Friston 2012), action-oriented predictions (Clark 2016), and active inference (Friston et al. 2011). I will elaborate on these concepts in the next section, before applying them to the phenomenon of social cognition.

## 3 Action-Oriented Predictive Processing

### 3.1 Embodied Inference

The notion of 'embodied inference' was introduced by Friston and Stephan (Friston and Stephan 2007) to express how PP is an instance of the free-energy principle (FEP). To see what this means, we first have to consider the situation an embodied organism is embedded in. According to the second law of thermodynamics, the entropy of a closed system increases with time. Biological systems, however, are considered open systems in that they exchange energy and matter with their environment. They thusly resist the second law of thermodynamics and sustain their order. How is that achieved? Friston and Stephan (Friston and Stephan 2007, p. 422) suggest that the "premise here is that the environment unfolds in a thermodynamically structured and lawful way and biological systems embed these laws into their anatomy." In this sense, we can talk about embodied systems as *being* models of the environment they live in (cf. Friston 2012, pp. 89–90), instead of talking about systems that *have* or *build* models of the world. This is what Friston (Friston 2012) calls "embodied inference". More specifically, this term expresses that the physiology of a system already presupposes the circumstances it lives in — an organism's phenotype determines its possible state space.

3 Please note that thus far, the discussion between cognitivism and phenactivism is of an almost fully theoretical nature. This is partly due to the fact that most empirical designs available are based upon cognitivist assumptions and that thus far, phenactivists have only introduced few empirical designs. Thanks to an anonymous reviewer for raising this point.

4 This is not to say that one should or could simply combine phenactivism and cognitivism. Due to some metaphysical incompatibilities, a straightforward combination of the two does not come easy. For a detailed discussion of this matter, see Quadt 2015.

5 One possible concern at this point is that PP is built on the conviction that cognition is computation. Since most phenactivists reject such a computational view of the mind, this could lead to a rejection of PP by proponents of phenactivism. There are two ways to tackle this worry. First it should be noted that the aim of this paper is to create a *new* position that merely integrates ideas from each side of the theoretical spectrum, but does not aspire to be fully compatible with both. On the other hand, it could be claimed that phenactivism is not obliged to reject computationalism — a topic that needs to be pursued elsewhere. Thanks to an anonymous reviewer for raising this concern.

## 3.2 Action-Oriented Predictions

The topic of representations is one of the most controversial in the debate between cognitivism and phenactivism, which is why the notion of action-oriented predictions is of such high importance. While it is almost impossible to imagine cognitivism without the concept of representations, phenactivism rejects it entirely. It will thus be vital for our goal of finding a middle-way to alleviate the tensions revolving around this topic. One compelling solution has been proposed by Clark, who started to lay out the concept of 'action-oriented representations' in his earlier work (Clark 1997) and continued to draw a picture of representations that defies the 'old-school' version of cognitivism. The concept of representation in radical cognitivism refers to (sub-)symbolic vehicles that carry a specific content and in this sense are thought to 'mirror' the external world. Clark offers several arguments in favor of the idea that the kind of representation that PP yields is in no way related to the stiff, passive-mirror-of-nature representation old-fashioned cognitive science talked about.

First, although internal models are a central part of PP, these models are fundamentally grounded in embodiment, in that they "allow a system to combine a real sensorimotor grip on dealing with its world with the emergence of higher-level abstractions that (crucially) develop in tandem with that grip." (Clark 2014, p. 242) Representations or internal models are not marooned from brain-external matter, they are *for* engaging the body and world, to elicit action and active navigation of the environment. At the same time, the concept of representations is not given up. To see how representations in this context are defined, allow me to cite Clark's idea at length:

> […] each PP level (perhaps these correspond to cortical columns — this is an open question) treats activity at the level below as if it were sensory data, and learns compressed methods to predict those unfolding patterns. This results in a very natural extraction of nested structure in the causes of the input signal, as different levels are progressively exposed to different re-codings, and re-re-codings of the original sensory information. These re-recodings […] enable us, as agents, to lock us onto wordly causes that are ever more recondite, capturing regularities visible only in patterns spread over space and time. Patterns such as weather fronts, persons, elections, marriages, promises, and soccer games. […] What locks the agent on to these familiar patterns is, however, the whole mutli-level processing device (sometimes, it is the whole machine in action). That machine works (if PP is correct) because each level is driven to try to find a compressed way to predict activity at the level below, all the way out to the sensory peripheries. These nested compressions, discovered and annealed in the furnace of action, are what I [...] would like to call "internal representations. (Clark 2015a, p. 5)

As I read Clark, the essence of his claim is that representations are abstractions of sensory signals. They are not the sensory data themselves, but carry information that has been compressed and abstracted, enabling a prediction of what the "sensory" data a level below could be. In this sense, it is useful to talk about internal models and representations. Predictions represent potential sensory input, becoming more and more abstract as one goes up the hierarchy.

These kinds of representations do not merely generate a picture of the world in our heads. If the central role active inference plays in FEP is taken seriously, representations engage the *whole agent* to extract hidden causes in the world. In this sense, Clark opts for talking about 'action-oriented'-predictions: "They will represent how things are in a way that, once suitably modulated by the precision-weighting of prediction error, also prescribes (in virtue of the flows of sensation they predict) how to act and respond." (Clark 2016, p. 133). Considering the role of internal models—to prepare systems to act upon their environment and enable them to do so—thus helps us tune the notion of representation towards a more embodied, flexible one. This is, in my view, a crucial step towards

finding the 'golden middle' between cognitivist and enactive theories. The notion of action-oriented predictions will also be of high importance for what I call 'interactive inference'.

### 3.3 Active Inference

Clark's interpretation of Friston's take on FEP entails that organisms strive to reduce free energy by opting for the most efficient way to do so. Efficiency, here, refers to finding the strategy that involves the least complex route towards prediction error minimization, while bringing the largest effect. The brain's task thus not only entails the construction of inner models, but also preparing an organism for its bodily exchange with the environment. This involves the estimation of which channel and which 'strategy' will most efficiently minimize prediction error — will it be better to change my models (perception) or use my body to bring forth a change in the environment (action)? The latter strategy refers to what is called 'active inference'.

The body thus has an indispensable role in action-oriented PP. As described in the previous section, the trick is to acknowledge that the task of predictive models (i.e., representations) is to find the most efficient, least costly route to success. This is what Clark (Clark 2015a, p. 9) refers to when he talks about the "productive laziness" of the brain; whenever the body or the world can be recruited to do a job, there is no need to compute complex inner models. Precision-weighting determines whether low-level modalities or high-level modeling will 'be in charge' to solve the task at hand — depending on how efficient the strategy is estimated to be. This strategy will more often than not involve the engagement of brain-external structures:

> The task of the generative model […] is to capture the simplest approximation that will support the actions required to do the job — this means taking into account whatever work can be done by a creature's morphology, physical actions, and socio-technological surroundings. […] There is thus no conflict with work that stresses biological frugality, satisficing, or the ubiquity of simple but adequate solutions that make the most of brain, body, and world. (Clark 2015a, p. 291)

Clark here endorses a central aspect of phenactive theories, namely the role of extra-neural structures for an agent's navigation of its environment. Active inference takes center stage in this interpretation of PP, in virtue of the fact that the function of predictive models is to distribute the cognitive workload and recruit embodied action whenever possible.

Such a view emphasizes that PP displays a deep and fundamental connection of mind and body. This leaves us with the following picture of the (social) mind. PP accounts for quite a spectrum of phenomena; on the one hand, it is a rather brain-bound view, since the generation of predictions and the precision weighting process is neurally implemented. In that way, perception is brought forth mainly by top-down processing and is determined internally. This side of PP neatly accommodates 'representation-hungry' processes like imagination, dreaming and also thinking about other people, which seem to occur without much brain-external help. To see this, consider that it is argued that the main task of the cortex is to generate predictions about incoming stimuli (Friston et al. 2012). This means, basically, that the brain is able to reconstruct "the sensory signal using knowledge about interacting causes in the world" (Clark 2016, p. 85). Once learned, the system will be able to process without actual input and thus bring forth imagination, dreams, or explicit theorizing.

On the other hand, even those more 'decoupled' phenomena have been shown to involve the body. Saccadic eye-movements, for example, may be the bodily 'grounds' for phenomenal experience in dream states (Metzinger 2014). In this sense, the body and environment are indispensable parts of cognition. It is this neat interplay of internal models, action, and the body that make PP the perfect fit for a theory that integrates both phenactive and cognitivist elements, providing a sound ground for a theory on social cognition.

# 4 Conceptual Tools

## 4.1 Embodied Social Inference (EmSI)

The first concept I wish to introduce is 'embodied social inference' (EmSI), which emphasizes that the physiology of an organism constrains the kinds of social interactions it can engage in. Recall that embodied inference means that the thermodynamical laws of an agent's environment are 'folded into' her morphology; that her body is built to keep her alive by resisting the second law of thermodynamics. In this sense, it can be said that the agent is a model of its world, because their physiology incorporates the physical laws the body needs to obey to ensure survival. This is related to the claim that the physiology of an organism constrains the kind of mind it has, because the laws that are relevant for this specific phenotype will be modeled by its body.

In the same way, it can be said that the kind of body an organism has determines the kind of social interaction and understanding it is capable of. While a herring strives to stay in its large fish school to ensure its survival, cats aim for much smaller groups or may even survive on their own. The human body needs a caretaker for an extended amount of time during childhood, not being able to sustain itself until a certain age. Further, while humans are able to use their speech-apparatus to communicate and interact, ants will have to rely on pheromones to send signals to each other. This can be seen as embodied social inference (EmSI); an organism's phenotype determines the kind of social abilities they possess. To be more specific, an embodied organism can be called a model of its social environment because their physiology incorporates possibilities for interaction; vocal cords make vocal communication possible, for example.

This is also important when discussing the role of similarity for social cognition. While there are very many individual differences, the gross anatomy and morphology of individual organisms of one species is rather similar. This similarity may provide a fundamental role in the attempt to recognize the other as 'one of us' and thus to understand them. The role of bodily similarity is twofold; it not only determines how well we understand another person, but it also opens up the possibility that there needs to be a general similarity for social processing to begin with. The claim that a certain degree of similarity is needed in order to understand each other has been famously formulated by a number of researchers. For example, Meltzoff (e.g., Meltzoff 2005; Meltzoff 2007; Meltzoff 2013) states in his 'like me' hypothesis that the development of understanding others hinges upon the fact that the infant perceives the other as 'like me'. In fact, it is claimed

> that the core sense of similarity to others is not the culmination of social development, but the precondition for it. Without this initial felt connection to others, human social cognition would not take the distinctively human form that it does. (Meltzoff 2013, p. 139)

Meltzoff's reasoning rests on the assumption that social cognition — especially in developmental terms — is enabled by matching visual to motor representations. The bedrock of his argument rests on many neonatal imitation studies by him and his colleagues (Meltzoff and Moore 1997). Although having no visual information about one's own face, newborn babies appear to be able to imitate an adult's behavior, such as tongue protrusion (Meltzoff and Moore 1977). It is thought that the visual information of the adult is 'matched' onto the proprioceptive information the newborn already acquired. This matching process then enables imitative behavior.

The 'like me' hypothesis gains additional support when viewed from a PP perspective. In accordance with a simulation model, Friston and Frith (Friston and Frith in press, p. 12) argue that "internal or generative models used to infer one's own behaviour can be deployed to infer the beliefs (e.g., intentions) of another — provided both parties have sufficiently similar generative models." In other words, similarity here is seen as a presupposition for mental state inference. Only when there is a suffi-

cient similarity of models, there can also be a big enough overlap which allows the application of one's own models to understand the other's behavior.[6]

This is important for several reasons. First, I claim that replicative interactive inference (RInI) largely draws on similarity. Secondly, similarity relates to the discussion of so-called 'shared representations'. What Friston and Frith refer to above is exactly this — models or representations that are sufficiently similar can be used for both self- and other-related processing. A famous example of social mechanisms that rely on shared representations is found in the mirror neuron system. Mirror neurons are known to fire not only when an individual executes, but also when she merely observes an action (e.g., Rizzolatti and Craighero 2004). They can thus be said to involve shared representations, because they function both for action execution (self-related) and action observation (other-related). Finally, action-oriented PP implies that representations, i.e., predictive models, are grounded in sensorimotor processes. The range of these processes, in turn, are constrained by the kind of body an organism has. As trivial as it seems, this basically means that bodies determine the range of (social) experiences one can have. Metzinger (Metzinger 2004[2003], pp. 160–161) picks up this point and formulates it as the 'single-embodiment constraint':

> Trivially, the causal interaction domain of physical beings is usually centered as well, because the sensors and effectors of such beings are usually concentrated within a certain region of physical space and are of limited reach. […] This functional constraint is so general and obvious that it is frequently ignored: in human beings, and in all conscious systems we currently know, sensory and motor systems are physically integrated within the body of a single organism. This singular "embodiment constraint" closely locates all our sensors and effectors in a very small region of physical space, simultaneously establishing dense causal coupling.

In making this statement, Metzinger clarifies that the behavioral space of an individual is limited and constrained by its body. The range of possible behavior and experiences shape our cognitive processing, an effect whose pervasiveness becomes clear when viewed through the lens of PP. PP depicts the neural and cognitive architecture as immensely flexible and ever-changing. If precision-weighting admits, any sensory signal can change predictions at any level of the processing hierarchy.

If it is furthermore true that anatomical as well as morphological features are the basis for a system's generative models, and if it is true that these models can only be used for both self- and other-related processing if they are sufficiently similar, it follows that the bodies of interacting individuals must be sufficiently similar, too. Put differently, if the bodily structure of individuals is grossly different, their models may not be sufficiently similar, thus restricting interaction and understanding. The relation to EmSI should be clear by now; the phenotype of an individual must exhibit some degree of similarity in order to make it possible to recognize others as 'like me' and thus to enable the matching of one's own models to the other's.

To sum up, EmSI refers to the determining and constricting role that bodies play for social cognition, and also for interaction. In this sense, it can be said that the very physiology of an individual determines its space of possible social interactions.

## 4.2 Interactive Inference

The notion of interactive inference is tightly related to active inference and can be described as the minimization of prediction error by engaging in an embodied interaction. Applying this line of thought to the realm of social cognition, I now wish to add that interaction can play the same role for social cognitive processing as action plays for general cognition. This amounts to gathering information about

6 While similarity is claimed to be crucial for social cognition, please note that it may not be necessary for all kinds of social understanding. Otherwise, we would not be able to understand that the dog's wagging tail is an expression of his excitement or that what the octopus is intending is to open the jar. Distinguishing between self and other, thus, is just as important as similarity. Thanks to an anonymous reviewer for raising this concern.

the social environment and in this way actively sculpting one's external and internal environment. I thus claim that just as *active inference* is central for general cognition, interactive inference (InI), as I will call the process, is as central for social cognition.

What exactly does 'interactive inference' mean? Recall that active inference can be described as minimizing prediction error in several ways, namely by actively changing an agent's inner and outer environment so to fulfill exteroceptive, proprioceptive and interoceptive predictions, and the disambiguation between competing predictive models (cf. Seth 2015, pp. 13–14). In a similar way, interactive inference can be described as minimizing prediction error while navigating the social environment. Instead of changing one's model about the other person in order to understand her (perceptual inference), InI serves to actively sample proof for predictions or to cancel out possible models about causes of the behavior or another person. The basic idea is that engaging in interactions with other people can be a means to minimize prediction error and thus offers a fast and fruitful way to understand others. In what follows, I will elaborate on the concept by further distinguishing two types of InI; replicative interactive inference (RInI) and complementary interactive inference (CInI). The distinction serves to distill and differentiate the manifold ways in which interaction can enrich and enable social cognitive processing.

### 4.3 Replicative Interactive Inference (RInI)

Turning towards two different types of InI, let us first consider what I will call replicative interactive inference (RInI). In RInI, the other's internal or bodily states are replicated, such as in mimicry, emotional contagion, or automatic imitation. This replication has two effects, both of which can be said to make prediction error minimization more efficient. First, it serves to make oneself more similar to the other; in other words, to 'put oneself into' the other's bodily state. Instead of generating brand new models about the other person and the possible causes of her behavior on the basis of exteroceptive social stimuli, it will be quicker to gather information by tuning into their bodily, i.e., interoceptive or proprioceptive, state. So far, we have discussed the role of similarity in terms of morphology. However, this referred to the basic possibility of understanding each other. RInI can now be said to enhance this similarity by replicating the other's *current* bodily state.

Secondly, RInI serves to give 'first-hand' information about the other person. In order to get a sense of the other, predictions about their current state are corrected in virtue of error signals. When replicating the other's bodily state, these error signals should be more reliable, since they come not only from one exteroceptive (e.g. visual) source, but also from an internal source (e.g., proprioceptive prediction error). Therefore, during RInI, the bodily state (e.g., posture, movements) of another person is mimicked in order to supplement exteroceptive information about them with interoceptive and proprioceptive information. Mimicry, synchronization and automatic imitation are instances of RInI that function to make predictions about the other more precise by increasing the number of signal sources that yield relevant information.

These phenomena occur automatically and involuntarily — even when people are explicitly asked to suppress these tendencies. There are, for example, many studies which show that individuals cannot help but synchronize their movements with the other person. This has been shown for several motor acts, such as finger tapping (Oullier et al. 2008), rocking in rocking chairs (Richardson et al. 2007) and body posture (Lafrance and Broadbent 1976). Chartrand and Lakin (Chartrand and Lakin 2013, p. 288) provide a comprehensive review on these effects and summarize them under the notion of 'the Chameleon effect': "[…] much like chameleons change their color to blend into their surrounding environment, humans alter their behavior to blend into their social environment." In a vast number of studies that are reviewed by the authors, it has been shown that mimicry and synchronization are accompanied by many facilitating factors and in turn also facilitate social interaction. For example, individuals are more likely to mimic another person when there are prior 'pro-social' factors, such as

in-group effects and prior rapport. Individuals with similar opinions and high empathy rates are more prone to mimicry and synchronization. Although there are also inhibitors of mimicry such as the wish to disaffiliate with the other person, the authors conclude that unconscious mimicry and synchronization seems to be a default for social interactions and occurs even when individuals face other tasks (cf. Chartrand and Lakin 2013, p. 290). Further, individuals that were told to keep still and suppress their tendency to replicate the other person's behavior perform worse at emotion detection tasks.

Furthermore, consider the following study conducted by Ainley and colleagues (Ainley et al. 2014) that links interoceptive awareness with the tendency to automatically imitate. They found that — contrary to their initial prediction — participants who scored higher for interoceptive awareness had a greater tendency to imitate. In other words, the more one is aware of one's interoceptive processing (in this study measured with the so-called 'heartbeat perception task'), the less one is able to inhibit automatic imitation. One possible (although rather speculative) interpretation of these results is that people with higher interoceptive awareness set the gain on interoceptive prediction errors higher. Ainley and colleagues (Ainley et al. 2014, p. 26) hypothesize that

> [g]iven that interoceptive awareness affects perception of the body, it is also likely to modulate action representations. It has recently been indicated that in order to avoid mirroring another person's actions it is essential to reduce the precision of proprioceptive prediction error (Friston, Mattout & Kilner, 2011). If people with high interoceptive awareness have initially precise proprioceptive prediction errors then their tendency to imitate others may be accounted for.

Put differently, in order to inhibit imitation and not to replicate the other's movement, gain on prediction error must be set low. Thus, weighting the precision of prediction errors high may result in the tendency to automatically imitate the other person. If this is correct, the processing steps underlying automatic imitation could be the following. First, contextual cues yield information that the current incoming signals originate from another person; thus representations about sensory consequences — which could be proprioceptive, interoceptive, or exteroceptive — are recruited. Next, depending on whether the gain on prediction error is set high or low, the observed state of the other person is replicated or not. As described above, highly precise errors would result in a replication of the other's state, while low-weighted prediction errors would result in the inhibition of automatic imitation.

This may not only be the case in motor imitation. Phenomena such as emotional contagion or the queasiness one feels when observing someone eating something truly disgusting could be cases in which gain on interoceptive prediction error is set high. This would lead to the replication of the other's interoceptive state and thus trigger 'shared bodily experiences'. Entering an actual interaction should provide all interacting individuals with more unambiguous cues to which predictive model has the highest posterior probability. To see this, recall that RInI serves to make the bodies of interacting individuals to be in more similar states. If it is true that higher-order representations are grounded in sensorimotor processes, this should also lead to a more similar representational state of the body model in both individuals.

Facial emotion recognition serves as another elaborative example of RInI. Several findings are of central importance here. First, it has been claimed that the face is likely the most significant body region for social cognition, since it provides the most relevant information when it comes to understanding others (cf. Farmer et al. 2014, p. 290). Secondly, a great number of studies have shown that the sight of emotional expressions leads to activation in brain areas with mirror properties (e.g., Wicker et al. 2003). Further, people tend to mimic facial expressions of their interaction partners (cf. Chartrand and Lakin 2013, p. 287). Above, I reviewed some of the research suggesting that mimicry not only occurs ubiquitously, but that it also has striking effects on social relationships. In turn, there is growing evidence that the tendency to mimic is considerably influenced by top-down effects and prior information about the other person (ibid.).

Putting these findings together, the following picture emerges: Visual signals of the other person's facial expression (plus contextual information) trigger generative models about the underlying emotional state — this is where shared representations enter the picture. These predictive models serve as a basis for generating proprioceptive predictions — that is, the motor commands underlying the facial expression — and also interoceptive predictions which refer to the internal bodily state the person must have been in to give rise to the emotion displayed on their face. Proprioceptive prediction error can be quashed by changing the state of facial muscles ourselves, thus mimicking the other person. Interoceptive prediction error can also be minimized by actively changing one's internal environment. Seth (Seth 2013) claims that emotions occur when prediction errors are cancelled out for exteroception, interoception and proprioception, thus disambiguating multimodal models generated in the insular cortex. The same may be true for emotion recognition; multimodal predictive models about the cause of incoming exteroceptive signals are confirmed or ruled out by quashing proprioceptive and interoceptive prediction error, inferring the most likely cause of the observed emotion. Mimicry, as an instance of RInI, is therefore a crucial and fast way to enhance this process of emotion recognition.

The rationale here is that greater bodily similarity will lead to greater social similarity and facilitate social understanding. Of course, whether or not interactive inference will be deemed a fruitful way to figure out the other person depends on prior beliefs and expectations about the other person. As already mentioned, top-down effects are pervasive and determine whether or not mimicry occurs. However, this fits nicely in the more general framework of PP, since the multidirectional interplay between bottom-up and top-down effects is of central importance.

### 4.4 Complementary Interactive Inference

The automatic replication of bodily and motor states is, of course, not the only process which happens between individuals during an interaction. Instead of replicating, it will often be necessary to perform complementary actions. This second case I will call complementary interactive inference (CInI). CInI refers to changing one's internal (i.e., bodily) or external environment in response to the other person without replicating the other's state. This has several functions.

First, it can serve to regulate another person's current bodily or emotional state. This can be achieved by changing one's own posture, movements, or gestures (e.g., giving the other an encouraging nod to make her continue talking), but also by altering one's interoceptive state. An intriguing example of this latter case can be found in so-called 'kangaroo care', which is often used for prematurely born (human) babies. During kangaroo care, mothers hold their infants in an upright position close to their body between their breasts and underneath their clothing. It has been found that this has many positive effects on both mother and baby. Most interestingly for the matter here are the physiological effects; mothers regulate their body temperature according to their infants needs and thereby also enhance self-regulation of the child. When the child has a fever, mothers lower their body temperature so to provide cooling for their infant. Further, if the baby has an irregular heartbeat, this can be counteracted and becomes more steady when their ear is placed on their mother's chest and they hear the mother's steady heartbeat (Ludington-Hoe et al. 2006; Nyqvist et al. 2010).

A second function of CInI could be to evoke behavioral response of the other person that serves as additional exteroceptive input in order to disambiguate social stimuli. Gestures, facial expressions or other movements are used to signal one's uncertainty and thus provoke a reaction of the other person, which then serves as additional information. I might, for example, shrug my shoulders or raise my eyebrows in order to signal you that I did not understand what you were saying. This signals to you, in turn, that I need additional information and may — if this interaction is successful — elaborate on your stance.

Thirdly, CInI serves to make oneself more predictable, thereby smoothing out social understanding, joint action, or coordination. For joint actions that require coordination, for example, Vesper and

colleagues (Vesper et al. 2010) have coined the term 'coordination smoothers' to describe the modulation of one's own behavior in order to make coordination with another person more simple:

> One way to facilitate coordination is for an agent to modify her own behavior in such a way as to make it easier for others to predict upcoming actions, for example by exaggerating her movements or by reducing the variability of her actions. (Vesper et al. 2010, p. 999)

In several studies it has been found that people indeed adjust their movement trajectories, their pace or use signaling or communicative actions in order to increase predictability. For example, piano players that are performing a duet exaggerate their finger movements or speed up in order to decrease variability (Keller et al. 2007).

What may be the mechanisms underlying all these functions of CInI? Vesper and colleagues claim that prediction and motor simulation are key to enabling the execution of complementary actions. Simulations are thought to be especially useful for joint actions, where they enhance timing and anticipation of sensory consequences. This comes naturally within a predictive processing framework, since it is assumed that predictive models represent sensory consequences of actions in a counterfactual manner (Seth 2014). Assuming that these predictive models can be shared — i.e., that they can be used for both self- and other-related processing — it becomes clear how they can be exploited to compute not only the consequences of one's own, but also the other's sensorimotor trajectory. Interestingly, the role of similarity becomes important one more time, for joint action is enhanced when the timing patterns of both agents are predictable. The predictability in turn is dependent upon how similar agents are, and how similar their motor experience is. This has been shown in several studies that show that mirror neuron activity increases when observing actions that already belong to one's own motor repertoire (Calvo-Merino et al. 2004). The mirror neuron system is therefore involved in both replicative action processing and the preparation of complementary actions. According to Pezzulo and colleagues (Pezzulo and Dindo 2011, p. 612), "this suggests that the brain can encode actions executed by others in an interaction-oriented way, and more broadly that action-perception mappings could be quite flexible and task-dependent."

Taken together it can thus be hypothesized that shared predictive models are not only useful for replicative, but also complementary interactive inference. Interaction is here used to solve problems with the other person, in virtue of making oneself more predictable, and using one's body to signal what is needed from the other. Framing interaction within PP allows to attribute an important role to interaction patterns between individuals to their social cognitive processing. The mutually unfolding predictions, actions, counteractions and perceptions are captioned by interactive inference and thus provide a new way to conceptually grasp how interactions matter for social cognition.

## 5 Conclusion

The current situation in the research field of social cognition has been depicted as problematic because the theoretical schemes of phenactivism and cognitivism alone do not yield a sound ground for a theoretical framework on the phenomenon. While the latter ignores the importance of embodied interaction, the former has been doubted to have sufficient conceptual and empirical back-up. At the same time, both theories account for important aspects of social cognition. The main goal of this paper was therefore to find a theoretical approach to combining these aspects, while circumventing the problems that come with phenactivism and cognitivism.

Action-oriented PP provides many opportunities for implementing both cognitivist and phenactivist elements in a theory on social cognition. Another aim in this paper was thus to exploit some of them and start to suggest ways in which PP can enlighten theoretical work on the phenomenon. Just like general cognition, social cognition heavily draws on the interaction of body, mind and world. PP

is therefore the perfect partner to highlight this dependency, since it appears that although a great part of the prediction error minimization machine is located in the brain, the body and action play an indispensable role for this mechanism. To see this, remember that while prediction generation clearly is the brain's job, the minimization of prediction error — the core of PP — heavily engages the body and world in virtue of active inference.

In this sense, it has been claimed that embodiment is fundamental to (social) cognition in at least two ways. First, the very morphology and phenotype of a system set the baseline of what are probable states for it to be in. To capture this idea for the social realm, the notion of embodied social inference (EmSI) has been introduced. EmSI expresses that our bodies define the kinds of social interactions we are able to engage in, and that a certain amount of morphological similarity is needed in order to enable social understanding. Second, as described above, active inference appears as a part of PP it cannot do without. This has consequences for our view on both general and social cognition. Concerning the latter, I coined the term of 'interactive inference' (InI) to describe replicative and complementary behavior that serves to cancel out prediction error via engagement in social interactions.

The perspective adopted in this paper has implications for future research. For example, it is asserted that differences in sensorimotor processes result in differences of predictive models, which can be shared and exploited for social cognition. From this, we can derive the prediction that large differences of sensorimotor processes between individuals will make social cognitive processes that rely on them more difficult. This has been shown for the case of autism. Cook (Cook 2016) argues that since the kinematics of movements in typical and autistic individuals deviate, they are less likely to resonate with each other. This may be one cause not only for the social impairments that come with autism, but also for the difficulties that typical individuals have in understanding autistic individuals. From the perspective of interactive inference, it can be assumed that processes of replication are disrupted, thus leading to an impaired inference process between individuals. It can be hypothesized that predictive models that are built on the basis of an individual's own motor repertoire are too different to support a stable inference mechanism. Future research should investigate at which level impairments occur and cause impaired social interactions between autistic and neurotypical individuals.

This also relates to the question of how individual differences influence social cognition. This question needs to be broken down into several sub-issues and more attention in future research. As described before, in the case of autism it has been hypothesized that differences in kinematic profiles between individuals with autism and typically developed individuals are one source of problems in social understanding (Cook 2016). It has further been shown that similar motor experience of individual enhances imitative behavior (Kilner et al. 2007). These findings can serve as a starting point to examine how important individual similarity and differences are for social cognitive processing. At the neural level, differences in precision weighting could influence the tendency to imitate. This would be predicted by the claim that precision optimization is a leading component in automatic imitation.

The considerations in this paper show that although PP puts forth a quite central role of the brain, it still integrates a deep sense of embodiment and relation with the environment in virtue of being an instance of FEP. As such, this theory displays a fundamental continuity of mind and life. Again, there lies a great opportunity to satisfy demands from phenactivism in taking this continuity seriously and explore its consequences for a theory of our social minds.

## References


Ainley, V., Brass, M. & Tsakiris, M. (2014). Heartfelt imitation: High interoceptive awareness is linked to greater automatic imitation. *Neuropsychologia, 60*, 21–28. https://dx.doi.org/10.1016/j.neuropsychologia.2014.05.010.

Arnau, E., Estany, A., González del Solar, R. & Sturm, T. (2014). The extended cognition thesis: Its significance for the philosophy of (cognitive) science. *Philosophical Psychology, 27* (1), 1–18. https://dx.doi.org/10.1080/09515089.2013.836081.

Auvray, M. & Rohde, M. (2012). Perceptual crossing: The simplest online paradigm. *Frontiers in Human Neuroscience, 6*. https://dx.doi.org/10.3389/fnhum.2012.00181.

Auvray, M., Lenay, C. & Stewart, J. (2009). Perceptual interactions in a minimalist virtual environment. *New Ideas in Psychology, 27* (1), 32–47. https://dx.doi.org/10.1016/j.newideapsych.2007.12.002.

Bick, J., Zhu, T., Stamoulis, C., Fox, N., Zeanah, C. H. & Nelson, C. A. (2015). Effect of early institutionalization and foster care on long-term white matter development: A randomized clinical trial. *JAMA Pediatrics*. https://dx.doi.org/10.1001/jamapediatrics.2014.3212.

Calvo-Merino, B., Glaser, D. E., Grézes, J., Passingham, R. E. & Haggard, P. (2004). Action observation and acquired motor skills: An fMRI study with expert dancers. *Cerebral Cortex, 15* (8), 1243–1249. https://dx.doi.org/10.1093/cercor/bhi007.

Chartrand, T. L. & Lakin, J. L. (2013). The antecedents and consequences of human behavioral mimicry. *Annual Review of Psychology, 64*, 285–308. https://dx.doi.org/10.1146/annurev-psych-113011-143754.

Clark, A. (1997). The dynamical challenge. *Cognitive Science, 21* (4), 462–481.

——— (2013). Whatever next? Predictive brains, siuated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–253. https://dx.doi.org/10.1017/S0140525X12000477.

———(2014). Perceiving as predicting. In D. Stokes, M. Mohan & S. Biggs (Eds.) *Perception and its modalities* (pp. 23–44). Oxford: Oxford University Press.

———(2015a). Embodied prediction. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570115.

———(2015b). Predicting peace: The end of the representation wars. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570979.

———(2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York, NY: Oxford University Press.

Cook, J. (2016). From movement kinematics to social cognition: The case of autism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 371* (1693). https://dx.doi.org/10.1098/rstb.2015.0372.

de Bruin, L., Van Elk, M. & Newen, A. (2012). Reconceptualizing second-person interaction. *Frontiers in Human Neuroscience, 6*. https://dx.doi.org/10.3389/fnhum.2012.00151.

De Jaegher, H. & Di Paolo, E. A. (2007). Participatory sense-making. *Phenomenology and the Cognitive Sciences* (6), 485–507. https://dx.doi.org/10.1007/s11097-007-9076-9.

De Jaegher, H., Di Paolo, E. A. & Gallagher, S. (2010). Can social interaction constitute social cognition? *Trends in Cognitive Sciences, 14* (10), 441–447. https://dx.doi.org/10.1016/j.tics.2010.06.009.

Di Paolo, E. A. & De Jaegher, H. (2012). The interactive brain hypothesis. *Frontiers in Human Neuroscience, 6*, 1–16. https://dx.doi.org/10.3389/fnhum.2012.00163.

Farmer, H., McKay, R. & Tsakiris, M. (2014). Trust in me: Trustworthy others are seen as more physically similar to the self. *Psychological Science, 25* (1), 290–292. https://dx.doi.org/10.1177/0956797613494852.

Fox, N., Almas, A. N., Degnan, K. A., Nelson, C. A. & Zeanah, C. H. (2011). The effects of severe psychosocial deprivation and foster care intervention on cognitive development at 8 years of age: Findings from the Bucharest Early Intervention Project. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 52* (9), 919–928. https://dx.doi.org/10.1111/j.1469-7610.2010.02355.x.

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

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

———(2012). Embodied inference: Or ''I think therefore I am, if I am what I think". In J. Kriz (Ed.) *The implications of embodiment: Cognition and communication* (pp. 89–125).

Friston, K. & Frith, C. (in press). A duet for one. *Consciousness and Cognition*. https://dx.doi.org/10.1016/j.concog.2014.12.003.

Friston, K. J. & Stephan, K. E. (2007). Free-energy and the brain. *Synthese, 159* (3), 417–458. https://dx.doi.org/10.1007/s11229-007-9237-y.

Friston, K., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics, 104* (1-2), 137–160. https://dx.doi.org/10.1007/s00422-011-0424-z.

Friston, K., Adams, R. A., Perrinet, L. & Breakspear, M. (2012). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3*, 151. https://dx.doi.org/10.3389/fpsyg.2012.00151.

Froese, T., Iizuka, H. & Ikegami, T. (2014). Embodied social interaction constitutes social cognition in pairs of humans: A minimalist virtual reality experiment. *Scientific Reports, 4*. https://dx.doi.org/10.1038/srep03672.

Gallagher, S. (2008). Direct perception in the intersubjective context. *Consciousness and Cognition, 17* (2), 535–543. https://dx.doi.org/10.1016/j.concog.2008.03.003.

Gallese, V. & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2* (12), 493–501.

Gopnik, A. & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language, 7* (1-2), 145–171. https://dx.doi.org/10.1111/j.1468-0017.1992.tb00202.x.

Herschbach, M. (2012). On the role of social interaction in social cognition: A mechanistic alternative to enactivism. *Phenomenology and the Cognitive Sciences, 11*, 467–486. https://dx.doi.org/10.1007/s11097-011-9209-z.

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

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C. & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology, 3* (3), 529–535. https://dx.doi.org/10.1371/journal.pbio.0030079.

Keller, P. E., Knoblich, G. & Repp, B. H. (2007). Pianists duet better when they play with themselves: On the possible role of action simulation in synchronization. *Consciousness and Cognition, 16* (1), 102–111. https://dx.doi.org/10.1016/j.concog.2005.12.004.

Kilner, J., Friston, K. & Frith, C. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing, 8* (3), 159–166. https://dx.doi.org/10.1007/s10339-007-0170-2.

Lafrance, M. & Broadbent, M. (1976). Group rapport: Posture sharing as a nonverbal indicator. *Group & Organization Management, 1* (3), 328–333. https://dx.doi.org/10.1177/105960117600100307.

Ludington-Hoe, S. M., Lewis, T., Morgan, K., Cong, X., Anderson, L. & Reese, S. (2006). Breast and infant temperatures with twins during shared kangaroo care. *Journal of Obstetric, Gynecologic, and Neonatal Nursing: JOGNN / NAACOG, 35* (2), 223–231. https://dx.doi.org/10.1111/j.1552-6909.2006.00024.x.

Meltzoff, A. N. (2005). Imitation and other minds: The "like me" hypothesis. In S. Hurley (Ed.) *Perspectives on imitation* (pp. 55–77). Cambridge, MA: MIT Press.

——— (2007). The 'like me' framework for recognizing and becoming an intentional agent. *Acta Psychologica, 124* (1), 26–43. https://dx.doi.org/10.1016/j.actpsy.2006.09.005.

——— (2013). Origins of social cognition. In M. R. Banaji & S. A. Gelman (Eds.) *Navigating the social world* (pp. 139–144). Oxford University Press. https://dx.doi.org/10.1093/acprof:oso/9780199890712.003.0025.

Meltzoff, A. N. & Moore, K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, 75–78.

——— (1997). Explaining facial imitation: A theoretical model. *Early Development and Parenting, 6*, 179–192.

Metzinger, T. (2004[2003]). *Being no one: The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

——— (2014). First-order embodiment, second-order embodiment, third-order embodiment: From spatiotemporal self-location to minimal selfhood. In R. Shapiro (Ed.) *The routledge handbook of embodied cognition* (pp. 272–286). Routledge.

Nyqvist, K. H., Anderson, G. C., Bergman, N., Cattaneo, A., Charpak, N., Davanzo, R., Ewald, U., Ibe, O., Ludington-Hoe, S., Mendoza, S., Pallás-Allonso, C., Ruiz Peláez, J. G., Sizun, J. & Widström, A.-M. (2010). Towards universal kangaroo mother care: Recommendations and report from the first European conference and seventh international workshop on kangaroo mother care. *Acta Paediatrica, 99* (6), 820–826. https://dx.doi.org/10.1111/j.1651-2227.2010.01787.x.

Oullier, O., de Guzman, G. C., Jantzen, K. J., Lagarde, J. & Kelso, J. A. S. (2008). Social coordination dynamics: Measuring human bonding. *Social Neuroscience, 3* (2), 178–192. https://dx.doi.org/10.1080/17470910701563392.

Overgaard, S. & Michael, J. (2013). The interactive turn in social cognition research: A critique. *Philosophical Psychology*, 1–25. https://dx.doi.org/10.1080/09515089.2013.827109.

Pezzulo, G. & Dindo, H. (2011). What should I do next? Using shared representations to solve interaction problems. *Experimental Brain Research, 211* (3-4), 613–630. https://dx.doi.org/10.1007/s00221-011-2712-1.

Quadt, L. (2015). Multiplicity needs coherence – Towards a unifying framework for social understanding. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND:* 26(C). Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958571112.

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R. L. & Schmidt, R. C. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science, 26* (6), 867–891. https://dx.doi.org/10.1016/j.humov.2007.07.002.

Rizzolatti, G. & Craighero, L. (2004). The mirror neuron system. *Annual Review of Neuroscience, 27* (1), 169–192. https://dx.doi.org/10.1146/annurev.neuro.27.070203.144230.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences, 17* (11), 565–573. https://dx.doi.org/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. https://dx.doi.org/10.1080/17588928.2013.877880.

—— (2015). The cybernetic Bayesian brain. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570108.

Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind.* Cambridge, Mass and London: Belknap.

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

Vesper, C., Butterfill, S., Knoblich, G. & Sebanz, N. (2010). A minimal architecture for joint action. *Neural networks: The Official Journal of the International Neural Network Society, 23* (8-9), 998–1003. https://dx.doi.org/10.1016/j.neunet.2010.06.002.

Wicker, B., Keysers, C., Plailly, J., Royet, J.-P., Gallese, V. & Rizzolatti, G. (2003). Both of us disgusted in my insula: The common neural basis of seeing and feeling disgust. *Neuron, 40* (3), 655–664. https://dx.doi.org/10.1016/S0896-6273(03)00679-2.

Zimmer, D. E. (1989). Wilde Kinder. In D. E. Zimmer (Ed.) *Experimente des Lebens* (pp. 21–47). Zürich: Haffmanns Verlag.

# The Problems with Prediction

## The Dark Room Problem and the Scope Dispute

Andrew Sims

There is a disagreement over the scope of explanation for predictive processing. While some proponents think that it is best motivated from—and indeed comprises an explanation of—biological self-organization, others maintain that it should only be a theory of neurocognitive function, or even just of some limited domain of neurocognitive function. Something that these theorists share is an interest in addressing the dark-room problem: at its most naïve, if action is driven by the minimization of surprise then why don't cognitive creatures act to minimize stimuli in general? The dark-room problem is in fact best conceived as a cluster of related concerns, rather than as a single argument against action-oriented predictive processing. These have to do with: i) whether PP (predictive processing) has any substantive empirical content when it is pitched in very general domains; ii) whether a specification can be given of the environmental niche that action moves the organism to occupy, and which is not the dark room; and iii) whether an adequate account can be given within this specification of exploratory and exploitative behaviours. There are interesting conceptual relations between the dark-room problem and the scope dispute. As the putative scope of predictive processing gets wider (culminating in the free energy principle), the resources that are available for answering the concerns about niche-specification become very rich. But increasingly puzzling problems arise as to the implementation of surprise-minimisation within non-paradigmatically cognitive biological systems. On the other hand, under more restrictive construals of the scope of predictive processing, there are new difficulties standing in the way of niche-specification, and new questions about the interface between surprise-minimisation and model-free cognition undermine the promise of predictive processing as a unifier of theories of neurocognitive function in subordinate domains. In this paper I make explicit the dialectic between proponents and critics in order to show how the two problems are related.

**Keywords**

Active inference | Aneural information processing | Dark-room problem | Explanatory scope | Free energy principle | Minimal cognition

### 1 Predictive Processing and the Scope Dispute

Predictive processing (PP) is the name for a class of theories in cognitive neuroscience which present the prospect of unifying various accounts of perception, action, and very many other ordinary and pathological cognitive phenomena. Roughly, the central idea is that both perception and action can be explained in terms of a mechanism whose sole function is the minimisation of surprise. There are an astounding range of ordinary and pathological cognitive phenomena that have possible explanations expressed in the theoretical vocabulary of PP, and the number of these continues to grow.

Surprise—often also called surprisal or self-information to distinguish it from phenomenological surprise (Clark 2013a)—is a quantity from information theory that describes how likely some event or set of events is, given some model that assigns probabilities over events. To make the distinction between information-theoretic and phenomenological surprise concrete, Clark gives the example of an elephant being smuggled onstage during a magic show. Given the context, this is not very surprising in an information-theoretic sense. Such things may be expected to happen in the context of a magic show. However, we may expect the audience to experience phenomenological surprise. In its broadest and least controversial formulation, PP states that some neurocognitive functions can be explained

Sims, A. (2017). The Problems with Prediction - The Dark Room Problem and the Scope Dispute.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 23. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573246

in terms of the minimisation of information-theoretic surprise. In stronger formulations of PP the explanatory scope of surprise minimisation becomes wider, from the extension of the mechanism to action (Friston 2009) all the way up to its application in understanding adaptive behaviour in general (Friston 2013). The details of PP will have already been discussed at length by the other contributors to this volume, and so I will not belabour the details except to remind the reader of its key features.

Consider then the problem of under-determination for visual perception. One pattern of retinal stimulation (a set of sensory states) is compatible with very many interpretations of its causes due to ambiguities that arise from distance, occlusion, and noise; the task in visual perception is to construct the best possible interpretation on the basis of that pattern. This means producing a visual scene on the basis of the retinal stimulation. One way to eliminate ambiguity is to interpret the sensory states on the basis of a model of the causes of those states. That resolves ambiguities by discounting interpretations of the states that are less likely, given that model. Thus, to give an informal example, we are able to perceive a particular pattern of sensory states as an old or young woman rather than as a truly ambiguous figure (Figure 1).

**Figure 1:** The "young woman/old woman" ambiguous figure (Boring 1930).

PP posits a computational architecture that is capable of successfully implementing a species of this basic top-down strategy (Rao and Ballard 1999). Imagine a model of the causes of sensory states that is *hierarchical*, and for which each level in the hierarchy is predicting the states of the level below it (it is *generative*, since it generates predictions of the sensory states it should encounter if the model is true). So at each level there is a comparison between the actual states which are propagated upwards from lower levels and laterally, and predicted states that are propagated downwards from higher levels and laterally. These can either match or not. If they do not match, then the actual state is propagated upwards as a *prediction error* signal, and the parameters at the higher level are updated on the basis of that prediction error in line with the norms of Bayesian belief-updating. So the higher-up parameters that generate predictions are treated as prior beliefs (in the form of probability distributions) about the way that states at the lower level will behave, and they are updated where they fail to predict those states accurately.[1] This means that the mechanism as a whole can slowly approach being an accurate model of the causes of sensory states, and that it does so by the minimisation of prediction error, which is an upper bound on surprise. PP states that neurocognitive systems implement a hierarchical generative model of this kind, with sensory states at the lowest level in the hierarchy, and that the model of the causes of those states is the content of perception—in this context, the visual scene.

1 "Prior belief" is a term of art here, however. It refers to whatever plays the functional role of top-down prediction in the hierarchical generative model. It's not to be taken as indicating the existence of a propositional attitude.

More expansive versions of this broad kind of proposal make room for multiple perceptual modalities in the hierarchy and include suggestions that it extends upwards into more abstract and amodal predictions. On this hypothesis we posit a more abstract hierarchical model about the causes of sensory states that helps inform perception in multiple modalities. For example, my interpretation of some set of sensory states in audition may be disambiguated by more general beliefs about the context of the states—I may interpret those states as being caused by the clanging of pots and pans on the basis of a prior belief that the origin of those states is from behind a door which I can see is clearly marked "KITCHEN," and on the basis of prior beliefs to do with kitchens more generally. Although it is not uncontroversial as to what each of the levels represents (Vance 2015), one way to make sense of the notion is that each level in the ascending hierarchy represents the world at an ever increasing level of spatiotemporal generality (e.g., Hohwy 2013, pp. 28-30).[2]

For a mechanism like this to work effectively, it needs to be able to distinguish between signal and noise. What that means is that it needs to evaluate how likely some datum is to be genuinely indicative of a causal regularity in the world. A hypothetical example of a situation in which a datum is *not* genuinely informative in the right kind of way can be given in the context of the measurement of population growth. If I am trying to measure long-term growth in tourism in Rio de Janeiro during the Olympics, for example, I will get a reading that is not genuinely informative about tourism growth in that city over a longer period of time. It is a noisy datum. Updating a Bayesian model on the basis of noisy data leads to overfitting; an overfitted model is greatly reduced in its predictive power because it contains redundant parameters and does not easily generalise to new data. So an effective Bayesian model needs to modulate its updating on the basis of the reliability of error, that is, how likely an error is to be genuinely informative about a causal regularity. In PP, this function is performed by the precision weighting of prediction error. A precise prediction error—a prediction error which has been directly assigned a high precision value—is considered very reliable, and so the model is updated on the basis of that error without much attenuation. But imprecise prediction error is treated as noisy and unreliable, and is therefore more likely to be attenuated if any updating occurs at all. The model therefore also needs to maintain a higher-order model of precision, in order that imprecise error can be treated as such; predictions are treated as more reliable than prediction error under conditions of sensory uncertainty.

It's possible to explain action according to this same basic computational architecture. Whereas perception is thought to function through an alteration of prior beliefs on the basis of prediction error, action is construed as the minimisation of prediction error through an alteration of the way that sensory states are sampled. That is to say that the world is sampled such that some set of predictions come out true (but this is not merely a corroborative but also a disambiguating process, as in the case of the visual saccade, for example, Friston et al. 2012a). In the simplest case the way this might occur is that the model predicts some set of proprioceptive states corresponding to the movement of the body, and then realises those proprioceptive states by the propagation of prediction through classical reflex arcs. Where prediction error is encountered (that is, at the stage when the predicted outcome of the movement does not yet obtain), this error is assigned a very low precision, and so resampled in such a way that the predictions come out true. But the same basic principle can be posited for more complex actions and higher levels of planning, so that the predictive-processing theorist can explain my walk to the cafeteria in terms of the visual, auditory, tactile, and other sensory states that I predict to encounter during the course of that action. It is the minimisation of prediction error that drives the action, but

2 What I mean by representation here should be spelled out. The extent to which PP requires representations, and of what kind, is controversial (see Clark 2016, §6.6, and Hohwy In Press, for diverging views). So I shall here assume a weak and inclusive notion: for these purposes the term "representation" describes an isomorphism between one or more physical particulars and a cause or structured set of causes within the world, so that the particular(s) carry information about their causes in a specific manner (Dretske 1981). That is a very weak notion, but it can be supplemented with further conditions, such as a history of playing a particular role for a system that consumes the representation (e.g., Millikan 1984). Similarly, given the possibly wide scope of PP (section 3.3), it seems prudent to assume a pluralism about representation, on which the content of representations may be fixed by different kinds of facts in different kinds of representational system (Shea 2013).

it is minimised by resampling, rather than updating. This way of distinguishing between perception and action within PP mirrors a distinction in the philosophy of science between theory-revision and experiment (see Hohwy 2013, p. 43, for a nice illustration of this).

Everything that I have explained so far is relatively uncontroversial, except for the way that some of the finer details should be elucidated, for example, whether the hierarchy should be conceived in terms of spatiotemporal or computational depth. But there is a more significant disagreement over the scope of the explanation (Sims 2016). That is to say that not everybody who endorses the general mode of explanation can agree on how much it is supposed to be an explanation for. On this matter, four very general positions can be distinguished (see Table 1). I should note at this point that I do not mean to imply that these positions are consistently held by any particular researchers, except perhaps implicitly. This taxonomy is supposed to be of heuristic value, in mapping out the conceptual possibilities. On the first position (minimal predictive processing), PP is only to be construed as a theory of any number of *perceptual* processes—perhaps visual perception, for example. This is a rare position in the philosophical literature, however, since one thing that seems to be broadly agreed on is that one of the reasons that PP is so interesting is that it can give a unified explanation of perception and action in terms of the same mechanism. This position is also not vulnerable to the dark room problem, and for these reasons it is not relevant for the purposes of my exposition. On the second position (mixed predictive processing), PP is to be interpreted as a theory of just some neurocognitive processes of both perceptual and motor kinds. That means that the explanatory burden for neurocognitive function is shared amongst both PP and other kinds of models in a "mixed" theory. For example, one may suggest that there are mechanisms involved in action which are "model-free," and do not require a representation of the causes of sensory states. One may insist, for example, that neurocognitive function includes "complex admixtures of strategies including the canny use of bodily form and various 'representation-lean' ploys." (Clark 2013b, p. 8) On the third position (maximal predictive processing), PP is to be construed as a complete theory of all neurocognitive function. That means that all neurocognitive function can be explained in terms of the minimisation of prediction error. A proponent of this position will insist on the "preposterous" nature of the hypothesis: "it leaves no other job for the brain to do than minimise free energy[3]—so that everything mental must come down to this principle." (Hohwy 2015, pp. 8-9)[4] On the fourth, and boldest, position, the mechanism described in PP is not only to be understood as a complete theory of neurocognitive function but of adaptive behaviour in general; all adaptive behaviour is understood in terms of the minimisation of surprise. It seems clear that Karl J. Friston intends the "free energy principle" to be taken in this way:

> Most treatments of self-organization in theoretical biology have addressed the peculiar resistance of biological systems to the dispersive effects of fluctuations in their environment by appealing to statistical thermodynamics and information theory. Recent formulations try to explain adaptive behavior in terms of minimizing an upper (free energy) bound on the surprise (negative log-likelihood) of sensory samples. This minimization usefully connects the imperative for biological systems to maintain their sensory states within physiological bounds, with an intuitive understanding of adaptive behavior in terms of active inference about the causes of those states. (Friston 2013, p. 1)

For all of these positions (except the minimal position) there is a conceptual difficulty with the explanation of action that has been called "the dark room problem." (Friston et al. 2012b). That problem is an apparent consequence of the explanation of action—and therefore motivation as well—on the sole basis of the minimisation of surprise.

3 Free energy is the long term average of prediction-error, and therefore an upper bound on surprise over the long term.

4 It's probable that Hohwy holds a stronger position than this—for example, it seems that he may be willing to countenance an interpretation of natural selection in terms of surprise minimisation (see in particular Hohwy 2015, p. 10)—but what he says here exemplifies well the commitments of the maximal predictive-processing theorist in general.

**Table 1:** Kinds of positions with respect to the scope of PP.

| Position | Scope |
|---|---|
| Minimal predictive processing | Some perceptual processes |
| Mixed predictive processing | Some perceptual and motor processes |
| Maximal predictive processing | All neurocognitive processes |
| Free energy principle | All biological processes, on multiple timescales |

## 2 Three Aspects of the Dark-room Problem

The way that the dark room problem puts pressure on PP and its extension to action can be captured in the following line of reasoning. First, the critic claims that the explanation of action on the basis of prediction-error minimisation entails that the agent is always acting to minimise surprise. Second, it is inferred that this basic principle means that the ideal surprise-minimising agent should be expected to seek out a place where it can be free of surprising and unanticipated stimuli. And this would be an environment that is free of any stimuli whatsoever; this is the "dark room" that gives the problem its name. But then, the critic goes on, this is an absurd consequence; this kind of behaviour would spell extinction for any agent which carried it out. That is because an environment without any stimuli is also an environment without any nourishment or opportunity to reproduce. A creature which behaved in this way would not survive, and it would not pass on its genes to offspring. It follows by *modus tollens*, then, that action must be driven by processes which are not surprise-minimizing. This is how the problem is spelled out in its most basic terms. But in fact as we shall see it is better conceived as a way to articulate a cluster of related concerns, rather than a unitary argument with a single conclusion. I will be arguing that there are three distinct difficulties that are raised here. The first, the negative problem, concerns the apparent result that surprise-minimisation entails that the agent seeks to rid itself of stimuli altogether; the second, the positive problem, concerns the presumed poverty of behaviour that could be produced by mere surprise-minimisation (the charge is that it fails to account for rich repertoires of exploratory, exploitative, and playful behaviour); the third, the problem of triviality, is a sceptical concern about whether the extension of scope means sapping PP of any empirical content, rendering it trivial. What unifies these problems under the "dark room" rubric is that they originate in the concerns about how to model motivation within the framework; this is an issue which is revealed starkly in the dark room scenario. The reason that the problem of empirical content is also related to the dark room scenario is that it is attempts to address the positive problem in terms of evolutionarily selected "deep" priors which gives rise to the charge of triviality; this will be made clear in section 3.3.

### 2.1 The Negative Problem

The negative problem is the aspect that is the most initially intuitive. It is the identification of surprise-minimisation with stimuli-minimisation in general. There is a sense in which all stimuli are minimally surprising, if one takes the baseline for stimulation to be the state of the organism prior to the stimuli. That is to say that any stimuli will be a change in what the organism's current state is, and be surprising in virtue of this difference, however minimal. On top of this, it appears to be that a consistently applied PP framework will place surprise-minimisation as the sole principle that drives action. This need not be the case—a contrary example would be the mixed models of section 3.1—but this is certainly the case for the more ambitious readings of PP on which it is supposed to suffice as a unified theory of all the processes underlying perception and action.

So with these two pieces in place, the negative problem states that insofar as all stimuli are minimally surprising, and insofar as a consistently applied PP must place surprise-minimisation as the sole principle driving action, then it seems that the surprise-minimising agent ought to minimise stimuli in general. But that seems wrong, or at least at odds with what we know about living things: they don't

seek to minimise stimuli in general. So something has gone awry. Either the analysis of all stimuli as minimally surprising is incorrect, or it's wrong to say that all action is the minimisation of surprise.[5] Here are two examples of the dark room problem thus characterised in the literature. The first is in a paper by Schwartenbeck and collaborators, who aim to give a formal treatment of the issue, and who are one of the first to explicitly distinguish the negative and positive problems: "Should we not, in accordance with the principle, prefer living in a highly predictable and un-stimulating environment where we could minimize our long-term surprise?" (Schwartenbeck et al. 2013) And another more recent example, from Clark's recent and comprehensive book-length treatment of the philosophical issues associated with PP:

> The hapless prediction-driven organism, the worry goes, should simply seek out states that are easily predicted, such as an empty darkened room in which to spend the remainder of its increasingly hungry, thirsty, and depressing days. This is the so-called 'Darkened Room Puzzle'. (Clark 2016, p. 262)

That is the baseline concern that one can take away from the dark room problem: that minimising surprise entails minimising stimuli. One may deal with this problem by rejecting either of the two premises that lead to that result: either that all stimuli are surprising or that all action is surprise minimisation. On the first strategy, we are owed an explanation of why some stimuli are surprising and why others are not; on the second, we are owed an account of the other mechanisms that drive action, and how they are related to PP.

## 2.2 The Positive Problem

Typically, then, the PP theorist rejects the premise that all stimuli are minimally surprising. She rejects this on the basis that surprise minimisation always occurs relative to a model which assigns probabilities over possible sensory states, based on a set of (Bayesian) prior beliefs about the structured causes which produce those sensory states. This hierarchical generative model predicts the states that the agent will encounter, and it's on this basis that the least surprising state is in fact not a 'blank slate', as the negative problem assumes, but rather the kind of environment that the agent already expects to encounter. For prediction-error minimising agents, action is driven by prediction error relative to a set of predictions about the optimum states for the agent to be in. And this, the PP theorist will say, is by no means the darkened room of the negative problem.

Now, the positive problem has to do with the role that is left for uniquely motivational states to play in PP, once their traditional role has been usurped by "prior beliefs" or "predictions" about the agent's sensory states, and which drive action on the PP framework. It may be that, by positing the influence of a model of causes of sensory states, the PP theorist can show how the negative problem is mistaken. She can do so by pointing out that the sensory states that the agent expects to be in are species specific and are specified on the basis of an agent's prior adaptation to a particular ecological niche. But this seems to imply that we are left without the classical distinction between representations with different directions of fit, because the way that action works within PP is by minimising surprise with respect to prior beliefs rather than maximising utility with respect to the agent's desires. Thus we see Pezzulo and colleagues claim that:

> […] our scheme for behavioural control is based on Bayesian inference and does not call on reward prediction errors for learning or inference. One advantage of this is that the concept of rewards is replaced by the realization of prior preferences. This means that epistemic value and pragmatic

5 Klein (in press) notes that this problem is recognised early on by Mumford (Mumford 1992). But in fact the problem is extant even *earlier* on; even Freud (Freud 1950 [1895]), who inherits the Fechnerian (Fecher 1873) conception of cognition as the minimisation of quantity, is compelled to posit the "reality principle" in order to address very similar issues.

> value (e.g., utility or reward functions) have the same currency and can be accommodated within the same (information hungry) Bayesian scheme […] (Pezzulo et al. 2015, p. 27)

However, this may render a large number of observable behaviours inexplicable – those behaviours that are playful, exploratory, and exploitative. In other words, reducing the states driving action to states that predict future states of affairs means that we can no longer make a cogent distinction between what is probable for an agent and what is of value for it, and we need the concept of value to explain why there are unlikely or uncertain states that are nonetheless valuable or desirable on the part of the agent. It's fairly intuitive to judge that states with high utility for an agent are not always those with high prior probability, and putative examples of this dissociation are becoming ever more widespread in the critical literature:

> Should the first amphibian out of water dive back in? If a wolf eats deer not because he is hungry, but because he is attracted to the equilibrium state of his ancestors, would a sudden bonanza of deer inspire him to eat only the amount to which he is accustomed? Should a person immersed in the "statistical bath" of poverty her entire life refuse a winning lottery ticket, since this would necessitate transitioning from a state of high equilibrium to a rare one? (Gershman and Daw 2012, p. 306)

In other words, even if the dark room scenario does not obtain (agents don't aim to minimise all stimuli) we should still expect an agent who only minimises surprise with respect to a model of the causes of sensory states to lack many of the playful, exploratory, and exploitative behaviours that we observe in the natural world—and perhaps more pertinently, in ourselves.[6] For example, it is not necessarily adaptive for an agent to consume resources in amounts predicted by past consumption—especially if past consumption was in meagre amounts. We would expect that agent, if adaptive, to *exploit* any sudden availability of resources. The criticism that surprise-minimisation does not predict such behavior remains live even after the negative problem is resolved by appeal to a model which relativises surprise minimisation to a particular set of prior beliefs.

## 2.3 The Triviality Problem

The triviality problem has to do with the empirical content of PP. Stated baldly, it is a concern that PP, when pitched at a sufficiently wide scope, may turn out to lack empirical content. This is a problem that is most often associated with the widest-scoped version of PP, the free energy principle. The free energy principle becomes relevant at this point because one way in which the positive problem may be overcome is to expand the scope of the model in order to include evolutionary influences. That would offer a possible way of addressing the issue because it can explain the existence of the problematic behaviours in terms of priors that are evolutionarily specified. It does not require that these priors be learned from the environment. It can do so because evolution is itself construed as the minimisation of surprise at phylogenetic timescales.

Although the triviality issue is very often raised informally at conferences, it is less widespread in print.[7] But I can provide two loci at which the issue is raised explicitly by critics. The first is in the original discussion of the dark room problem itself, in the dialogue between Friston, Thornton, and Clark. In responding to the notion that the negative problem is dissolved by appeal to the model of the causes of sensory states, Thornton makes the charge that: "If we allow unlimited rein over the interpretations [i.e., models] agents are assumed to apply, the dark room problem can be eliminated. But the hypothesis then seems to be stating something that is true by definition." (Friston et al. 2012b, p. 1)

6 Of course, there are various ways in the literature to deal with this concern, see especially section 3.2 below.

7 Though more so with respect to Bayesian models of cognition in general, cf., Bowers and Davis 2012.

Indeed, this sometimes seems to be the case, and a tautological reading of the free energy principle is even sometimes endorsed quite explicitly. Look at what Friston himself says in the same article: "The tautology here is deliberate […] Like adaptive fitness, the free-energy formulation is not a mechanism or magic recipe for life; it is just a characterization of biological systems that exist." (Friston et al. 2012b, p. 2) The concern here is therefore that the characterisation of all action as surprise-minimisation reduces to a definition or tautology, and does not constitute an empirical hypothesis that generates substantive predictions or explains causes in the world.[8]

One way to interpret this is a concern about falsifiability. That would be to say that the free energy principle is not falsifiable, because it is compatible with every state of affairs. It therefore lacks empirical content. Although the falsifiability interpretation of the triviality problem is a common critical refrain, is has been noted that the objection lacks some force (Hohwy 2015, p. 14-15). Firstly, it is a widespread view that falsifiability is not sufficient nor even necessary for something to be a genuine scientific explanation. Secondly, the relationship between biological function and natural selection seems similarly definitional or conceptual in this way, but almost nobody denies that the theory of natural selection is an empirical theory that describes a causal process (Ruse 2008, pp. 44-45).

So with the natural selection analogy in mind, one way to deal with the triviality problem has been to argue that the value of PP will come out of its pragmatic value in constraining more local empirical hypotheses that describe mechanisms more limited in scope. And it may not only act as a general constraint (perhaps in the same way that the laws of physics (controversially) constrain explanations of particular systems), but also perhaps be suggestive of explanations in subordinate domains like systems biology and abnormal psychology.

That then leads us to the second place where this concern is raised at length in the literature, and to the more subtle version of the triviality problem that it constitutes, in a soon-to-be-published treatment by Klein (Klein in press). Klein notes that even if it is the case that PP and the attendant free energy principle admits of pragmatic value in producing hypotheses, the fact that its proponents quite openly admit the tautological nature of the wider scoped hypotheses is problematic:

> Appeal to apparent tautologies should trouble you. For whatever tautologies do, they don't explain why things happen. At best, they give us reason to believe that something is the case. But philosophy of science has moved away from epistemic conceptions of explanation and towards ontic ones […] Good explanations detail a causal story, and it is not obvious that [the free energy principle] does so. (Klein in press)

That leaves no empirical content for the theory itself – it is rather what Klein, after McMullin (McMullin 1985), calls a "Galilean idealisation." Like the frictionless plane, the content of the free energy principle would on this view be literally false, though perhaps useful in formulating empirical hypotheses if taken with a grain of salt. It remains to be seen what consequences this has for the way we should think about the theory, both in an epistemic (does it make endorsement of it less justified?) and ontological (what entities and processes does it imply?) sense.

Such is the dialectic that leads us through the three aspects of the dark room problem. The initial intuition is that replacing conventional motivational imperatives with the imperative to minimise surprise entails absurd consequences—that the best strategy for surprise minimisation would be the minimisation of sensation in general. That produces the appearance of a dilemma, with two corresponding ways to reply: either sensation is not minimally surprising or not all motivation is surprise-minimisation. The latter is relatively undesirable, since it undermines the unificatory appeal of PP. But the former is easily followed, given that PP includes the notion of a model of causes of sensory

8 An anonymous reviewer suggests that this is a straw man. But it is a view clearly held by critics: "[Thornton:] we can certainly view the process by which agents adapt to their environments as a process by which they reduce their surprise. The problem is we can also view it the other way around, seeing the situation in terms of agents reducing their surprise by adapting to the environment." (Friston et al. 2012b, p. 1)

states within which there are states that are less surprising than an absence of states altogether. But even with the model taken into account, one may doubt whether the imperative of surprise-minimisation can produce complex behaviours like play and exploration. Last, there also seems to be a problem of triviality for wider-scoped free energy principle that is also related to the dark room problem. That emerges out of the appearance of tautology in the claim that all adaptive behaviour is free energy minimisation, and challenges the ability of free energy theorists to provide substantive empirical hypotheses regarding specific mechanisms.

## 3 Three Replies to the Dark-Room Problem

### 3.1 Mixed Predictive Processing

On first gloss, it seems that the issues associated with the dark room problem pushes us towards a view that is mixed. A view that is mixed is a view that makes room for cognitive processes that are not surprise-minimisation. (Clark 2016) is the most thorough working out of a view like this that exists in the philosophical literature, though it is (Pezzulo et al. 2015) that have given a more thorough mechanistic account; I will base my discussion around both of these. However, I should begin my explanation of these mixed predictive processing theories by saying something about the distinction between model-free and model-based processes, since one way to develop a mixed view of predictive processing is to have model-free processes play the role of motivating action, and thereby giving an answer to the (negative) dark room problem that takes the horn of the dilemma on which action is driven by processes *other* than surprise-minimisation of the kind posited in PP.[9]

The distinction between model-based and model-free processes originates in the study of reinforcement learning, where it is used to distinguish between a process that learns the value of available options by trial and error, and without a model of the causal structure of the environment—this is model-free—and learning that assigns value on the basis of a model of how rewarding events in the world are statistically related to other events—this is model-based (Gläscher et al. 2010). PP would fall unambiguously under the "model-based" category of learning processes. Model-free processes are attractive in the context of the dark room problem because they may be construed as offering a set of imperatives to action which could be mixed with the standard PP mechanism of surprise minimisation in order to yield imperatives to action that look genuinely "motivational," and which therefore entail action that defeats the dark room problem. For example, it may be that there is a mechanism underlying action that causes an agent to indiscriminately seek out and consume sources as reward on the basis of availability, and drives action in this way. Ainslie's (Ainslie 2001) behavioural findings of hyperbolic discounting could be indicative of such a mechanism. He has found in studies both in animal models and human participants that the way rewards are valued increases steeply as they approach in time, and that smaller but more immediate rewards tend to be consumed in preference to temporally distant but larger rewards in decision-making tasks. It may be that whatever mechanism produces this effect works in independence from any kind of model of the causes of reward, that is, it is model-free (cf., Clark 2016, pp. 252-256).

If this is the case, then a mixed theorist can give an answer to the dark problem which deals with all three aspects at a single stroke. The negative problem is clearly no issue, since the model-free mechanisms that drive action are not minimising surprise, they are reward-seeking and therefore attempt to bring the agent into contact with the appropriate stimuli. And cutting off the dark room problem this early in the dialectic also means that the other two issues (the positive problem and the triviality problem) do not come up, since those problems result from the sole appeal to a model in order to deal

9 An anonymous referee has advised me that model-free processes can be understood within the context of predictive processing (e.g., Pezzulo et al. 2015; also Clark 2016, § 8.6), and thereby constitute a complement to predictive processing rather than a competitor. That's true. But it's not necessary to do so, and to construe model-free processes in this way just means that the model is not genuinely mixed; it collapses into the maximal view.

with the negative problem; this answer takes the other horn of the dilemma, which doesn't lead to those two issues.

However, it is unlikely that many PP theorists are going to want to take this path. That is because it undermines one of the most attractive features of PP: the way in which it serves to unify the mechanisms underlying perception and action within a single theory. Indeed, one might observe that it solves the problems associated with PP by ceasing to be a PP theory; it fails to unify perception and action in the way that is extolled in the philosophical literature:

> [PP] is a proposal that has already been applied to a large—and ever-increasing—variety of phenomena. It thus serves as a powerful illustration of the potential of some such story to tackle a wide range of issues, illuminating perception, action, reason, emotion, experience, understanding other agents, and the nature and origins of various pathologies and breakdowns. (Clark 2016, p. 10)

This may be one reason why authors have formulated mixed views on which the model-based processes *themselves* play an arbitrating role. That is to say that it is the models themselves which determine when model-free processes drive action, and when learning is instead contextualised within a model of the causes of sensory states. Again, Andy Clark holds a view like this: he thinks that "[…] a kind of meta-model […] would be used to determine and deploy whatever [model-based or model-free] resource is best in the current situation, toggling between them when the need arises." (Clark 2016, p. 253) This is an attractive view for other reasons, as well. There is evidence to suggest that model-free learning processes do not exist in isolation from those that are model-based, but rather that they are highly integrated (Daw et al. 2011).

Clark's proposal, more specifically, is that model-free and model-based processes need to be understood as situated along a scale where the latter kind of learning is dominated by top-down influence within the generative hierarchical model and the former is dominated by bottom-up sensory influences. A mechanism of this sort is outlined more formally in Pezzulo et al. (Pezzulo et al. 2015). They envision the relationship between model-based and model-free in terms of a hierarchy where higher-levels within the model contextualise lower levels, and that learning is to be considered "model-free" when the higher levels fail to contextualise those lower. What determines whether contextualisation occurs or not are assignments of precision within the model; when prediction errors are assigned higher precision values then they drive action in ways that are less contextualised, because the higher levels of the model exert less of an influence.

Notice, however, that it seems to be that that *this* kind of mixed PP view entails that all learning is minimally inferential and model-based, because there is no learning that is entirely independent of the meta-model. Certainly, Pezzulo et al. (Pezzulo et al. 2015, p. 32) seem to recognise this: "Strictly speaking […] habitual behaviour is not completely model free in that it continues to depend on the (simplest) type of predictive model, of the kind 'because there is a stimulus, I expect a response'." So a mixed-view, when it is properly elaborated, appears to in fact be a species of "maximal predictive processing" theory. That is because after all is said and done, surprise minimisation nonetheless remains the sole imperative driving action, even in putatively "model-free" modes of learning. Therefore, whether or not this view is successful in addressing the dark room problem depends on whether these maximal views are so successful. Let's consider that now.

## 3.2 Maximal Predictive Processing

Maximal predictive processing is the view that prediction-error minimisation is *all* that the brain ever does; all neurocognitive function can be explained in terms of the minimisation of surprise. So we aim to explain all cognition, and thereby all mental phenomena, in terms of prediction-error minimisation. The maximal theorist claims that the mammalian brain works (and only works) by minimising

prediction error, but is agnostic on the question of whether or not other biological entities or systems function in this way.

In facing up to the negative aspect of the dark room problem, the maximal PP-theorist chooses to take the horn of the dilemma on which not all stimuli are minimally surprising. The way this works is to demonstrate that the minimisation of surprise is carried out with respect to a hierarchical generative model of the causal structure of the world. This model assigns probabilities to sensory states—that is, it generates predictions—such that the agent anticipates that it will be in states that reflect the ecological niche to which it is adapted, which is that within which it can harvest reward and pass on its genes. Given that ecological niche is species-specific, it appears that there needs to be some kind of story given here about the origins of the priors which specify that niche. In other words, there must be a relation between the neural and surprise-minimising morphology of the agent and the non-neural but niche-specifying morphology of the agent, such that the non-neural morphology can play an appropriate role within the model without itself being surprise-minimising. If non-neural morphology is in fact directly (and not vicariously) surprise-minimising, then this view collapses into the much stronger free energy principle which I discuss in section 3.3.

Now, there are a number of ways in which this general strategy may be pursued. One is to say that the relevant morphological traits are themselves represented within the model, and that this allows them to play a role in prediction-error minimisation without themselves minimising prediction error. With this in mind, a first attempt at such an account of morphological representation within the surprise-minimising brain might focus upon the interoceptive prediction of the internal milieu (Craig 2003). Interoceptive systems monitor the physiological states of the body such as "[…] those relating to heart rate, glucose levels, build-up of carbon dioxide in the bloodstream, temperature, inflammation, and so on." (Barrett and Simmons 2015, p. 419) The prediction of these sensory states leads us to perceive them as feelings about those states. So, for example, we might interoceptively perceive dehydration as thirst. There are influential attempts in the literature to account for interoception within the scope of PP (Seth 2013; Barrett and Simmons 2015). Within the PP framework the homeostatic states are those that are predicted, and deviation from those predicted states (deviation from interoceptive states associated with satiation, for instance) will lead the creature to take action in order to bring itself back into line with those states.

That seems a satisfactory first pass at how the basic PP story may be extended to take the morphology of the agent into account. When it comes to the negative aspect of the dark room problem, this affords the following answer. The agent does not stay in the dark room because doing this would lead it to occupy sensory states that are surprising, states that are interoceptively perceived as hunger and thirst. Therefore, staying in this impoverished environment does not in fact minimise surprise, but rather elicits it. In other words, staying in that environment is a very poor strategy for the minimisation of surprise. A much better strategy is to actually get out and exploit richer environments for nourishment so that the interoceptive states can be brought back into line with prior expectations. The bottom line is that on this view we should not expect the dark room scenario to obtain. That is because the dark room problem does not take into account interoceptive sources of surprise, but only exteroceptive sources like vision and hearing. Furthermore, we could conceivably extend this undeveloped account to encompass those exteroceptive senses. For instance, the mammalian eye is structured in such a way that it is receptive to changing patterns of light. It is not unconceivable that part of the model reflects this morphological trait, eliciting surprise when such stimuli are absent. In other words, an absence of visual stimulation would itself be surprising.

Now, the answer to the negative aspect of the dark room problem that is given here is one which generates the positive problem. The stipulation of a model which specifies the expected states (states that are not the dark room) is subject to the issue that motivation is solely driven by prior beliefs, leaving no room for pro-attitudes as traditionally conceived. It is thought that this crowding out of pro-attitudes by prior beliefs would lead to an impoverished behavioural repertoire that would fail to

include the ubiquitous tendencies towards play, exploration, exploitation, and behaviours do not seem to be best conceived as the minimisation of surprise. There are two quite general answers to be given at this point.

The first, as set out in Schwartenbeck et al. (Schwartenbeck et al. 2013), has to do with the way that prediction-error minimisation is formalised within PP. More specifically, when an action policy is selected amongst alternatives, the fact of uncertainty about outcomes will dictate that an agent attempts to visit many varied states with equal probability. That is because it will not be clear for the agent which states actually have the highest utility (construed in terms of prior beliefs). So there will be a shift between occupying the least surprising states and many novel states, depending on the level of uncertainty: "[…] when the differences in the expected utilities of outcomes become less differentiable, agents will try to visit several states and not just the state that has highest utility." (Schwartenbeck et al. 2013, p. 3) There will be a context-sensitive weighting of these two different kinds (exploitative and exploratory) kinds of strategy, where this weighting is influenced by the estimation of uncertainty through the assignment of precision as described in the first section of this paper.

In fact, this may seem to mirror the distinction between model-free and model-based modes of learning, as they are understood by both Pezzulo et al. (Pezzulo et al. 2015) and Clark (Clark 2016). That is because for them that distinction also appears to be a trade-off between two kinds of strategy that is arbitrated by the dynamics of precision assignment. Here, the agent would switch between exploratory and exploitative modes of engagement with the environment on the basis of how reliable their information is considered to be vis-à-vis the states that are least surprising (have the highest "value"). When such information is assigned very high precision values, then the agent engages in exploitative behaviours because there are states that are unambiguously more valuable than other states. But when such information is assigned low precision values, then the agent engages in exploratory behaviours because it is not sure which state will be most valuable (probable).

The other answer to be given is that the minimisation of prediction error takes place within the context of a hierarchical model, which means that the minimisation of surprise is an optimisation process that occurs over very many levels of spatiotemporal generality. With this in mind, it may well be the case that intuitive appeals to putative counter-examples where there are unlikely events that have very high utility do not sufficiently take into account deeper imperatives, or a balance between those imperatives and others in the hierarchy. For example, in response to the question of Gershman and Daw (Gershman and Daw 2012, p. 306), "[s]hould a person immersed in the "statistical bath" of poverty her entire life refuse a winning lottery ticket[?]", we might respond that although it is indeed true that for someone with a long history of poverty the state of sudden riches would be surprising, the deeper prior belief that compels the person towards keeping themselves fed or to acquire resources means that they will try to get themselves out of that situation of poverty if given such a chance. They won't refuse the ticket. So perhaps we can account for apparent counter-examples of this kind by appeal to distinctions between prior beliefs at different levels of hierarchical depth, or different levels of spatiotemporal generality. The deeper those beliefs are, the more likely they are to look like states with high utility rather than high prior probability.

There is a vexing question that comes up at this point of appeal to evolutionarily selected "deep" prior beliefs. One may first suppose that some of these deep priors need not be genetically innate but can be extracted from the environment itself. These would be priors that are extracted from highly consistent regularities within the environment, for example, the inability of two solid objects to simultaneously occupy the same space (cf., Hohwy et al. 2008, p. 692). However, it is unlikely that all such deep priors can be accounted for in this way. That is because many such priors will be idiosyncratic to whatever species the agent belongs to. Since the regularities that are extracted from the environment are presumably *in* the environment, they must remain constant across species. So if the deep prior in question is idiosyncratic to species (e.g., some prior or set of priors that produces a behavioral disposition to seek out dark environments in troglofauna, for instance), then it appears that it cannot

be learned from the environment. It must be innate. So the question I have in mind can be posed as follows.

We think that we know that brains perform many of their tasks in virtue of their performing active inference. Now it appears that in order for us to be able to explain motivation within a pure active-inference framework we need to posit innate priors. These innate priors are determined by the total morphology of the organism, insofar as this morphology is isomorphic with an optimum trajectory or a set of good-enough trajectories through state space. So the question is this: how does the non-neural morphology play the right role in active inference? This can't just be through its representation in active inference, because then there's no reason those innate priors don't just update when they encounter prediction error—why they are recalcitrant. Having the priors themselves be fixed morphological traits or processes (metabolism, for instance) addresses this issue, but then we have a problem about the computational interaction between neural and non-neural morphology. How is such interaction to be explained?

I will be arguing that the free energy principle, as developed and applied by Friston and his collaborators, provides one kind of answer to such questions. This is a kind of PP that is expanded in scope in order that it constitutes an explanation of biological adaptation in general, and on various timescales. In order to give a full reply to the dark room problem, the maximal-PP theorist is obliged to go further and either: i) embed PP within the wider scoped free energy principle; or ii) give an alternative account. I am open to the idea that there is an alternative account available, but in the rest of this paper I will be exploring (i).

### 3.3 The Free Energy Principle

It looks as though a satisfactory answer to questions about how we came to have the priors that we have can be given by embedding the maximal PP story within the wider scoped free energy principle. The free energy principle gives a surprise-minimisation account of biological processes in general, which affords us a way to explain the origin of the prior beliefs that are relevant to answering the dark room problem and the way in which those prior beliefs are related to morphological facts about the agent. But to see why this is so it's first necessary to set out the theory in sufficient detail.

We can start by explaining its initial motivation. The free energy principle is usually motivated by a much more general reflection on a putative distinction between biological and non-biological self-organising systems (e.g., Friston and Stephan 2007, §2.2). An example of the former kind might be a bird, or a bacterium. An example of the latter kind might be a snowflake, or a hurricane. Both of these kinds of complex system exhibit self-organisation; that means that they both spontaneously arrange themselves into an ordered pattern or structure without the intervention of an outside agent (Ashby 1962). However, Friston and Stephan (Friston and Stephan 2007) note a qualitative difference between them. The difference is that biological systems are *adaptive*. In the case of a snowflake, for instance, it will cross a phase boundary and melt with the change of temperature. But the "[…] key aspect of biological systems is that they act upon the environment to change their position within it, or relation to it, in a way that precludes extremes of temperature, pressure or other external fields." (Friston and Stephan 2007, p. 422)

The distinction may not be as stark as these authors suggest. After all, given sufficiently rapid and intense changes in temperature, biological systems also dissipate into the environment. That is to say that systems like snowflakes are not unique in this respect. Conversely, one may give potential examples of self-organising systems that act on their environment but that are non-biological. Aggregates of biological systems (like societies) act on their environments in some way, but it would be controversial to label these biological in the same way as their constituents are. But critique of this kind lacks propriety; Friston and Stephan are not doing conceptual analysis, they are suggesting constraints on the

behaviour of biological systems for heuristic purposes. Their question is this: how is it possible for a biological system to avoid dissipation? For this purpose, their loose distinction is sufficient.

One way to understand that capacity is in terms of an exchange of energy between the organism and its environment. This reflects the traditional biophysical understanding of biological systems as energetically open systems: they take in energy and matter in a low-entropy form as nutrition (construed broadly) and excrete it back into the environment in the form of relatively high-entropy waste. This allows us to reconcile the increase of complexity and order in living systems with the second law of thermodynamics, which states that entropy is always increasing in closed systems. The biological system is an open system, which allows it complexity and order at the expense of its surrounding environment (Schrödinger 1944).

Another way to understand this capacity of regulating the relationship to environment is in terms of information—it is to understand the capacity as that of moving around within a particular set of sensory states to which the system is suited. That set of states is implicitly specified by the phenotype of the system, because the system is already evolutionarily adapted to some specific environmental niche. A ferrophilic bacterium, for example, is adapted to a solution which contains specific levels of iron and oxygen. As such its phenotype will bear some substantive relation—perhaps representational (Shea 2012)—to this niche, and must alter its relationship to the environment in line with that relation. If the relation is construed in terms of representation, for instance, then it must regulate its relationship to the environment so that the propriety-conditions of that representation are satisfied.

These two ways of understanding biological systems—thermodynamic and informational—are complementary. Even though the environment is always becoming more and more disordered, it nonetheless behaves in a regular and lawful way, and the exploitation of this regularity makes it possible for the biological system to embed that regularity into its physical structure: "organisms could maintain configurational order, if they transcribed physical laws governing their environment into their structure." (Friston and Stephan 2007, p. 422)

How do biological systems manage to do this? On the free energy principle, the task is construed as a problem of Bayesian inference. The inference in question occurs across a boundary that segregates the internal states of the system and its external environment—this boundary is called a Markov blanket. Markov blankets consist of two kinds of state: sensory states and active states. The prototypical example of some such boundary is the cell wall. The task of the biological system is to infer the causes that act on it from the outside, with access only to the sensory states in the Markov blanket. The way that it does so can be modelled with the very same formalisms that govern active inference and belief updating in PP as applied to neurocognitive function. That is to say that the system approximates a model of the causes of its sensory states in two ways: by updating its internal states where those states fail to correspond to sensory states, and by acting on its environment in order to change the way that the outside causes generate sensory states.

These very abstract considerations can be illustrated more concretely with reference to the example of circadian rhythmicity (Bechtel 2011; Sheredos 2012). Circadian rhythms are periodic cycles which regulate other processes (metabolic, behavioural, genetic, and so on) on a roughly 24 hour period. These are sensitive to external cues (so-called Zeitgebers) in calibrating the clock, but the rhythm is endogenously produced, which means that its periodicity will remain in effect even in the absence of any Zeitgebers. Sheredos (Sheredos 2012) has argued that the circadian rhythms of cyanobacteria are cognitive in the minimal sense specified by the free energy principle. That is to say that the systems which perform circadian rhythmicity perform prediction-error minimisation. The circadian rhythm in the cyanobacterium regulates two metabolic processes that are chemically incompatible: photosynthesis (day-time) and nitrogen fixation (night-time). Roughly, the endogenous tendency of the system to a default period and phase of rhythm can be construed as the priors, and the sensitivity to Zeitgebers can be construed as producing prediction error in the system. In the cyanobacterium, the function of signalling prediction error is realised by high levels of phosphate, which are both produced

during photosynthesis and play a central role in transforming the protein which regulates the circadian cycle. So if the bacterium is unexpectedly performing photosynthesis at a time when it predicts there should be low levels of ambient light, the relatively high levels of phosphate will phosphorylate the protein regulating the circadian cycle, and this will recalibrate the clock. This feedback mechanism instantiates the hierarchical and bidirectional feedback mechanism that the free energy principle (and PP) describes, but it does so within a non-neural system. According to the free energy principle, all adaptive behaviour is like this.

The free energy principle may also be considered to apply over longer time scales (Friston 2013; Friston et al. 2015; Hobson and Friston 2016). One may distinguish functional (metabolism, learning, and inference), developmental, and even phylogenetic time scales in this regard. Natural selection takes place over very many generations of individual phenotypes; persistent self-organisation emerges from a fluctuating environment with a general tendency towards disorder, and at the expense of the order in that environment. Hobson and Friston (Hobson and Friston 2016) therefore suggest that natural selection can be explained under the same mathematical formalism as in PP when applied to brain function. Namely, natural selection is to be construed as a process wherein a genotype models the causal regularities (in this case, selection pressures) that impinge upon it from without, and in doing so embeds this causal structure into the phenotype across multiple generations. Each generation is understood to be a Bayesian "update" on the basis of prediction error that corresponds to selection pressures that the phenotype is not yet adapted to:

> […] in natural selection, each new generation corresponds to a Bayesian update, converting a prior distribution over phenotypic characteristics into a posterior distribution. […] this means that evolution is the process of predicting which phenotypes are best adapted to their econiche. (Hobson and Friston 2016, p. 247)

Again a specific example will be helpful. Circadian rhythmicity has a genetic component, as described in (Bechtel 2011). One could construe the endogenous tendency to rhythmicity as the "prior" which is updated on the basis of prediction error that results if Zeitgebers are out of sync with that prior. For instance, if a cyanobacterium is introduced into an environment where the day-night cycle is different (e.g., a different time zone), then its periodicity will be updated on this basis (the phase may well remain the same, if the length of the day itself is the same). However, we may ask questions about how it is that the periodicity has this endogeneity in the first place. The obvious answer is that it is genetic, but this just labels the problem and leaves the mechanism obscure. Bechtel (Bechtel 2011, p. 145) has shown that—in much the same way as the circadian rhythm itself—the genetic basis of the rhythm can be understood as instantiating a feedback mechanism that is regulated on the basis of error. The empirical research which is the basis of his discussion targets the circadian rhythm in fruit flies (Hardin et al. 1990). That research demonstrates that the genetic basis (the gene *per*) of the prior is down-regulated by high levels of the protein (PER) that it expresses; when there is a buildup of PER, this acts on the gene to prevent further generation of the protein. The period of this process is circadian: it occurs over a roughly 24-hour period.

With the foregoing example in mind, here's how the introduction of the free energy principle aids us in answering the questions about priors driving action. The first thing to point out is that the mechanism of free energy minimisation is not arbitrarily limited to instantiation in a *neural* realisation base; any realisation base that is sufficiently complex will be enough. The example of the circadian rhythms suggests that a chemical basis may be sufficient to realise active inference, and that a neural realisation is probably not necessary. The second thing to note is that the functional continuity between various otherwise distinct systems (genetic, cognitive, and behavioural) means that one can locate the origins of priors in different systems and across different timescales; one need not have a single centre of cognition which somehow respects facts about the morphology of the agent by means

of representing those facts in some way. Within this context, the relevant non-neural facts about the agent that make it so that the agent moves into and around a specific environmental niche are not themselves "dumb" or computationally inert, but are built into the computational machinery of active inference itself. So, in fact, there is no hard interface between the brain and the morphological facts that it requires access to in order to drive action in the right ways. So, in principle, both the origins of prior beliefs can be given along with a constraint on the account of how they input into processes of learning, inference, and action.

On the basis of these lines of reasoning, it seems to me that maximal PP that is grounded within the free energy principle is well-equipped to handle the concerns associated with the dark room problem. But the problem of triviality is yet significant. Clark gives voice to these concerns when he suggests that to excessively widen the scope of PP threatens "[…] to over-intellectualize large swatches of adaptive response in both human and non-human animals." (Clark 2013b, p. 8) But of course, the danger of the problem depends on what exactly this over-intellectualisation amounts to. If it commits us to saying that bacteria or genomes have attention, or imagine, or suffer from schizophrenia, then of course this seems like a debauched extension of anthropic notions into domains where they do not belong. But if we are simply placing functional requirements on those simpler systems, then this anxiety seems out of place.

Perhaps on this point we need to bite the bullet of Klein's suggestion – perhaps the free energy principle is a Galilean idealisation. However, something similar may be true of the laws of physics (Cartwright 1986), and these are nonetheless considered to be a significant advancement in scientific knowledge and highly valuable in constraining scientific models in more specific contexts. The same may be true of the free energy principle, which could serve to constrain theorising about specific mechanisms in cases like that of the circadian rhythm, neurocognitive function, and perhaps in future even social and aggregate entities (cf., Friston and Frith 2015). In that case its criteria for endorsement would be largely pragmatic.

## 4 Conclusion

The dark room problem is both plural and significant. I've tried to show here how the various concerns that constitute the problem emerge in the dialectic between PP-theorists and PP-critics. The initial puzzle—the negative problem—is an intuitive and naïve one. The question can be phrased like this: if action is just the minimisation of surprise, then why don't we try to minimise all stimuli? The answer to this question must either devalue the role of surprise-minimisation or explain why not all stimuli are surprising. Mixed PP views take the first horn of this dilemma by specifying mechanisms underlying action which do not work via surprise minimisation. One way to do so, I argued, is to appeal to "model-free" learning processes. These are reinforcement-learning schemes that do not require any representation of the way that events are statistically related to one another; they learn the value of different actions through trial and error.

Another way to construct a mixed PP view is to have both model-free and model-based processes integrated within a 'meta-model': within the predictive-processing architecture itself. Then, I argued, this just collapses into a maximal predictive processing view—that is the view that predictive processing is all that the brain ever does, and so all neurocognitive function must be explained in terms of surprise minimisation. If this view is endorsed, then the PP-theorist is taking the horn of the dilemma on which not all stimuli are surprising; that is because some are assigned a high probability within a model of the causes of sensory states. Then the maximal-PP theorist is obliged to respond to the positive aspect of the dark room problem: how does surprise-minimisation account for behavioural repertoires which include exploration and exploitation?

I suggested that the maximal-PP theorist can give two related answers. The first, following a suggestion by Schwartenbeck et al. (Schwartenbeck et al. 2013), is that exploratory and exploitative be-

haviours will be selected according to a trade-off that is driven by the dynamics of precision assignment. When beliefs about which states are "valuable" are imprecise, then the agent will try to occupy all of them (and find novel ones) in the exploratory mode; when beliefs about such states are precise, then the agent will just occupy those which are most valuable, in the exploitative mode. Second, appeals to prior beliefs which are deeper in the hierarchy can help explain why some states appear to have low probability but high value: it is because they entail deeper states that *do* have a high probability (winning a lottery entails having access to resources).

However, this raises puzzles about the origins of "deep" priors as well as how genetic information might interface with priors that are active in learning and inference in ontogenetic time. The free energy principle provides some suggestions here, though there is still much to be done in this regard. The example of circadian rhythms demonstrates how functional continuity can be established between free energy minimisation in both phylogenetic and ontogenetic time, thereby suggesting a relatively robust account of the way in which tendencies to particular kinds of action can originate in evolutionary processes. However, this raises questions about whether expanding Bayesian active inference to so wide a scope does not sap the free energy principle of any substantive empirical content. I think it is possible that this may be the case. But if it is, we may well go on to ask what such triviality amounts to if the account is both explanatory and of use as a heuristic. Certainly it must not be trivial in any sense that should worry us. But that is a challenge that may be taken up by the critics of PP in future.

## References


Ainslie, G. (2001). *Breakdown of will.* Cambridge: Cambridge University Press.

Ashby, W. R. (1962). Principles of the self-organizing system. In H. H. von Forster & G. W. Zopf (Eds.) *Principles of self-organization: Transactions of the university of Illinois symposium* (pp. 255-278). London: Pergamon Press.

Barrett, L. F. & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nature Reviews Neuroscience, 16*, 419-429.

Bechtel, W. (2011). Representing time of day in circadian clocks. In A. Newen, B. Bartels & E.-M. Jung (Eds.) *Knowledge and representation* (pp. 129-162). Paderborn: Mentis.

Boring, E. G. (1930). A new ambiguous figure. *American Journal of Psychology, 42*, 444-445.

Bowers, J. S. & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138*, 389-414.

Cartwright, N. (1986). *How the laws of physics lie.* Oxford: Clarendon Press.

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

———(2013b). The many faces of precision (Replies to commentaries on "Whatever next?"). *Frontiers in Psychology, 4* (270).

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

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

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*, 1204-1215.

Dretske, F. I. (1981). *Knowledge and the flow of information.* Cambridge, MA: MIT Press.

Fecher, G. T. (1873). *Einige Ideen zur Schöpfungs- und Entwickelungsgeschichte der Organismen.* Leipzig: Breitkopt & Härtel.

Freud, S. (1950 [1895]). Project for a scientific psychology. In J. Strachey (Ed.) *The standard edition of the complete psychological works of Sigmund Freud* (pp. 281-391). London: The Hogarth Press and the Institute of Psychoanalysis.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*, 293-301.

———(2013). Life as we know it. *Journal of the Royal Society Interface, 10*.

Friston, K. & Frith, C. (2015). A duet for one. *Consciousness and Cognition, 36*, 390–405. https://dx.doi.org/10.1016/j.concog.2014.12.003.

Friston, K. J. & Stephan, K. E. (2007). Free-energy and the brain. *Synthese, 159*, 417-458.

Friston, K., Adams, R. A., Perrinet, L. & Breakspear, M. (2012a). Perceptions as hypotheses: Saccades as experiments. *Frontiers in Psychology, 3* (151).

Friston, K., Thornton, C. & Clark, A. (2012b). Free-energy minimization and the dark-room problem. *Frontiers in Psychology, 3* (130).

Friston, K., Levin, M., Sengupta, B. & Pezzulo, G. (2015). Knowing one's place: A free-energy approach to pattern regulation. *Journal of the Royal Society Interface, 12*.

Gershman, S. J. & Daw, N. D. (2012). Perception, action and utility: The tangled skein. In M. I. Rabinovich, K. J. Friston & P. Verona (Eds.) *Principles of brain dynamics: Global state interactions* (pp. 293-312). Cambridge: MIT Press.

Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66*, 585-595.

Hardin, P. E., Hall, J. C. & Roshbash, M. (1990). Feedback of the Drosophila period gene product on circadian cycling of its messenger RNA levels. *Nature, 343*, 536-540.

Hobson, J. A. & Friston, K. J. (2016). A response to our theatre critics. *Journal of Consciousness Studies, 23*, 245-254.

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

——— (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt am Main: MIND Group.

———(In Press). The predictive processing hypothesis and 4e cognition. In A. Newen, L. Bruin & S. Gallagher (Eds.) *The Oxford handbook of cognition: Embodied, embedded, enactive and extended*. Oxford: Oxford University Press.

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

Klein, C. (in press). What do predictive coders want? *Synthese*. https://dx.doi.org/10.1007/s11229-016-1250-6.

McMullin, E. (1985). Galilean idealization. *Studies in the History and Philosophy of Science Part A, 16*, 247-273.

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism*. Cambridge, MA: MIT Press.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics, 66*, 241-251.

Pezzulo, G., Rigoli, F. & Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology, 134*, 17-35.

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*, 79-87.

Ruse, M. (2008). Darwinian evolutionary theory: Its structure and its mechanism. In M. Ruse (Ed.) *The Oxford handbook of philosophy of biology* (pp. 34-63). Oxford: Oxford University Press.

Schrödinger, E. (1944). *What is life? The physical aspect of the living cell*. Cambridge: Cambridge University Press.

Schwartenbeck, P., FitzGerald, T., Dolan, R. J. & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology, 4* (710).

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

Shea, N. (2012). Inherited representations are read in development. *British Journal for the Philosophy of Science, 64*, 1-31.

———(2013). Naturalising representational content. *Philosophy Compass, 8*, 496-509.

Sheredos, B. (2012). Reductio ad bacterium: The ubiquity of Bayesian "brains" and the goals of cognitive science. *Frontiers in Psychology, 3* (498).

Sims, A. (2016). A problem of scope for the free energy principle as a theory of cognition. *Philosophical Psychology, 29*, 967-980.

Vance, J. (2015). Review of the predictive mind. *Notre Dame Philosophical Reviews*.

# Affective Value in the Predictive Mind

Sander Van de Cruys

Although affective value is fundamental in explanations of behavior, it is still a somewhat alien concept in cognitive science. It implies a normativity or directionality that mere information processing models cannot seem to provide. In this paper we trace how affective value can emerge from information processing in the brain, as described by predictive processing. We explain the grounding of predictive processing in homeostasis, and articulate the implications this has for the concept of reward and motivation. However, at first sight, this new conceptualization creates a strong tension with conventional ideas on reward and affective experience. We propose this tension can be resolved by realizing that valence, a core component of all emotions, might be the reflection of a specific aspect of predictive information processing, namely the dynamics in prediction errors across time and the expectations we, in turn, form about these dynamics. Specifically, positive affect seems to be caused by positive rates of prediction error reduction, while negative affect is induced by a shift in a state with lower prediction errors to one with higher prediction errors (i.e., a negative rate of error reduction). We also consider how intense emotional episodes might be related to unexpected changes in prediction errors, suggesting that we also build (meta)predictions on error reduction rates. Hence in this account emotions appear as the continuous non-conceptual feedback on evolving — increasing or decreasing — uncertainties relative to our predictions. The upshot of this view is that the various emotions, from "basic" ones to the non-typical ones such as humor, curiosity and aesthetic affects, can be shown to follow a single underlying logic. Our analysis takes several cues from existing emotion theories but deviates from them in revealing ways. The account on offer does not just specify the interactions between emotion and cognition, rather it entails a deep integration of the two.

**Keywords:**

Affect | Dark room | Emotion | Prediction error | Predictive processing | Reward | Uncertainty | Valence

**Acknowledgements**

Sander Van de Cruys is a postdoctoral fellow of the Research Fund Flanders (FWO). Partly supported by a Methusalem grant by the Flemish Government (METH/08/02) to Johan Wagemans. I want to thank Johan Wagemans, Jan Lauwereyns, Andrey Chetverikov, Agnes Moors, Tomer Fekete, Chris Trengove, Omer Van den Bergh, and three anonymous reviewers for helpful comments on an earlier draft of this paper. I am very grateful to the organizers and the participants of the MIND23 workshop (Frankfurt, May 2016) for inspiring discussions, and to the editors and three anonymous reviewers for constructive comments. The views expressed (and any remaining errors) are solely my own responsibility.

*Happiness is neither virtue nor pleasure nor this thing nor that, but simply growth. We are happy when we are growing.*
*— W.B. Yeats*

In his seminal work on the relation between cognition and emotion, Robert Zajonc (Zajonc 1980) wrote that "preferences need no inferences". If this were true, affect would never find a place in the currently popular attempt at a unified model of cognition, called predictive processing (PP), that holds that inference is all the brain does (Clark 2013; Friston 2010). Indeed, there is so far little work on affective value and affective experience in PP (but see Barrett and Simmons 2015; Seth 2013). Nevertheless, PP has in recent years been shown to hold a lot of promise in blending perception, action and cognitive beliefs into a coherent, well-founded framework, pleasingly taking down the walls between these subfields. Although plenty of fundamental issues concerning its computational articulation and biological implementation remain (e.g., see commentaries on Clark 2013), as a unified theory of cognition it arguably fares better than any other alternative we have. Crucially, if it is to become an overarching framework of the mind-brain, emotions, so central to existence and survival, somehow

Van de Cruys, S. (2017). Affective Value in the Predictive Mind.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 24. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573253

have to fit in, including their experiential aspects. However, the unifying logic of PP—a single computational principle for the whole brain—seems directly opposed to the popular notion in emotion theorizing that emotions are a bricolage of modules adapted to very specific challenges in our ancestral environment. Rather than built around a single neat, optimal logic, emotions are assumed to be a messy, ad hoc bag-of-tricks. In practice, it has, however, proven difficult to distinguish different emotion 'modules' in the brain, even at the subcortical level, which has led some emotion theorists (e.g., Barrett 2014; Carver and Scheier 1990; Moors 2010; Russell 2003) to abandon this route in favor of a view that assumes fewer fundamental affective 'building blocks'. The aim of the current paper is to show that this movement may afford new ways to integrate emotion in PP. Much of this is, as we will see, thanks to the clear evolutionary rationale that is at the core of PP. At first blush, this may seem to lead to a concept of value or emotion that seems rather alien or counter-intuitive to how we usually think about emotional value, but it will turn out to have much in common with existing theories of emotions. Most importantly, cognition (information processing) and emotion will be shown to be entangled from the start.

First a brief note about terminology. Throughout this paper, we will generally use the broader term "affect" because it does not constrain our explanandum to so-called "basic emotions", to conscious "feelings", to "moods" or to "reward". While we will suggest how to differentiate between these affective concepts in the course of the paper, we assume a basic "affective value" that is shared by all of these. It is the cause of this core of emotions that we attempt to explain. The exercise we undertake in this paper is to examine how the affective dimension of our mental life can be understood *without* positing more principles than those provided by PP. To anticipate a key thesis of this work, we propose the origin of emotion does not lie in being able to infer or predict (the causes of) bodily changes, as accounts of emotion as "perception of the body" argue (Barrett and Simmons 2015; Seth 2013). Rather, it is situated in how the brain succeeds or fails to do so over time (i.e. prediction error dynamics). We do not deny the importance of bodily arousal in the resulting emotions, but we identify error dynamics as the fundamental cause of emotions. Given that, following PP, prediction errors are ubiquitous processing products, the implication will be that emotions can emerge from any processing, not just that about the body.

## 1 Predictive Processing

PP holds that an organism is constantly, proactively predicting the inputs from its environment. Since it has no independent access to objective features in the world, all an organism can do is learn patterns in its input generated by statistical regularities in the environment and by its own actions (Clark 2013). While in principle there may be different ways in which prediction could modulate perceptual processing, PP proposes a well-defined computational scheme and a single guiding principle (Friston 2010). The scheme posits that every level of the perceptual hierarchy predicts activity in the level below, in effect explaining away input that is consistent with it such that only mismatches remain. These mismatches, called prediction errors, are sent upwards to update future top-down predictions. Much of the brain's predictive activity has a limited time frame. It predicts current input by inferring assumed causes of these inputs. The higher up in the hierarchy, the more time and space predictions can span, because they can work with regularities defined on lower levels. In this way lower level predictions model the faster changing dynamics, while those higher up track and recreate slower changing dynamics.

PP thus completely inverts the classical bottom-up view of the perceptual hierarchy. The brain actively generates the perceptual world (predictions are based on generative models, i.e., models that can generate the input), and perceptual input is in fact the feedback on how good these constructed models are. Although anatomically prediction errors are conveyed by feedforward connections, functionally they are the feedback, sanctioning the models we construct of the outside world, in line with

the view of perception as controlled hallucination (e.g., Horn 1980). The fundamental underlying principle guiding this process of iterative, hierarchical matching of predictions with inputs is that of prediction error minimization (PEM). Perception is inference to the best prediction, the one that minimizes prediction errors. Simultaneously, learning will use remaining prediction errors to home in on the best predictions for the current context, thereby reducing future prediction errors. Hence, we perceive what led to successful predictions in the past (see also Purves et al. 2011).

## 1.1 Action

There is one other, complementary way of minimizing prediction errors, which does not focus on improving predictions, but rather on modifying the things predicted, through action. In this framework, movements serve to bring the input closer to our prior expectations (often called 'active inference'). More specifically, they are induced by their expected exteroceptive and proprioceptive consequences (Friston et al. 2010), much in line with James' "anticipatory image" (James 1890) and with the ideomotor principle (Hoffmann 2003). Like object-level, conceptual predictions ("an apple") unpack to a myriad of lower-level featural predictions ("green", "curved", …), so can high-level expected states ("goals") be unpacked in specific component predictions and eventually in expected proprioceptive states. When the latter are compared to afferent signals of muscle spindles at the spinal level, they generate sensory prediction errors to be reduced by motor neuron activation, in a classical reflex arc. Hence, motor commands are replaced by expectations about the state of proprioceptive sensors. At a higher level these 'commands' stem from beliefs about state transitions (active inference). A certain perceptual stimulus may be predictive of a state transition through the agent's intervention (an affordance, if you will), that can be actualized by unpacking this prediction to proprioceptive states.

Bear in mind that, from this inference system's perspective, there is no intrinsic difference between the external and the internal milieu. With the same predictive machinery, generative models can be learned about changes in interoception, based on input from somatovisceral sensors (Seth 2013). Likewise, internal 'actions', such as autonomic responses, are brought about by similar principles as 'external' actions. They consist of changing a bodily set-point or expectation (e.g., temperature) so autonomic reflexes (e.g., shivering) can be elicited.

We limit ourselves to this very brief sketch of PP and refer to the many in-depth resources for more details about its computational mechanisms and how these may map onto neural circuits and their plasticity (Bastos et al. 2012; Friston 2003, Friston 2010). Further implications of the framework will be discussed to the extent that they connect to value and emotional relevance.

## 1.2 Prediction and Homeostasis

Organisms maintain their own organization (order) in the face of constant fluctuations in the environment through homeostasis. The bioenergetic regulation of homeostasis in essence makes sure that the organism is bounded in the physiological states it can be in, which allows it to resist the dispersive effect of the second law of thermodynamics (Schrödinger 1992). This can be considered the most fundamental goal of any organism, though of course, it is not an intentional one. It is just a result of the fact that organisms that do not tend to homeostasis will lose existence as a unit. One can view homeostasis as a limited set of *expected* interoceptive states 'discovered' by evolution, because they have proven to enable continued existence (Friston 2010; Pezzulo et al. 2015). Therefore, survival "depends upon avoiding surprising encounters and physiological states that are uncharacteristic of a given phenotype" (Friston 2009).

The key problem is that organisms generally do not have direct ways to change these internal states (Pezzulo et al. 2015). They may be able to shiver when their body temperature drops, but they cannot replenish glucose levels without interaction with their environment. This implies that they actually need to build a generative model for these expected internal states, by learning about the different

ways these internal states can be 'caused'. It is only because the organism needs to go through its environment to fulfill the interoceptive expected states, that it needs (exteroceptive) perception and action.

It needs perception to infer hidden states of affairs in the world that may cause the expected internal states. And it needs action to control those states of affairs. As hinted in the section on action above, actions are also represented as predictions or beliefs, specifically about transitions that one happens to be able to control (control beliefs). As we saw earlier, this generative model, in service of the body, is constructed with the PP scheme. The models needed for this may be rudimentary and fixed or complex and flexible, depending on complexity and volatility of the organism's *Umwelt*. For some organisms the causal chain of expected internal states might go through very high-level and context-dependent states (e.g., social interactions). At that point, it pays to build deep models incorporating this strong contextualization of interoceptive states (Pezzulo et al. 2015).

Simpler organisms may not need such strong, flexible contextualization, because internal states are reliably caused by stable states in their environment. Even in that case homeostasis is not actually static or merely reflexive. If reliable predictive information is present, it is more efficient to anticipate changes with compensatory action (Heylighen and Joslyn 2001). For example, when the single-celled gut bacterium E. coli is ingested by mammals it will respond to the temperature shift by not only upregulating heat shock genes (to compensate for temperature) but also by downregulating genes for aerobic respiration. They use the temperature information (when entering the mouth) to predict that they will end up in a low oxygen environment, i.e., gastrointestinal tract (Freddolino and Tavazoie 2012). They encode something about the cause of this particular stimulation (ingestion), however rudimentary. As Freddolino & Tavazoie (Freddolino and Tavazoie 2012, p. 369, my emphasis) describe: "microbial behaviors are as much responses to the *meaning* of a specific environmental perturbation (viewed in the context of their evolved habitat) as they are responses to the direct consequences of that perturbation".

For E. Coli the predictive "learning" of these regularities does not take place within the organism but within populations. Through natural selection the predictive environmental relation can be embodied in the molecular regulatory networks of the cell (Freddolino and Tavazoie 2012). Environmental regularities left their imprint on the organism's constitution, just because a constitution embodying these regularities increases fitness in a Darwinian sense. In analogy to PP, evolution can be considered as an error-correcting code, except that the errors are not represented at the level of a single organism[1]. But note that the normative character—the value—originates in the organism (that maintains its internal states better or worse), not in the process of evolution (Deacon 2011). The boundedness in the homeostatic set, the 'mother-value of all values' (Weber and Varela 2002) also gives the whole predictive endeavor its normativity. Once the organism engages itself to make a prediction, there is something at stake (cf. value), because of the link from the quality of predictions to basic organismic functioning. There is a vested interest for the prediction to materialize. In complexer animals, "the gross bodily form, biomechanics, and gross initial neural architecture of the agent all form part of the (initial) 'model' and […] this model is further tuned by learning and experience" (Friston et al. 2012a), using the general-purpose PP mechanisms. Reducing prediction error can be a proxy for fitness, because prediction error minimization is the proximal, local mechanism that makes sure that in the long run organisms stay within physiological bounds (Friston 2010).

## 2 Prediction and Reward

The fact that predictive models are grounded in homeostasis by no means implies that only predictions about favorable outcomes can be formed. For example, even though some perceptual predictions may seem not to be consistent with biologically expected states (e.g. "the rattling that I hear is caused by a snake nearby"), it is all the more important to make them accurately, and not hallucinate more

1 Recent work suggests there is a deeper, formal equivalence between PP and evolutionary population dynamics (Harper 2009).

agreeable alternatives. A negative stimulus only has really detrimental consequences for survival if the system was not able to adequately prepare for it, by marshaling the necessary compensatory mechanisms—if necessary by acting to avoid it. Once this is taken care of, what could be a threatening stimulus for bodily integrity, becomes harmless. Conversely, predicting (and hence preparing for) a future negative stimulus that turns out not to occur, is often very wasteful for an organism. So we see that there are good biological reasons for why prediction confirmation should be good, while failures should be bad. In fact, even an appetitive stimulus such as food could, for an unprepared body, be unpleasant.

In a PP account, reward stimuli are just expected or familiar sensory states (Friston et al. 2012b). Intuitively we feel that we avoid punishment or seek reward and therefore visit these states less or more frequently, respectively. PP turns this intuition around, describing frequently visited states as rewards because they are expected. The reward value of a stimulus can be defined as the frequency with which it is chosen (Moutoussis et al. 2015). Rewards do not "attract behavior", but attainment of rewards is the result of prediction error minimization, exactly as described for perception and action in general. Specifically, while in classical reinforcement learning goal-directed decision making consists of finding the policy that maximizes expected reward, when framing it in terms of Bayesian inference one assumes reward attainment and finds the policy (state-action pair) that best explains or causes that effect (see Schwartenbeck et al. 2014; Solway and Botvinick 2012; Srivastava and Schrater 2015). If one redescribes utility of outcomes as prior beliefs about states one will end up in, one can use the same PP machinery to minimize errors along the road to the expected state. This boils down to building a generative model of rewards (same as for any other stimulus). Importantly, it requires that we have prior beliefs about what the world will be like and about expected final states or goals (Moutoussis et al. 2014). The latter are the alternative outcomes that we expect to be reachable with policies we can apply. The key is to reduce the discrepancy between the likely and the expected outcomes. Note that within this approach, one could still make a distinction between "greedy" and epistemic actions (Friston et al. 2015). Greedy or pragmatic actions use prediction errors to directly fulfill expected "rewarding" states, e.g. food consumption. This is possible when there is little uncertainty about the path leading up to the expected state. In case there *is* considerable uncertainty, epistemic actions are directed at acquiring more information, that allows greedy action in the future. This implies that action may lead to increase of the distance to a goal (prediction error), in order to move to a familiar position where one can approach the goal with larger certainty (Friston et al. 2015). However, any actual behavior will have a combination of pragmatic and epistemic elements, with prediction errors as the common currency.

The repertoire of innate expected states is specified and extended by learning throughout an animal's life. In fact, within this view, there are no distinct reward or punishment stimuli (Friston et al. 2009; Wörgötter and Porr 2005). Any sensory signal has a cost, namely the prediction error. It tells something about the success (failure) of the generative model we used for predicting the input. This also implies that habits or 'rituals', i.e., predictable sequences of behavior, are in fact a form of reward. There is usually no tremendous pleasurable experience to habits (we will come back to this point later on), but not performing habits when the appropriate eliciting context is present seems to produce some negative affect. It speaks to the self-sustaining nature of habits (Egbert and Barandiaran 2014; Egbert and Canamero 2014). Indeed, for over-learned behavioral patterns, devaluation of the reinforcer that was originally used to establish the behavior will not lead to reduction in behavior (Wood and Neal 2007). The wider implication is that organisms do not only preserve their life (homeostatic predictions) but also their *way of life*, as a set of expected (preferred) behaviors (Di Paolo 2003).

## 2.1 Problems with the Classical Reward Concept

Several developments started eating away at the concept of reward as absolute, stable representation of utility, that guides decision making. We highlight two theoretical problems, and two empirical ones.

First, Friston et al. (Friston et al. 2012b) criticize the inherent circularity in the definition of reward. Reward is often defined as a stimulus that elicits (reward seeking) behavior (Schultz 2007). Evidently, one cannot invoke rewards to explain that same behavior later on. Second, recent theorizing suggests that rewards and punishments are always subjective and internal, meaning they "are constructions of the subject rather than products of the environment" (Dayan 2012, p. 1089). They are dependent on the position relative to expected states. Reward is not something in the environment, much less an external critic such as often assumed in computational reinforcement learning (see a similar critique in Singh et al. 2009). Psychological theories too, often incorporate a semi-hidden homunculus. Here, the value (or cost) function applied to perceptual or cognitive output hides an 'evaluator', an unanalyzed 'agent' that can assign the values, within an allegedly objective, quantifiable construct. Misled by our intuition that rewards are self-evident, these homunculus remnants too often go unquestioned.

Third, Chater & Vlaev (Chater and Vlaev 2011) argue on empirical grounds that, similar to sensory judgment in psychophysics, value is a not represented as an absolute magnitude but rather as a comparison, relative to the local context. Chater & Vlaev conclude that "to the extent that people have a grasp of their own, more global, values, this must be inferred from sampling their own past choices and other memories, thus revealing their preferences" (Chater and Vlaev 2011, p. 96). In other words, humans can easily infer the reward value based on experience sampling, but these values are constructed predictions that best explain the sampled experiences. Generally, however, these representations are not necessary to enable adaptive behavior.

Finally, after conditioning reward-modulated activity is found throughout the visual hierarchy, including the primary visual cortex and the lateral geniculate nucleus (Gershman and Daw 2012; Serences 2008). Conversely, "neutral" perceptual prediction errors elicit activity in striatal and midbrain regions, usually connected to reward/punishment and motivational functions (e.g., Schiffer et al. 2012; Iglesias et al. 2013; den Ouden et al. 2009). Other studies find that dopamine neurons also code for sensory uncertainty in a rewarded sensory decision-making task (de Lafuente and Romo 2011), possibly because expected reward will decrease when sensory uncertainty increases (Bach and Dolan 2012). In sum, these developments indicate that the strict segregation of probabilities (perceptual processing) and utilities (cost-reward processing) is untenable (Gershman and Daw 2012), and suggest that the PP concept of reward merits further examination, because it does not suffer from these shortcomings.

## 2.2 Pleasant Surprises and Other Objections

The PP framing of reward does not mean that learning or behavior is not as constrained as in conventional models of reward and punishment. To take the extreme example, even if at the agent-level a pain stimulus is perfectly expected, across all levels of predictions this will never become an expected state. Tissue damage can be seen as a violation of a bodily expected state (bodily integrity) that is not compatible with continued existence. On the other hand, this approach has no difficulty explaining why humans seem to find reward in endlessly varying idiosyncratic 'niches', based on the wide flexibility in predictions they can generate.

An obvious counter-argument to the thesis that prediction errors always have a negative value is the existence of 'pleasant surprise', e.g. when one receives an offer that is better than expected. However, even in such cases there is some evidence in humans and monkeys that the initial reaction to prediction error or surprise is generally negative (however short-lived) (Knight et al. 2013; Noordewier and Breugelmans 2013). But note that the agent-level emotion of surprise encompasses more than a single, momentary prediction error (surprisal) at some level of the brain. As we will see in later sections, the dynamics of the failures and successes in prediction are more important here.

Still, intuitively, the identification of prior probabilities (predictions) with utilities seems wrong-headed. For example, Gershman & Daw (Gershman and Daw 2012, p. 306) ask: "Should a

person immersed in the 'statistical bath' of poverty her entire life refuse a winning lottery ticket, as this would necessitate transitioning from a state of high equilibrium probability to a rare one?" To start to defuse this argument, one has to acknowledge that in such complex cases there is not just one prediction (e.g., of poverty) at play, but rather a complete predictive hierarchy. The person growing up in poverty does not lose his or her expectation to be well-fed and to provide for kin. There might be interiorized social expectations as well, that could also urge the person to accept the winning lot. That said, once accepted, the new situation may create quite some prediction errors given a predictive system unadapted to that new state of affairs (indeed, most lottery winners like to continue their life, including job, as before; Kaplan 1987). Later sections will hopefully shed a different light on these forms of 'upward mobility' (Gershman and Daw 2012).

A related counter-intuitive idea in this proposal is the lack of distinction between core concerns, desires, needs or goals of an organism versus just any predictions. Part of the answer has to be found in the hierarchy, with likely states being about what happens when I see (or do) this. In contrast, desired states are about what I can, more abstractly, expect given my biological constitution, experience and sensorimotor capacities (i.e., predictive models)[2]. For example, while I, at some level, expect a food reward (desire), the way to "generate" this involves lower-level sensorimotor predictions about states and state transitions (some of which we control), that are navigated through by PEM. The assumption is that some expectations (e.g., about homeostatic states that, for example, a food reward can fulfill; see the section on homeostasis) are hardwired, installed by evolution. So another part of the answer must lie in the lower flexibility of goals/desires, compared to sensorimotor predictions which are flexibly updated when inputs change. However, because the whole hierarchical model is relevant to the organism (if it was not it would not have been formed), any prediction has cognitive (belief) and conative (desire) elements which cannot be disentangled (Millikan 2004). Predictions are never motivationally neutral, but represent states and direct behavior at the same time. More work will be needed to explain our intuitive distinction between likely and desired states if both are ultimately predictions, but the hierarchy is bound to play an important role. The lower level likely states mostly pertain to faster changing dynamics in inputs (regularities in shorter time frames). The higher level desired states link to slower changing dynamics (e.g., 'I am a good person'). If evidence mounts that undermines the latter type of predictions, a full-blown existential crisis may occur. Luckily, there are often ways to shield such prediction errors, i.e., to explain them away with 'auxiliary predictions'.

At this point one might object that reward (or positive affect) as defined so far is too 'conservative' a concept[3]: we basically aim to return to familiar, overlearned states or situations and resist anything that deviates from those expected states (but see dissonance or conflict theories of emotion; Mandler 2003; Festinger 1962; Hebb 1946). This approach may explain the familiarity bias (cf. the mere exposure effect) that is often reported (Lauwereyns 2010), but it does not even remotely seem to capture our *experience* of reward in general. We easily get bored—a loss of reward value—with very familiar or repetitive stimuli. More so, we seem to actively explore departures from well-trodden paths and expected situations. How do we explain that our motivations often lie outside of predictable ruts? And how can we more fully account for rewards as hedonic, pleasurable experiences derived from these different situations? That is what we will discuss next.

## 3 Affective Valence

If we define drives as prediction errors or discrepancies between current and expected state (Keramati and Gutkin 2011), we end up a new way of looking at the affective, experiential value of rewards as

2 At least in humans this seems to have an important social comparative component as well: our predictions are formed based on what people that one considers to be similar to oneself, could attain.

3 But it is far from a passive notion: to keep the organism within some expected range of a variable often means elaborate and vigorous activity (cf. allostasis; Egbert et al. 2013).

well. Reward value is directly dependent on drives in the sense that the reward value of, say, a drop of water depends on the internal drive state of the organism (e.g., a thirsty rat). It is easy to see that what is critical then is the change in prediction errors (drive states). Hence, positive value is defined as a decrease of prediction errors, while negative value can be equated with an increase in prediction errors.

More generally, we propose that the affective valence is determined by the change in (or first derivative of) prediction error over time (Joffily and Coricelli 2013; Van de Cruys and Wagemans 2011)[4], with positive valence linked to active reduction of prediction errors, and negative to increasing prediction errors. This makes sense because these temporal dynamics signal whether the organism is making progress (or regress) in predicting its environment, which in the long run translates in proper functioning of the processes of life (fitness) (Damasio 2003). It is easy to see that the reward value of food very much depends on how large the prediction error initially was (i.e., how hungry you were), and hence how big a change the food consumption induced, but we propose this is a general pattern.

Importantly, emotional valence is not something added to these error dynamics, it is those dynamics. They are a reflection of quality of processing, so they do not have to be evaluated in turn. We connect positive and negative affect here to general purpose processing characteristics, detached from particular utility or motivations. They are purely determined by how the organism interacts with its environment (see also Polani 2009).

This goes beyond the simple view that prediction confirmation results in positive affect, while violations of predictions are negative. Once homeostasis, rather than being reactive, relies on predictive models, errors often do not have direct effect on homeostasis (or fitness). It then becomes equally important to monitor prediction error dynamics, as it is to monitor the errors as such. Mere presence of instantaneous prediction error does not seem to be an adequate basis of emotional valence. Positive affect might still occur for a large instantaneous error as long as this error is (or has been) in the process of being reduced.

It is no stretch for humans to imagine that making progress in predicting various sensorimotor domains can be very rewarding (e.g. see Hsee and Ruan 2016). More challenging is to show those 'informational' rewards in nonhuman animals. However, Bromberg-Martin & Hikosaka (Bromberg-Martin and Hikosaka 2011) have managed to show that monkeys too are prepared to work to receive cues that reduce their uncertainty (reduce errors), even though their choice had no influence whatsoever on the actual reward subsequently received. The animals even chose the information cue more consistently than they typically choose a high probability reward over a low probability reward (Niv and Chan 2011). Moreover, these informational gains elicited dopaminergic neural activity in midbrain regions similar to that for conventional rewards. Our account would predict that such effects generalize to other animal species, but of course, for there to be changes in prediction errors there need to be predictions formulated. Therefore, the specific instances of predictive gain will depend on the kind of models an animal constructs about its world.

Behavioral testing of these ideas is challenging because these dynamics are subject to learning and because it can be difficult to determine the predictions participants apply. Suggestive evidence comes from a recent study looking at the affective consequences of conflict resolution (Schouppe et al. 2014). These authors build on the priming study by Dreisbach & Fischer (Dreisbach and Fischer 2012) which showed that incongruent Stroop stimuli, as opposed to congruent ones, can prime people to more quickly evaluate negative words or pictures than positive ones (an indirect measure of negative affect). Schouppe et al. (Schouppe et al. 2014) report that, while incongruent stimuli are indeed aversive, once they are successfully solved more positive affect will follow than for congruent stimuli. The original prediction error (conflict) seems conducive to later reward from resolution, consistent with what we

4 Note that the model by Joffily & Coricelli strictly speaking is not about prediction errors but rather about the more general concept of (variational) free energy.

propose here. Future studies should attempt to induce, violate and resolve new predictions in the lab to see if these dynamics have the hypothesized emotional effects.

### 3.1 Specifying Predictive Progress

Even though the current view entails that emotions can arise wherever errors are compared, there are good computational and ecological reasons why change in errors is computed and compared within the limits of one and the same input domain. Comparing errors from very different perceptual levels or sensorimotor situations would be very demanding to the system, and, more importantly, unproductive. As Oudeyer, Kaplan & Hafner (Oudeyer et al. 2007, p. 8) remark with regard to an artificial agent, such a system may "attribute a high reward to the transition between a situation in which a robot is trying to predict the movement of a leaf in the wind (very unpredictable) to a situation in which it just stares at a white wall trying to predict whether its color will change (very predictable)." PP proposes that specialization (functional segregation) in the brain stems from conditional independence of different representations—representations that have predictive relations organize into regions with tight interconnections (Friston et al. 2013; Stansbury et al. 2013). This architecture may also be used to evaluate changes in errors relating to predictions that actually belong to the same domain.

Predictive progress has already been used to understand and implement intrinsic rewards in the domain of artificial intelligence (Kaplan and Oudeyer 2007; Schmidhuber 2010). More recently a decrease in prediction errors (or equivalently a predictive learning gain) was assumed to underlie intrinsic rewards in humans as well (Kaplan and Oudeyer 2007). Agents that at each point try to maximize predictive progress, will avoid losing time in regions of sensorimotor space that are too difficult to predict with the current capacities and regions that do not contain any learnable differences anymore, either because the domain is known or because what is left is noise variation. Hence, they will automatically focus on situations and stimuli that contain learnable differences, just above their current state of predictive knowledge, where the largest gain can be made. This guiding principle enables the agent to explore and proceed through stages of increasing predictive difficulty ('developmental phases').

There is some debate about the extent to which such an imperative to maximize prediction error reduction and PEM are one and the same thing (Clark 2013; Froese and Ikegami 2013; Little and Sommer 2013). Proponents of the 'maximizing learning gain' position contend that an organism driven by PEM will seek a dark room and stay there, because prediction error is maximally reduced there. However, a dark room is not actually a maximally expected situation, or does not stay so for long, in a PP framework (see also Friston et al. 2012a). Prediction errors are always computed relative to an agent's possibly complex, embodied model, with its specific organism-defining expectations, quickly rendering the dark room unexpected. While this seems to answer the 'negative' objection (why not stay in the dark room), can PEM also fully account why we humans 'positively' seek out prediction errors? This seems to depend on the kind of multi-level and second-order predictions we generate. As an example, if, at an abstract level, you expect yourself to be friendly, confirmation of this prediction will sometimes entail prediction errors on other, possibly lower levels. The key is to predict the violations as well, such that their impact can be reduced (see discussion on precision above). Similarly, if you expect to be a good darts player, you will need to tolerate some lower level sensorimotor errors to get there, usually because you can also reasonably expect the errors encountered to be reducible, based on previous experience. In short, a good predictive agent will always expect to be surprised. We seek prediction errors that are reducible, given our models, including the actions (as beliefs about inputs we control) we can rely on.

## 3.2 Non-Conceptual Metacognition

An operation performed on the prediction errors can be considered a form of metacognition. Similar to precision, the temporal comparison of prediction errors is a second-order operation. In the first-order process, prediction errors are information used to update predictions, while in the second-order process the prediction errors as outputs of the first order process are in turn compared in time, which provides new information that, we argue, is phenomenally experienced as valence and that may become available for processes beyond the predictive chain that created the errors. The result is a form of nonconceptual information about uncertainty that increases or decreases in the current situation. It is not about the (propositional) content but about the content-forming processes. The thesis here is that emotions are the qualitative experience (quale) of this kind of nonconceptual information. In a related view, Reisenzein (Reisenzein 2009)[5] argues that emotions non-conceptually convey important changes in an experiencer's belief system in interaction with the world. This is indeed what prediction errors signal. Their dynamics are a form of feedback on the system's own functioning as it deals with external and internal challenges, so a conception of affect as a continuous "neurophysiological barometer of the individual's relationship to an environment at a given point in time" (Duncan and Barrett 2007) is nicely consistent with this. Similarly, Frijda (Frijda 2006, p. 82) notes: "pleasure is the positive outcome of constantly monitoring one's functioning". For affect and motivation, the attainment of the "object" is of lesser importance, considering that predictive progress is zero then. This provides an interesting perspective on what Cantor and Kihlstrom (Cantor and Kihlstrom 1987, p. 179) called the paradox of goal-setting, namely "that people are often less intrigued or impressed with an end-state the closer they come actually to achieving it".

In the current view, the non-conceptual information is available in terms of the positive or negative affective tone of experiences. Note the connection with the concept of cognitive or perceptual fluency (Reber et al. 2004) as the ease with which stimulus material is processed. In its different operationalizations (e.g., by increasing the symmetry and contrast of visual stimuli, or the readability of words), it has been repeatedly shown to positively affect the appreciation of stimuli. Fluency should also be seen as a metarepresentation (Alter and Oppenheimer 2009) and is arguably well characterized as the experience of actively reducing prediction errors (and disfluency as increases in prediction errors). Moreover, if one identifies emotion with the way of processing rather than end-products, perceptual (dis)pleasures and 'proper' emotions might be subsumed under the same principles. Specifically, (dis) fluency with regard to approaching high-level goals or biological concerns (bodily expected states) is what we usually associate with emotions. This idea is barely new. In a very influential control-theoretic approach to emotions, Carver & Scheier (Carver and Scheier 1990) linked dynamics in mismatch between goals and actual state of affairs to dynamics in emotion. They described how multi-level goals should be interpreted as hierarchical reference values, from abstract idealized goals (e.g., having a self-image of a good person at the highest level), to more concrete actionable expectations (e.g., shoveling snow off of walks). In PP terms, actions have to make sure that the agent can harvest the inputs that conform to "trickled down" expectations. So, analogously to PP, these expected values can generate errors at every level. Our own actions (or external circumstances) cause changes over time in discrepancies relative to these values. Carver and Scheier already argued that emotion is about monitoring the rate of discrepancy (prediction error) reduction, as we propose above. However, their analysis suggests a pertinent extension of what we presented so far. They suggest that the rate of mismatch reduction is in turn subject to a control loop, comparing actual with expected rate of change. Only when the current rate of prediction error reductions deviates from the expected rate of reduction, so Carver and Scheier argue, one experiences emotion. This will of course be positive affect if the rate of progress to the goal is higher than expected, negative if it is lower than expected.

5 Note that in Reisenzein's theory desires still have a status categorically different from other beliefs.

Based on PP, this makes a lot of sense. As we described, prediction error minimization is the way we perceive and act, so we are reducing errors all the time, e.g., when we successfully use our sensorimotor system to walk the street. Generally, little positive or negative emotion is involved despite these constant error reductions. This may mean that these changes in sensorimotor errors are not large enough, but most likely what rate is substantial depends on the expected rate of reduction for the current sensorimotor context. Where do the predictions of rates come from? These might very well be contextually learned through the same predictive machinery as for 'first-order' predictions. In fact, PP already includes second order expectations about precisions, which can be considered as (inverse squared) expected prediction errors (Mathys et al. 2014). The higher order prediction errors that are used to update these expected precisions (so-called volatility prediction errors or VOPEs) compare predicted prediction errors (predicted uncertainty) with observed prediction errors (actual uncertainty). For example, if first level prediction errors (also called value prediction errors or VAPEs[6]) are reduced, but the corresponding VOPEs do not decrease, the error continues to be lower than expected, possibly providing a basis for positive affect. It still needs to be clarified to what extent the temporal derivatives and expected rates as described here can be realized using expected volatilities per se (Joffily and Coricelli 2013), but these developments at least suggest learning about such second order states is possible within a PP system.

However, genetic factors might also contribute to these expected error reduction rates. Individual differences in expected rates of error reduction may account for certain dispositional affective traits. Indeed if the predicted rate of progress is set too high, an individual will tend to experience more negative affect than positive, because the prediction will rarely be matched (Carver and Scheier 1990). This may happen, even if this person's actual rate of progress is very high. Furthermore, if the expected rates of progress are indeed at least partly learned specifically for different sensorimotor situations, this may constitute a form of emotion regulation. Specifically, the system may, through updating the expected rates, remain within a given range of emotional experience by adapting this criterion of expected rate of change (the neutral point).

Interestingly, once an agent can track and learn to expect certain rates of change in prediction errors, it arguably will show a distinct propensity to explore and learn. This continuous, active search for reducible prediction errors (satisfying an expected error reduction rate) may in turn have enabled the development of rich social relations and culture. As such, this may form another counter-argument for the dark room objection against the principle of PEM: There will never be a stationary stimulus or situation satisfactory for an agent that expects some non-zero rate of prediction error minimization.

## 3.3 Varieties of Affect

Looking back, we have first encountered reasons to attribute emotion to prediction errors (mismatch) or confirmation as such, then we have shown it may be better attributed to changes over time in prediction errors, and finally to errors about expected rates of change. Importantly, these three can be independent. Borrowing an analogy from Carver & Scheier (Carver and Scheier 1990); if we make the parallel with distance, speed (first derivative of distance over time) and acceleration (second derivative), we can see that any rate of progress can be associated with any instantaneous prediction error, and further any change in rate of progress can co-occur with any instantaneous rate. The rate of error minimization seems to provide the necessary signal for valence. However, in mammals, especially humans, rate may be subjected to predictions of its own, moving important emotional dynamics to that level. Still, rate may determine the continuous hedonic tone of what is sometimes called 'background emotions' (Damasio 2000). A steady rate of progress may induce a diffuse feeling of well-being, a sense of properly functioning bodily and sensorimotor systems, akin to what is sometimes described as experience of flow (Csikszentmihalyi 1996). The usually brief episodes of (intenser changes in) emotions

6 This is not about emotional value, but about a quantitative mean value of the estimated state.

in our daily life, i.e., emotions as commonly understood, seem linked to unexpected changes in rates of progress.

Emotions are notoriously volatile, comparative, and subject to habituation (Frijda 2006 [1988]). These characteristics naturally follow from the current framework. By definition, prediction errors and their temporal dynamics are dependent on learning. Pleasures from increased rates of predictive progress only last as long as this progress is possible. Kaplan & Oudeyer (Kaplan and Oudeyer 2007) note that "progress niches are nonstationary". Meanwhile, the contrastive property of emotion entails that a suboptimal state (sizable prediction error) may still be pleasurable depending on the starting position, because there is a positive, possibly higher than expected, rate of error reduction. Emotions emerge as situated in perpetually moving regions of state space in a system that grounds them in predictive dynamics. Rather than being associated with particular "target" objects or states, they are the concomitant of successful (or unsuccessful) striving (Duncker 1941). Note that this type of system does not aim to maximize the frequency of positive affect (nor would that be particularly adaptive; Carver and Scheier 1990). Rather, it may redistribute frequencies of positive and negative affect so as to preserve the range.

One might object that the view we propose runs the risk of 'intellectualizing' emotions. Indeed we essentially described affect as a specific form of cognition. In that sense, it is somewhat related to previous emotion theories about the 'need to resolve uncertainty', most aptly formulated by Kagan (Kagan 1972). But conceptualizing this as the 'wish to know' (Kagan 1972) or the 'need for cognition' (Cacioppo and Petty 1982) seems to suggest that this capacity is aimed at finding out some 'ground truth' (and exclusive to so-called higher animals). Knowledge captured in the predictive models is always subjective and constructed (Heylighen and Joslyn 2001), i.e., the agent has no direct access to the 'real world', but can only 'negotiate' its conditions by actively predicting (constructing) its characteristics. As von Glasersfeld (von Glasersfeld 1995) stated we meet the world only in our failures. In contrast to these existing related approaches, ours centers on prediction errors, rather than *any* uncertainty, and more specifically their dynamics (rather than static uncertainty). Still, our account underscores the role of uncertainty and unpredictability in emotion and motivation (Anselme 2010; Jackson et al. 2014; Whalen 2007). For example, rats seem more motivated to work for a reward in conditioning experiments that introduce some uncertainty in the predictive link between conditioned stimulus and unconditioned stimulus (reward) (Anselme et al. 2013). Conversely, the exacerbating effect of uncontrollability[7] and unpredictability on stress and anxiety is well-documented (Hirsh et al. 2012; Mineka and Hendersen 1985).

Error dynamics are common to all processing, be it interoceptive, exteroceptive, abstract goal-related or low-level sensorimotor. This may better account for the very broad range of situations that can engender positive or negative affects. For example, apart from biologically relevant things, positive emotions may be experienced from scary movies, abstract perceptual stimuli (e.g., in art), acquired tastes such as piquant foods (Rozin and Kennel 1983) or painful stimulation such as masochistic pleasures (Klein 2014). These instances may be difficult to explain from the viewpoint that pleasure is only attached to biologically instrumental situations (or appetitively conditioned stimuli). Below we review those emotions, subtle and intense, that are usually considered to be atypical, for that reason. We try to show that, when taking into account the error dynamics relative to (learned) expected states, they are very representative emotions.

### 3.3.1 Intrinsic Pleasure and Curiosity

Development is a rich source of emotions. For example, the baby that wants to keep on playing peek-a-boo (Parrott and Gleitman 1989) till predictions of object constancy are fully formed and the situation

7 In the current account the distinction between unpredictability and uncontrollability largely dissolves—actions (to exercise control) are predictions as well, with concomitant expected levels of prediction error decrease or increase.

contains virtually no dynamics in prediction errors anymore. Or the child that is excited to hear the same bed-time story again and again, until errors are driven down by learning its structure (not only of the plot but also of lower level sensory patterns, as is for example clear from toddlers' preference for repetitive rhymes). Later in life, emotion theorists emphasize the centrality of the emotion of interest, for development and beyond (Izard 2007; Silvia 2001). The two factors that have been shown to determine interest can easily be translated to our approach, arguably gaining some specificity in the process. First, only new, unexpected or complex stimuli ("novelty-complexity appraisal") can elicit interest (Silvia 2008), implying that prediction errors are required[8]. The second factor needed to evoke interest is roughly described as comprehensibility or coping potential (Silvia 2008), an appraisal of one's capacity to deal with or understand the (unexpected) stimulus. In our terms, this would be an expectation of a positive rate of error reduction for the current sensorimotor context. We continually, implicitly probe our coping potential by predicting performance (sensory consequences of actions) and computing errors. In fact, making progress (actively reducing errors) in predicting a certain activity domain would be a good indicator of adequate coping potential in this domain in the near future. Hence, the importance of expected rates of progress. In this way, important elements of appraisal theories of emotion can fit within this PP account (Ellsworth and Scherer 2003; Moors 2010).

One of the most influential views on curiosity and exploration is Berlyne's optimal level account (Berlyne 1970). He argued that organisms seek out stimuli with medium level complexity or novelty, to keep their arousal at an optimal, pleasing level. This preference for optimal level of complexity is corroborated in experiments with infants that looked longer at stimulus items that were neither very simple nor very complex (Kidd et al. 2012). Rats too, prefer to spend time in arms of a maze of which the patterns on the walls were slightly more complex relative to the walls they preferred earlier (Dember et al. 1957). The latter studies emphasize the crucial role of experience, which can lower complexity (increase predictability). We would argue that organisms are very much tuned to reducible uncertainty in input. They explore stimuli with medium levels of prediction errors, because they predict a positive rate of error reduction in these inputs. Indeed, they have had experience of error reduction with slightly simpler but similar inputs. In agreement with this, 18 month old children already attend longer to learnable compared to unlearnable linguistic grammar, strongly suggesting they make good estimates of their future predictive learning progress (Gerken et al. 2011).

In adults, through experience, these dynamics, and the pleasures or displeasures derived from them, are not so much situated on the purely perceptual level, but rather on the conceptual level, e.g., stories, jokes or soaps. Although a complete treatment of social emotions will not be given here, observe that they often involve a convergence or divergence in opinions or 'worldviews' (expected states and beliefs). We make models of ourselves and others, like we do for the rest of our environment (Moutoussis et al. 2014), so similar error dynamics are in play in this context. Moreover, the rewarding sense of (em)power(ment) can be interpreted as the result of actively bringing about anticipated sensory effects through action execution (Polani 2009). This seems consistent with our idea that rewards derive their rewarding capacity from reductions in pre-existing prediction errors. Beneficial, motivational effects of a sense of control ('mastery') (Klein and Seligman 1976) may similarly be explained as positive affect from a high expected rate of error reduction.

### 3.3.2 Humor

In general, the positive emotional mark on unexpected progress towards predicted states is stronger (than just progress). This is consistent with the view proposed here, that a higher than expected rate of error reduction determines positive emotion. This is best illustrated in laughter. In a poignant analysis, Sroufe & Waters (Sroufe and Waters 1976) observe that laughter results when a rapid, maximal tension build-up is followed by a rapid 'release' or 'recovery'. The ill-defined term 'tension' was often used to

8 Note that complexity is also dependent on predictability.

denote some incongruity in perceptual input, assumed to cause some negative arousal. Of course, prediction error can take its place, gaining not only specificity, but also integration in a plausible theory of cognitive processing. More important to stress is that a steep, sudden gradient of prediction error will lead to a prediction of low rate of error reduction. If errors can in fact be reduced, e.g. through an appeal to different predictions (restructuring of input), the reduction rate will be much higher than expected, resulting in intensely positive affect (laughter). This is the typical processing profile, not only for peek-a-boo-like fun in children, but for instances of humor in general (Rozin et al. 2006). Consistent with our approach, both the gradients and the unexpectedness are crucial. In earlier work (Van de Cruys and Wagemans 2011), we developed a similar account for "aesthetic emotions", where artists allow for unexpected increases in error reduction rates for greater appreciation, consistent with the emphasis on *relative* fluency in recent experimental psychoaesthetics (Topolinski and Reber 2010; Wänke and Hansen 2015).

## 4 Feelings and Function

Our account implies a fundamental misattribution in emotion. The intuition that emotions are entirely caused by objects in the world is misguided, because in fact they are linked to processing characteristics (see also Reber et al. 2004) rather than content of processing itself. The evaluation of error dynamics seems to provide a parallel (affective) dimension to experience, that is not strictly linked to the content (predictions) or particular prediction errors taking part in those dynamics. Still, the specific and diverse forms emotions can take, seem largely dependent on the conceptual context (sensorimotor or cognitive domain) in which the error dynamics appear. But such attributions are always constructions, they will never be directly about what caused the emotions, the (changes in) error reduction rates. One might argue that conscious emotions (or feelings) thereby acquire an intentional object or propositional content, but this does not seem to be a strict requirement, as the existence of conscious moods illustrate. As a related side note, Picard (Picard 2013, p. 2496) reports on two patients experiencing a feeling of intense bliss during epileptic seizures originating in the insula. One patient describes the experience: "...all the ordinary facts about the environment seem suddenly to become infused with certainty and a sense of inevitability... The sense that I had when I was experiencing some of these seizures was not unlike a continuous series of profound "a-ha!" moments. [...] Instead of merely being justified by one or several other considerations or observations, [my beliefs] seemed to be irrefutably supported by literally everything in the world." Such reports suggest first that certainty (lack of prediction errors) is an important factor in bliss, and second that those affective experiences can happen without concrete object or propositional content. Both conclusions are very much in line with the current view. This raises the empirical question of whether the insula, which is known to be both involved in uncertainty or risk processing and in emotions (Singer et al. 2009), might be responsible for some of the computations on error (or uncertainty) dynamics that we propose underlie the reports of intensely positive aha-experiences.

In the current view, the full-fledged (semantically rich), conscious emotions, in all their heterogeneity and object-directedness, appear as a form of construction, a "making sense of" underlying affect (Barrett 2014; Russell 2003). This could already be seen as a form of coping, a reaction to affect: categorizing or labeling an emotion to make it predictable (explaining away errors). The underlying affect consists, in our reasoning, of the changes in error reduction rates. The first-order errors that determine these dynamics can be multisensory (combined interoceptive and exteroceptive). Conscious feeling will amount to finding predictions that best explain the co-occurrence (regularity) of situational context (exteroceptive input) and bodily states, together with changes in rates of error reduction (second-order). Hence, the intentional content of feelings is the product of inferences, but the generation of emotion lies in error dynamics. Differently put, if emotions are categorized, a kind

of understanding is attained, which explains away part of the unexpected changes in errors, hence removing some of the emotionality.

To give one simple but concrete example of how a similar emotion might result in different feelings: an unexpected increase of prediction errors may be associated with both fear and shame (given that both are negative emotions). But on the basis of the different conceptual, situational context (e.g., shame probably concerns internalized social expectations, fear not necessarily so) they are differently interpreted and experienced. Interestingly, just by conceptualizing it as shame, one might activate coping strategies that in related situations helped returning to more expected states. The shame prediction for this constellation of inputs is predictive of certain actions or thought strategies that are in turn predictive of a reduction of prediction errors (e.g., actions to restore one's reputation with others).

## 4.1 Functions of Affect

The computations outlined above should be understood as building a model about how uncertainties evolve in the current context. It seems plausible that these models are crucial in guiding choices (implicitly or explicitly) about whether to continue to engage with the current sensorimotor activities or whether to disengage and switch. Specifically, (unexpected) decreases in prediction errors should raise predictive engagement, in line with how emotion motivates us to remain involved in activities. Confidence[9], as a rather emotional form of metacognition (Chetverikov and Filippova 2014), also seems to stem from these processes. In contrast, (unexpected) increases of prediction errors may change the balance in favor of performing actions to control input (conform to predictions; assimilation), instead of a continued search for revised predictions (accommodation). An action to avoid perceptual input (by averting the eyes), or even a mental switch to leave a certain way of thinking can also be ways to (temporarily) return to a more expected rate of error reduction.

This is where arousal and action tendencies, often considered to be core components of emotion (e.g., Frijda 1987), come in. Rather than being causally constitutive components, we would put them at the output side. If, as we argued, emotions are caused by (unexpected) changes in prediction errors, these computations indeed seem especially important in tipping the balance from updating predictions—a strategy that may be inadequate when confronted with increasing, precise errors—to acting to change the things predicted. Arousal is then derived from such action preparations. Of course, dynamics in autonomic and action-related prediction errors can give rise to emotional valence as well, given that they are governed by the same PP principles. In fact, it seems that the closer to action or autonomic responses these error dynamics are situated, the more intensely negative or positive emotions induced by these dynamics are. This may, however, have more to do with the precision of the predictions than with discrete differences in weight or importance in these predictive systems.

The brain predicts external stimuli in service of the body. It allows anticipation of what the body will need in terms of resources. Hence, it is important to accurately represent bodily states and their causes (Hohwy 2011). However, just recruiting bodily resources, or representing bodily states (and their causes) isn't emotion. If the body perfectly predicts the need for resources based on external input and prior knowledge, there can be bodily activation (arousal) without much emotion. Again, momentary prediction errors do not imply much, it is the changes in (in this case somatovisceral) prediction error, especially the unexpected ones, that should lead to notable emotion. So, while we agree with the models by Seth (Seth 2013), Gu, Hof, Friston & Fan (Gu et al. 2013) or by Barrett & Simmons (Barrett and Simmons 2015) that hold that emotions have to do with somatovisceral prediction errors, we stress that those accounts may not sufficiently explain the *causes* of emotions. The distinction should be clear: those accounts argue emotion is exactly like perception except of somatovisceral instead of exteroceptive inputs. Emotion is then inference to causes that explain (generate)

9 We mean confidence as the common sense personal-level phenomenological construct here, not the subpersonal computational concept.

somatovisceral inputs. We do not deny that these somatovisceral models are constructed, but focus on the dynamics in discrepancies of bodily state as the causes of emotions. The origin of emotion lies not in being able to infer or predict (a cause of) bodily changes, but rather in how we succeed or fail to do so over time (error dynamics).

Apart from actually instigating action (preparation), another function of negative affect (increasing rates of prediction errors) is inducing disengagement from current predictive activity in order to move to a more predictable 'set', what one could call compensatory progress and order. Preliminary evidence for this idea has come from studies finding increased predictable pattern perception when confronted with ambiguity, inconsistency or lack of control (Greenaway et al. 2013; Proulx and Heine 2009; van Harreveld et al. 2014; Whitson and Galinsky 2008). More broadly, uncertainty or inconsistencies may lead people to reaffirm their own (predictable) worldviews, such as nationality, ideology or religion (Inzlicht et al. 2011; Proulx et al. 2010). The negative affect thought to drive these effects is, according to our theorizing, a direct reflection of the higher than expected increase in prediction errors. These examples may then all boil down to efforts to return to an expected, positive rate of uncertainty reduction. The reverse may also hold: positive mood seems to induce a greater reliance on default prior or top-down knowledge, as indicated by an increased influence of prior judgments, scripts or stereotypes in event or person perception (Bless 2000; Bodenhausen et al. 1994). This dovetails with the proposed view that positive mood is linked to high predictive progress, implying that the models the organism has about its world have improved and so are adequate. A rational conclusion for the system would then be to increase reliance on these prior, top-down models (and reduce the influence of prediction errors).

A last function of these affective computations relates to learning and attention. Joffily & Coricelli (Joffily and Coricelli 2013) formally show that the first derivative over time of prediction errors can fulfill a similar function as the one usually assigned to the precision mechanism. An increase in prediction errors (negative valence) may indicate that actual, important changes in the world have taken place, so input (incoming prediction errors) should be weighted more heavily compared to top-down predictions (that apparently need to be updated). In other words, the error rates can be used as a meta-learning signal, tuning the learning rate for new inputs, depending on whether there is much too learn (i.e., in a changing world) or not. What we defined as expectations of error reduction rates then take on the role of expectations on learnability of particular input domains. These are models about what we do not know yet about the structure of the world (Joffily and Coricelli 2013) and how these uncertainties will evolve, i.e., to what extent we estimate these uncertainties will be reducible. Joffily & Coricelli argue that a model that uses rate of change in errors is more parsimonious than one including precisions (a conventional PP model), but more work will be needed to clarify both differences in computational realizability and biological plausibility.

## 5 Outlook

Throughout history, visual perception —the 'noblest sense'—was considered our main route to find the "ground truth" about the world out there. But physical differences in the world only become information (meaning) by the way we probe them, with our organism-specific predictions. This means that value and information are intertwined by construction —courtesy of our existence as biological organisms. Taking the organism as an (extensible) model of its environment, epistemic coherence is paramount and emotions emerge as the dynamics of attaining this predictive coherence or error reduction.

In large part, the plausibility of this framework for affect hinges on the success of PP proponents in pinning down the physiological basis of the computational scheme. As discussed, prediction errors are sub-personal processing products, which means that we will need to rely on neural measures for tracking these dynamics. So far, there is no direct neural evidence for the existence of the proposed

computational operations (or their products). There is only very preliminary and indirect evidence to date for the separable error and prediction populations of neurons (de Gardelle et al. 2012), as postulated by PP. However, the general idea that there are different levels in the hierarchy, with separable prediction errors has recently received support (Diuk et al. 2013; Wacongne et al. 2011). Once we succeed in properly localizing those on different hierarchical levels of processing, we can start looking for dynamics in these errors and neural populations or regions that track these changes and generate expectations of error reduction rates. Most likely, these computations are performed distributed in the brain (similar to first-order PP), given the widely distributed encoding of uncertainty in the brain depending on the domain concerned (Bach and Dolan 2012). In this regard, the overlap in regions found to be important in processing uncertainty and those active for emotional processing, is promising.

All in all, this approach shows some promise for the PP framework to become a common, well-specified language for psychology, from low-level sensorimotor issues to emotional and existential issues. The convergence between computational neuroscience and psychology as seen through the PP account is encouraging. However, this is only a preliminary exploration of how emotions and related aspects of experience may be reframed within PP. Many challenges lie ahead, but the question of whether the brain indeed tracks error increases and decreases and forms predictions about those, is open to empirical and computational investigation. If an emotion theory along the lines presented here is right, we might be getting some formal grasp on affective value and intrinsic motivation, key characteristics of proactive, living organisms.

## References


Alter, A. L. & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Pers. Soc. Psychol. Rev., 13* (3), 219–235.

Anselme, P. (2010). The uncertainty processing theory of motivation. *Behav. Brain Res., 208* (2), 291–310.

Anselme, P., Robinson, M. J. F. & Berridge, K. C. (2013). Reward uncertainty enhances incentive salience attribution as sign-tracking. *Behav. Brain Res., 238*, 53–61.

Bach, D. R. & Dolan, R. J. (2012). Knowing how much you don't know: A neural organization of uncertainty estimates. *Nat. Rev. Neurosci., 13* (8), 572–586.

Barrett, L. F. (2014). The conceptual act theory: A précis. *Emot. Rev., 6* (4), 292–297.

Barrett, L. F. & Simmons, W. K. (2015). Interoceptive predictions in the brain. *Nat. Rev. Neurosci., 16* (7), 419–429.

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.

Berlyne, D. E. (1970). Novelty, complexity, and hedonic value. *Percept. Psychophys., 8* (5), 279–286.

Bless, H. (2000). The interplay of affect and cognition: The mediating role of general knowledge structures. *Feeling and thinking: The role of affect in social cognition* (pp. 201–222). New York, NY: Cambridge University Press.

Bodenhausen, G. V., Kramer, G. P. & Süsser, K. (1994). Happiness and stereotypic thinking in social judgment. *J. Pers. Soc. Psychol., 66* (4), 621–632.

Bromberg-Martin, E. S. & Hikosaka, O. (2011). Lateral habenula neurons signal errors in the prediction of reward information. *Nat. Neurosci., 14* (9), 1209–1216.

Cacioppo, J. T. & Petty, R. E. (1982). The need for cognition. *J. Pers. Soc. Psychol., 42* (1), 116–131.

Cantor, N. & Kihlstrom, J. F. (1987). *Personality and social intelligence.* Englewood Cliffs, N J: Prentice-Hall.

Carver, C. S. & Scheier, M. F. (1990). Origins and functions of positive and negative affect: A control-process view. *Psychol. Rev., 97* (1), 19–35.

Chater, N. & Vlaev, I. (2011). The instability of value. In M. Delgado, E. A. Phelps & T. W. Robbins (Eds.) *Decision making: Attention and performance XXIII* (pp. 81–100).

Chetverikov, A. & Filippova, M. (2014). How to tell a wife from a hat: Affective feedback in perceptual categorization. *Acta Psychol., 151*, 206–213.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci., 36* (03), 181–204.

Csikszentmihalyi, M. (1996). Flow and the psychology of discovery and invention. New York: Harper Collins.

Damasio, A. R. (2000). *The feeling of what happens: Body and emotion in the making of consciousness.* Harvest Books.

——— (2003). *Looking for Spinoza: Joy, sorrow, and the feeling brain.* New York: Houghton Mifflin Harcourt.

Dayan, P. (2012). How to set the switches on this thing. *Curr. Opin. Neurobiol., 22* (6), 1068–1074.

de Gardelle, V., Waszczuk, M. & Egner, T. A. (2012). Concurrent repetition enhancement and suppression responses in extrastriate visual cortex. *Cereb. Cortex*.

de Lafuente, V. & Romo, R. (2011). Dopamine neurons code subjective sensory experience and uncertainty of perceptual decisions. *Proc. Natl. Acad. Sci. U. S. A., 108* (49), 19767–19771.

Deacon, T. W. (2011). *Incomplete nature: How mind emerged from matter.* W. W. Norton & Company.

Dember, W. N., Earl, R. W. & Paradise, N. (1957). Response by rats to differential stimulus complexity. *J. Comp. Physiol. Psychol., 50* (5), 514–518.

den Ouden, H. E. M., Friston, K. J., Daw, N. D. A. & Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cereb.Cortex, 19* (1460-2199 (Electronic)), 1175–1185.

Di Paolo, E. A. (2003). Organismically-inspired robotics: Homeostatic adaptation and teleology beyond the closed sensorimotor loop. *Dynamical Systems Approach to Embodiment and Sociality*, 19–42.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M. & Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *J. Neurosci., 33* (13), 5797–5805.

Dreisbach, G. & Fischer, R. (2012). Conflicts as aversive signals. *Brain Cogn., 78* (2), 94–98.

Duncan, S. & Barrett, L. F. (2007). The role of the amygdala in visual awareness. *Trends in Cognitive Science, 11* (5), 190–192.

Duncker, K. (1941). On pleasure, emotion, and striving. *Philosophy and Phenomenological Research, 1* (4), 391–430.

Egbert, M. & Barandiaran, X. E. (2014). Modeling habits as self-sustaining patterns of sensorimotor behavior. *Front. Hum. Neurosci., 8*.

Egbert, M. & Canamero, L. (2014). In H. Sayama, J. Rieffel, S. Risi, R. Doursat & H. Lipson (Eds.) *Habit-based regulation of essential variables*. Cambridge, MA: MIT Press.

Egbert, M., Virgo, N., Egbert, M. D., Froese, T., Kampis, G., Karsai, I. & Szathmáry, E. (2013). For biological systems, maintaining essential variables within viability limits is not passive. *Constructivist Foundations,* 9 (1), 109–111.

Ellsworth, P. C. & Scherer, K. R. (2003). Appraisal processes in emotion. *Handbook of Affective Sciences, 572*.

Festinger, L. (1962). Cognitive dissonance. *Sci. Am., 207* (4), 93–107.

Freddolino, P. L. & Tavazoie, S. (2012). Beyond homeostasis: A predictive-dynamic framework for understanding cellular behavior. *Annu. Rev. Cell Dev. Biol., 28*, 363–384.

Frijda, N. H. (1987). Emotion, cognitive structure, and action tendency. *Cogn. Emot., 1*, 115–143.

——— (2006). *The laws of emotion.* Mahwah, N.J: Psychology Press.

Friston, K. J. (2003). Learning and inference in the brain. *Neural Netw., 16* (9), 1325–1352.

——— (2009). The free-energy principle: A rough guide to the brain? *Trends Cogn. Sci., 13* (7), 293–301.

——— (2010). The free-energy principle: A unified brain theory? *Nat. Rev. Neurosci., 11* (2), 127–138.

Friston, K. J., Daunizeau, J. & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLoS One, 4* (7), e6421.

Friston, K. J., Daunizeau, J., Kilner, J. & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biol. Cybern., 102* (3), 227–260.

Friston, K. J., Thornton, C. & Clark, A. (2012a). Free-energy minimization and the Dark-Room problem. *Front. Psychol., 3*.

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M. A., Brown, H., Dolan, R. J., Moran, R. A. & Bestmann, S. (2012b). Dopamine, affordance and active inference. *PLoS Comput. Biol., 8* (1), e1002327.

Friston, K. J., Schwartenbeck, P., Fitzgerald, T. M., Behrens, T. & Dolan, R. J. (2013). The anatomy of choice: Active inference and agency. *Front. Hum. Neurosci., 7*, 598.

Friston, K. J., Rigoli, F., Ognibene, D. M., Fitzgerald, T. & Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci., 6* (4), 187–214.

Froese, T. & Ikegami, T. (2013). The brain is not an isolated ''black box," nor is its goal to become one. *Behav. Brain Sci., 36* (03), 213–214.

Gerken, L., Balcomb, F. K. & Minton, J. L. (2011). Infants avoid 'labouring in vain' by attending more to learnable than unlearnable linguistic patterns. *Dev. Sci., 14* (5), 972–979.

Gershman, S. J. & Daw, N. D. (2012). Perception, action and utility: The tangled skein. *Principles of Brain Dynamics: Global State Interactions*, 293–312.

Greenaway, K. H., Louis, W. R. & Hornsey, M. J. (2013). Loss of control increases belief in precognition and belief in precognition increases control. *PLoS One, 8* (8).

Gu, X., Hof, P. R., Friston, K. J. & Fan, J. (2013). Anterior insular cortex and emotional awareness. *J. Comp. Neurol., 521* (15), 3371–3388.

Harper, M. (2009). The replicator equation as an inference dynamic.

Hebb, D. O. (1946). On the nature of fear. *Psychol. Rev., 53* (5), 259–276.

Heylighen, F. & Joslyn, C. (2001). Cybernetics and second order cybernetics. *Encyclopedia of Physical Science & Technology, 4*, 155–170.

Hirsh, J. B., Mar, R. A. & Peterson, J. B. (2012). Psychological entropy: A framework for understanding uncertainty-related anxiety. *Psychol. Rev., 119* (2), 304–320.

Hoffmann, J. (2003). Anticipatory behavioral control. In M. V. Butz, O. Sigaud & P. Gérard (Eds.) *Anticipatory behavior in adaptive learning systems* (pp. 44–65). Springer Berlin Heidelberg.

Hohwy, J. (2011). Phenomenal variability and introspective reliability. *Mind Lang., 26* (3), 261–286.

Horn, B. K. P. (1980). *Derivation of invariant scene characteristics from images* (pp. 371–376). New York, NY, USA: ACM.

Hsee, C. K. & Ruan, B. (2016). The pandora effect: The power and peril of curiosity. *Psychol. Sci., 27* (5), 659–666.

Iglesias, S., Mathys, C., Brodersen, K. H. A., Piccirelli, M. & den Ouden, H. E. M. A. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron, 80* (2), 519–530.

Inzlicht, M., Tullett, A. M. & Good, M. (2011). The need to believe: A neuroscience account of religion as a motivated process. *Religion Brain Behav., 1* (3), 192–212.

Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspect. Psychol. Sci., 2* (3), 260–280.

Jackson, F., Nelson, B. D. & Proudfit, G. H. (2014). In an uncertain world, errors are more aversive: Evidence from the error-related negativity. *Emotion*.

James, W. (1890). The principles of psychology. *Harvard UP, Cambridge, MA*.

Joffily, M. & Coricelli, G. (2013). Emotional valence and the Free-energy principle. *PLoS Comput. Biol., 9* (6), e1003094.

Kagan, J. (1972). Motives and development. *J. Pers. Soc. Psychol., 22* (1), 51–66.

Kaplan, H. (1987). Lottery winners: The myth and reality. *J. Gambl. Stud., 3* (3), 168–178.

Kaplan, F. & Oudeyer, P.-Y. (2007). In search of the neural circuits of intrinsic motivation. *Front. Neurosci., 1* (1), 225–236.

Keramati, M. & Gutkin, B. S. (2011). A reinforcement learning theory for homeostatic regulation. *Advances in neural information processing systems* (pp. 82–90).

Kidd, C., Piantadosi, S. T. & Aslin, R. N. (2012). The goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS One, 7* (5), e36399.

Klein, C. (2014). The penumbral theory of masochistic pleasure. *Rev.Phil.Psych., 5* (1), 41–55.

Klein, D. C. & Seligman, M. E. (1976). Reversal of performance deficits and perceptual deficits in learned helplessness and depression. *J. Abnorm. Psychol., 85* (1), 11–26.

Knight, E. J., Klepac, K. M. & Kralik, J. D. (2013). Too good to be true: Rhesus monkeys react negatively to better-than-expected offers. *PLoS One, 8* (10), e75768.

Lauwereyns, J. (2010). *The anatomy of bias: How neural circuits weigh the options.* Cambridge, MA: MIT Press.

Little, D. Y.-J. & Sommer, F. T. (2013). Maximal mutual information, not minimal entropy, for escaping the "dark room". *Behav. Brain Sci., 36* (03), 220–221.

Mandler, G. (2003). Emotion. In D. K. Freedheim & I. B. Weiner (Eds.) *History of psychology*. John Wiley and Sons.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J. & Stephan, K. E. (2014). Uncertainty in perception and the hierarchical Gaussian filter. *Front. Hum. Neurosci., 8*, 825.

Millikan, R. G. (2004). *Varieties of meaning: The 2002 Jean Nicod lectures.* Cambridge, MA: Mit Press.

Mineka, S. & Hendersen, R. W. (1985). Controllability and predictability in acquired motivation. *Annu. Rev. Psychol., 36* (1), 495–529.

Moors, A. (2010). Automatic constructive appraisal as a candidate cause of emotion. *Emot. Rev., 2* (2), 139–156.

Moutoussis, M., Fearon, P., El-Deredy, W., Dolan, R. J. & Friston, K. J. (2014). Bayesian inferences about the self (and others): A review. *Conscious. Cogn., 25*, 67–76.

Moutoussis, M., Story, G. W. & Dolan, R. J. (2015). The computational psychiatry of reward: Broken brains or misguided minds? *Front. Psychol., 6*, 1445.

Niv, Y. & Chan, S. (2011). On the value of information and other rewards. *Nat. Neurosci., 14* (9), 1095–1097.

Noordewier, M. K. & Breugelmans, S. M. (2013). On the valence of surprise. *Cogn. Emot., 27* (7), 1326–1334.

Oudeyer, P.-Y., Kaplan, F. & Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput., 11* (2), 265–286.

Parrott, W. G. & Gleitman, H. (1989). Infants' expectations in play: The joy of peek-a-boo. *Cogn. Emot., 3* (4), 291–311.

Pezzulo, G., Rigoli, F. & Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Prog. Neurobiol., 134*, 17–35.

Picard, F. (2013). State of belief, subjective certainty and bliss as a product of cortical dysfunction. *Cortex, 49* (9), 2494–2500.

Polani, D. (2009). Information: Currency of life? *HFSP J., 3* (5), 307–316.

Proulx, T. & Heine, S. J. (2009). Connections from Kafka exposure to meaning threats improves implicit learning of an artificial grammar. *Psychol. Sci., 20* (9), 1125–1131.

Proulx, T., Heine, S. J. & Vohs, K. D. (2010). When is the unfamiliar the uncanny? Meaning affirmation after exposure to absurdist literature, humor, and art. *Pers. Soc. Psychol. Bull., 36* (6), 817–829.

Purves, D., Wojtach, W. T. & Lotto, R. B. (2011). Understanding vision in wholly empirical terms. *Proceedings of the National Academy of Sciences, 108* (Supplement_3), 15588–15595.

Reber, R., Schwarz, N. & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Pers. Soc. Psychol. Rev., 8* (4), 364–382.

Reisenzein, R. (2009). Emotional experience in the computational belief-desire theory of emotion. *Emot. Rev., 1* (3), 214–222.

Rozin, P. & Kennel, K. (1983). Acquired preferences for piquant foods by chimpanzees. *Appetite, 4* (2), 69–77.

Rozin, P., Rozin, A., Appel, B. & Wachtel, C. (2006). Documenting and explaining the common AAB pattern in music and humor: Establishing and breaking expectations. *Emotion, 6* (3), 349–355.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychol. Rev., 110* (1), 145–172.

Schiffer, A.-M., Ahlheim, C. & Wurm, M. F. S. (2012). Surprised at all the entropy: Hippocampal, caudate and midbrain contributions to learning from prediction errors. *PLoS One, 7* (5), e36445.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Trans. Auton. Ment. Dev., 2* (3), 230–247.

Schouppe, N., Braem, S., Houwer, J. D. A., Verguts, T. & Ridderinkhof, K. R. N. (2014). No pain, no gain: The affective valence of congruency conditions changes following a successful response. *Cogn. Affect. Behav. Neurosci.*, 1–11.

Schrödinger, E. (1992). *What is life? With Mind and Matter and Autobiographical Sketches.* Cambridge University Press.

Schultz, W. (2007). Reward. *Scholarpedia J., 2* (3), 1652.

Schwartenbeck, P., FitzGerald, T. H. B., Mathys, C., Dolan, R. & Friston, K. (2014). The dopaminergic midbrain encodes the expected certainty about desired outcomes. *Cereb. Cortex*, bhu159.

Serences, J. T. (2008). Value-based modulations in human visual cortex. *Neuron, 60* (6), 1169–1181.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci., 17* (11), 565–573.

Silvia, P. J. (2001). Interest and interests: The psychology of constructive capriciousness. *Rev. Gen. Psychol., 5* (3), 270–290.

——— (2008). Interest—The curious emotion. *Curr. Dir. Psychol. Sci., 17* (1), 57–60.

Singer, T., Critchley, H. D. & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends Cogn. Sci., 13* (8), 334–340.

Singh, S., Lewis, R. L. & Barto, A. G. (2009). *Where do rewards come from* (pp. 2601–2606).

Solway, A. & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychol. Rev., 119* (1), 120–154.

Srivastava, N. & Schrater, P. (2015). Learning what to want: context-sensitive preference learning. *PLoS One, 10* (10), e0141129.

Sroufe, L. A. & Waters, E. (1976). The ontogenesis of smiling and laughter: A perspective on the organization of development in infancy. *Psychol. Rev., 83* (3), 173–189.

Stansbury, D. E., Naselaris, T. & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron, 79* (5), 1025–1034.

Van de Cruys, S. & Wagemans, J. (2011). Putting reward in art: A tentative prediction error account of visual art. *Iperception, 2* (9), 1035–1062.

van Harreveld, F., Rutjens, B. T., Schneider, I., Nohlen, H. U. & Keskinis, K. (2014). In doubt and disorderly: Ambivalence promotes compensatory perceptions of order. *J. Exp. Psychol. Gen., 143* (4), 1666–1676.

von Glasersfeld, E. (1995). *Radical constructivism: A way of knowing and learning. studies in mathematics education series: 6.* Bristol, PA: Falmer Press, Taylor & Francis Inc..

Wacongne, C., Labyt, E., Wassenhove, V., Bekinschtein, T., Naccache, L. & Dehaene, S. (2011). Evidence for a hier-

archy of predictions and prediction errors in human cortex. *Proc. Natl. Acad. Sci. U. S. A., 108* (51), 20754–20759.

Weber, A. & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenol. Cognitive Sci., 1* (2), 97–125.

Whalen, P. J. (2007). The uncertainty of it all. *Trends Cogn. Sci., 11* (12), 499–500.

Whitson, J. A. & Galinsky, A. D. (2008). Lacking control increases illusory pattern perception. *Science, 322* (5898), 115–117.

Wood, W. & Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychol. Rev., 114* (4), 843–863.

Wänke, M. & Hansen, J. (2015). Relative processing fluency. *Curr. Dir. Psychol. Sci., 24* (3), 195–199.

Wörgötter, F. & Porr, B. (2005). Temporal sequence learning, prediction, and control: A review of different models and their relation to biological mechanisms. *Neural Comput., 17* (2), 245–319.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *Am. Psychol., 35* (2), 151.

# Action Prevents Error

## Predictive Processing without Active Inference

Jona Vance

According to predictive processing, minds relentlessly aim at a single goal: prediction error minimization. Prediction error minimization is said to explain everything the mind does, from perception to cognition to action. Here I focus on action. 'Active inference' is the standard approach to action in predictive processing. According to active inference, as it has been developed by Friston and collaborators, action ensues when proprioceptive predictions generate prediction error at the motor periphery, and classical reflex arcs engage to quash the error. In this paper, I raise a series of problems for active inference. I then offer an alternative approach on which action prevents error, rather than quash it. I argue that the action prevents error approach solves all the problems raised for active inference. In addition, I show how the alternative approach can be independently motivated by further commitments of predictive processing and that it is compatible with other prominent approaches to sensorimotor psychology, such as optimal feedback control.

**Keywords**

Action | Active inference | Deafferentation | Optimal control | Proprioception

**Acknowledgments**

Thanks to two anonymous referees for feedback on an earlier draft. Thanks also to Michael Anderson, Maria Brincker, Christopher Burr, Karl Friston, Jakob Hohwy, Bryce Huebner, Max Jones, and Alex Kiefer for discussions of this material. Special thanks to Thomas Metzinger and Wanja Wiese for very helpful comments and for their patience.

### 1 Introduction

In predictive processing (PP) frameworks, prediction error minimization explains everything the mind does, from perception to cognition to action. The mind constructs models containing information about objects and properties of all kinds. To test its models, the mind predicts the sensory inputs it's likely to receive if the models are accurate. Then it compares the predicted inputs with the inputs it actually receives. If there's a match between the predicted and actual inputs, the models are confirmed. If there's a mismatch, prediction error occurs.

To minimize prediction error the mind has two options. It can either revise its models to better conform to the actual signal, or it can act on the world to change its sensory inputs, bringing them into conformity with predictions. Call the former, model-altering form of prediction error minimization revision-PEM. Revision-PEM provides the central recipe for perception. Revision-PEM plays a key role in allowing organisms to modify their models to reliably track changing features of the world. Call the latter input-altering form of prediction error minimization action-PEM. It provides the central recipe for action (conceived liberally to include saccades and other involuntary behavior). This paper addresses the standard approach to action in PP and offers an alternative.[1]

The paper proceeds as follows. Section 2 introduces the standard approach to action in PP and contrasts it with an alternative approach that is prominent in motor control theory outside PP. Sections 3-5 raise problems for active inference. Section 6 describes an alternative to active inference that is compatible with the core commitments of PP and argues for its superiority.

1 My focus here is only on bodily action. For discussion of mental action in PP, see Metzinger 2017.

Vance, J. (2017). Action Prevents Error - Predictive Processing without Active Inference.
In T. Metzinger & W. Wiese (Eds.). *Philosophy and Predictive Processing:* 25. Frankfurt am Main: MIND Group. doi: 10.15502/9783958573260

## 2 Active Inference and an Alternative

The standard approach to action in predictive processing is active inference (Adams et al. 2013; Brown et al. 2013; Clark 2013; Clark 2016; Friston 2009; Friston 2011; Friston et al. 2010; Hohwy 2013; Shipp et al. 2013). In active inference, motor control is carried out in a probabilistic (Bayesian) hierarchy of numerous layers. Each layer, except the lowest, sends predictions to the layer below. These predictions are carried by top-down connections between the layers. The connections encode the system's generative model, specifying probabilistic relationships between the features represented at adjacent layers. In addition, each layer sends bottom-up error signals to the layer above. Error signals are generated by comparing predicted activity at a layer with actual activity at that layer and encoding the discrepancy. Finally, each layer also makes top-down precision estimates of the reliability of the error signal it receives from below. In these respects, active inference entails that the motor hierarchy has deep structural similarity with the perceptual hierarchy responsible for exteroceptive processes such as visual and auditory processing.[2]

In active inference, motor control and proprioception are deeply functionally and anatomically unified. Suppose an agent deliberately reaches for a cup. In active inference, the process is characterized as follows. The motor system encodes a hierarchically distributed set of prior probabilities over possible trajectories. The intention to reach for the cup is realized by activation at high levels of the motor hierarchy. These high level activations initiate further activations at lower layers of the motor hierarchy. The downward process carries predictions guided by the brain's somatomotor generative model of the body and world. The downward predictions unpack the subject's high level intention into expected proprioceptive sensory inputs; here, the inputs that would be received if the body moved so as to carry out the intended action of reaching for the cup. When the intention is initially formed, the body is not currently moving along the desired trajectory: the hand is not yet moving toward the cup. As a result, proprioceptive and other sensory inputs associated with the expected trajectory are not received. The mismatch between the expected sensory inputs and the actual inputs results in prediction error. According to active inference, this prediction error is encoded by alpha motoneuron activity (Adams et al. 2013), which compares downward predictions from primary motor cortex with upward Ia afferent inputs. Having encoded the discrepancy between these top-down and bottom-up signals, alpha motoneurons send the resulting error signal outward to enervate relevant muscle fibers via classical reflex arcs, which quash the prediction error and entrain movement along the expected trajectory. These are the basic details of active inference.

The term 'active inference' arguably has multiple meanings in the literature. Friston et al. 2010 (p. 232) characterize active inference as "the interplay of perception and action". Similarly, Brown et al. 2013 characterize active inference as the combination of perceptual inference and action. They write, "The brain can minimize prediction error in one of two ways. It can either change its predictions to better cohere with sensory input, or change the sampling of the environment such that sensory samples conform to predictions. The former process corresponds to perceptual inference...the latter to action: together they constitute 'active inference'" (p. 614). By contrast, Hohwy 2013 characterizes active inference more like Brown et al.'s notion of action (pp. 89-90). On his characterization, active inference contrasts with perceptual inference, rather than including it (Cf. p. 83). Similarly, Anil Seth writes that active inference is "Classically conceived of as the minimization of prediction error by performing actions that confirm sensory predictions" (Seth 2015, p. 2). Seth then argues that active inference should also be characterized as including the performance of actions to disconfirm predictions and the performance of actions to disambiguate among competing hypotheses. Finally, for Rescorla 2016, 'active inference' denotes the specific theory of motor control developed by Friston and colleagues, as outlined in this section and in the citations therein. It is in this final way that I use the term 'active inference' in this paper: to name the specific approach to motor control advanced by Fris-

2 For further discussion of top-down influences in perception in PP, see Vance and Stokes 2017.

ton and colleagues; including its integration with perceptual processing; its probabilistic, hierarchical modeling; and its commitment to the role of outward error signals from alpha motoneurons and reflex arcs at the motor periphery.

Although active inference is the standard approach to action in predictive processing, it is not the standard approach to action in sensorimotor psychology more generally.[3] One dominant approach more generally is optimal feedback control (Todorov 2004; Todorov 2009; Todorov and Jordan 2002). Like active inference, optimal feedback control adopts a hierarchical Bayesian approach to motor control. Active inference and optimal feedback control are also both probabilistic, hierarchical approaches to action. And both can be used to model motor control as a probabilistic inference problem (Friston et al. 2010; Todorov 2009). However, the two approaches also have significant differences (Adams et al. 2013; Rescorla 2016). Several differences are worth highlighting for our purposes.

First, the two approaches differ in how they characterize the functional architecture of the motor control and proprioception. In optimal control theory, motor pathways and proprioceptive pathways interact, but they are distinct. By contrast, in active inference motor pathways double as proprioceptive pathways. According to active inference, motor control and proprioception are implemented using one and the same generative model. The generative model is encoded in the weights between nodes of a belief net. Nodes are realized by populations of neurons. Since, in active inference, the network connections that carry efferent motor signals are one and the same as the network connections that carry proprioceptive predictions, motor command pathways and proprioceptive pathways are one and the same. Thus, in active inference — but not optimal control — the neural channels that implement motor control are, to a large extent, the very same channels that implement proprioception.

Second, the two approaches differ in how they characterize downward motor signals. In active inference, downward motor signals are sensory predictions that are unpacked using the somatomotor generative model mentioned above. Thus, in active inference there are no motor commands per se, only proprioceptive predictions that serve as (or implement) a kind of implicit motor command. By contrast, in optimal control theory, downward motor signals are modeled as commands carried by an inverse model (or controller). The controller assigns the motor command required to achieve the desired outcome.

Third, motor efference copy plays an important role in optimal control theory, whereas active inference does without efference copy. In optimal control theory, each time a motor command is given, the system sends a copy of the command back to the Bayesian estimator, which allows the system to estimate the likely effect of the command on the environmental state using a forward model. Engaging the estimator through efference copy allows the system to compare expected sensory inputs (associated with the intended action) with actual inputs, make online corrections as needed, and update the model through learning. In active inference, there are no motor commands, so there's nothing to copy. Instead, there are only sensory predictions carried downward by the generative model. Proprioceptive predictions terminating at alpha motor neurons implement (or play the role of) 'implicit' motor commands (Clark 2016, pp. 127-128). Sensory predictions terminating at other parts of the periphery constitute "corollary discharge", without efference copy (Clark 2016, pp. 125-127).

Fourth, the two approaches differ in how they characterize the outward signal from alpha motoneurons to muscle fibers. In active inference, alpha motoneurons generate the outward signal to muscle fibers by comparing Ia afferent feedback and downward predictions from motor cortex, encoding the mismatch, and sending the resulting error as an outward message (Adams et al. 2013). By contrast, in optimal control theory, downward signals from motor cortex engage alpha motoneurons, which further engage muscle fibers directly, without first comparing the downward signal with Ia afferent feedback. That is, active inference differs from most other approaches (including optimal control) by characterizing outward signals from alpha motoneurons to muscle fibers as error signals.

3 I take the term 'sensorimotor psychology' from Rescorla 2016. Sensorimotor psychology is often called motor control by its practitioners.

In the following sections, I raise a series of problems for active inference. In subsequent sections, I show how PP can be combined with optimal control theory to solve the problems.

## 3 The Action/Revision Problem

In active inference, action-PEM and revision-PEM both occur via proprioceptive predictive pathways. What explains why proprioceptive predictions at a given location in the hierarchy sometimes entrain action-PEM and at other times drive revision-PEM? Call this the action/revision problem for active inference. There are various statements of the action/revision problem in the literature. For example, Hohwy puts the problem as follows:

> [T]his overall account [active inference] creates a puzzle about how action is triggered, that is how the agent shifts from perceptual to active inference. This is because there will be competition between assessment of the actual proprioceptive input and the counterfactual proprioceptive input. Rather than changing the world to fit with the counterfactual predicted input, the system could just adjust its proprioceptive prediction in light of the actual input — it could realize that it is not actually in that state. This would prevent action from arising. A mechanism is thus needed to ensure agency (Hohwy 2013, p. 83).

Hohwy's formulation implies that the problem is to explain how and why the system shifts from, in my terminology, revision-PEM to action-PEM. The implication is that revision-PEM is the primary or default form of prediction error minimization, and that the system must be moved out of that default mode to engage in action. However, one need not assume that revision-PEM is the default form of PEM. And, indeed, some prominent recent developments of the PP framework eschew that assumption (Clark 2016; Seth 2015). In setting up the action/revision problem for PP, I do not assume that one form of PEM is primary or default.[4]

It's useful to contrast the action/revision problem with the so-called "dark room problem" for PP. The dark room problem asks: if the system aims at minimizing prediction error why doesn't the subject just go into a dark room where there is little sensory information thereby reducing the chance of error? If there is little sensory information in the first place, then, one might think, avoiding prediction error will be easier. A promising solution to the dark room problem does not solve the action/revision problem. The proposed solution entails that some predictions about endogenous states are unrevisable, such as predictions about blood sugar levels (Shea 2013, p. 229). When blood sugar levels drop below predictions, the systems cannot revise predictions; increasing blood sugar is the only way to minimize prediction error. The proposal works well when applied to subsystems that monitor blood sugar levels and inflexibly predict target levels. But that is not how most action works. Most goals are highly flexible and require constant revision. Action-oriented processing will have to involve plenty of revision along the way. Any solution to the action/revision puzzle must allow for revision during action.

Proponents of active inference hold that, in my terminology, both action-PEM and revision-PEM occur simultaneously (Clark 2016, p. 124). This is possible at different places in the network. However, note that engaging in action and revision simultaneously at the same location in the network would not minimize prediction error. For example, if one simultaneously engaged peripheral proprioceptive layers in action and revision, action-PEM would throw the body into motion in the predicted way while the revised predictions anticipate that the body was not moving in the predicted way. This would create new prediction error because there would be a new mismatch of predicted and actual state. The

4 Another formulation is due to Brown et al. (Brown et al. 2013, p.411) who write: "[W]e can either change our predictions to explain sensory input through perception. Alternatively, we can actively change sensory input to fulfill our predictions…. However, this creates a conflict between action and perception; in that, self-generated movements require predictions to override the sensory evidence that one is not actually moving." This formulation does not assume any priority of revision- over action-PEM. Cf. Clark (Clark 2016, p. 215).

problem for PP is to explain when, how, and why the system engages in action-PEM or revision-PEM at a given time and location in the network.

I now consider and criticize two proposed solutions to the action/revision problem. The first proposal appeals to counterfactual or subjunctive content to distinguish predictions that entrain action-PEM from those that do not. Here is Hohwy:

> Now consider how action comes about...The representations of predicted sensory input are counterfactual in the sense that the say how the sensory input *would* change if the system *were* to act in a certain way (Hohwy 2013; p. 82, italics original).[5]

The suggestion is that action-entraining predictions are distinct from others by involving counterfactual content. For example, to raise one's arm, one predicts something in the vicinity of: if I were to raise my arm, I would receive inputs Y. Now, generative models do encode counterfactuals by encoding information about the relationship between sensory inputs and features of the body or environment. And these counterfactuals are utilized in generating action according to active inference. But action-entraining proprioceptive predictions cannot themselves be conditionals. They are not in the subjunctive form: if X were the case, Y would be the case. Whenever a counterfactual conditional is false, it's because, at the nearest world where the antecedent is true, the consequent is false. So, for example, the conditional "if I were to raise my arm, I would receive inputs Y" is false if and only if at the nearest world where my arm is raised, I don't receive inputs Y. But the falsity of this conditional does not generate the prediction error needed for movement. Such a conditional is encoded in the generative model; if it's falsified, the model must be revised to encode which inputs are really connected with a raised arm. The prediction that can incite action is one such that raising one's arm fulfills the prediction. Raising one's arm would not make true the counterfactual conditional "if I were to raise my arm, I'd receive inputs Y".

According to Clark 2016, the predictions that entrain action refer to states of affairs that are non-actual. This strikes me as the right way to put things. In active inference, when the somatomotor system predicts some non-actual state of affairs, action can ensue to make that state of affairs actual, thereby quashing prediction error. For example, if I want to raise my arm, my somatomotor system can make a prediction that my arm is raised, even though it is not. This prediction generates further predictions down the somatomotor hierarchy yielding more specific proprioceptive predictions of non-actual sensory inputs. When reflex arcs are engaged to move the arm in the appropriate way, the predictions are fulfilled and the prediction errors are quashed. But predicting non-actual states of affairs is not unique to action, nor can it explain why action ensues. Predicting non-actual states of affairs is typically what generates prediction error in every modality, from vision to audition and so on. When the predicted state of affairs and the actual state of affairs differ, prediction error occurs. The upshot is that there is nothing about whether the content is counterfactual or non-actual that can provide a satisfactory response to the action/revision puzzle.

I now turn to the most prominent solution to the action/revision problem. It appeals to precision weighting along proprioceptive pathways. Recall that in active inference proprioceptive predictions in the very same neural pathways sometimes entrain action-PEM and at other times drive revision-PEM. On the precision balance view, the shift between these two forms of PEM results from shifts in the relative weighting of proprioceptive precision expectations. On Brown et al.'s characterization, the balance is between precision expectations for proprioceptive error signals and the precision of pro-

5 Clark initially appears to characterize things as Hohwy does. He writes, "PP...already subverts the traditional picture with respect to perception...The same story applies...to the motor case. The difference is that motor control is, in a certain sense, *subjunctive*. It involves predicting the non-actual proprioceptive trajectories that would ensue were we performing some desired action" (Clark 2016, p. 121, italics original). However, Clark arguably does not make Hohwy's mistake. Clark appeals to predictions of the non-actual proprioceptive trajectories that would ensue, but he does not imply that the prediction is itself a counterfactual or subjunctive conditional. He says the predictions are subjunctive only 'in a certain sense'.

proprioceptive predictions.[6] On the precision balance view, action ensues when the precision of proprioceptive predictions is higher than the precision estimate for the proprioceptive error signal.[7]

The precision balance approach to active inference initially appears to solve the action/revision problem. However, it faces an extension of the problem in accounting for proprioceptive revision during action. The precision balance view concedes that when the precision balance is set to favor action-PEM at the proprioceptive periphery, proprioceptive revision-PEM at the periphery cannot occur. This is because, in active inference, motor predictions and proprioceptive predictions occur in the very same neural pathways, and only one form of PEM can obtain at a given time and layer. Now, during an action, the motor system sends multiple messages to the periphery. Sending multiple motor messages controls the degree and duration of muscle activation (Knierim 1997, Ch. 3). This means that, on the precision balance view, at various times during a single movement, the precision balance will have to be set to favor action-PEM and proprioceptive predictions of expected trajectories will have to be sent down the motor hierarchy. However, in addition, there is significant evidence that proprioceptive processing engages in revision-PEM regularly during action in order to deliver feedback to the motor system so that it can make online adjustments at various points during movement. Such evidence comes from neurological studies of the activation of feedback pathways during action (Azim et al. 2014) as well as behavioral studies of prescribed adjustment during action (Liu and Todorov 2007). As a result, the precision balance approach to active inference must explain how action-PEM and revision-PEM can both occur during a single action.

On the precision balance view, the most natural suggestion requires that the precision balance shifts multiple times during a single action, so that action-PEM and revision-PEM can both occur. But the view requires more than repeated shifts to the precision balance in proprioceptive pathways during action. Recall that the action/revision problem is pressing in part because on active inference approaches, the very same generative model (and neural pathways) are responsible both for action-PEM through motor control and proprioceptive revision-PEM. Proprioceptive predictions that entrain action-PEM will have significantly different content from proprioceptive predictions that allow for useful feedback about the actual state of the body. As a result, the shift proposal requires that the system make desired-trajectory predictions to initiate action; then, during action, the same system must adopt a different set of predictions aimed at correctly predicting the actual-current-state of the body, and so on, each time a new motor prediction is issued. Since the alternation is supposed to be realized in the same proprioceptive pathways, activity at the relevant nodes will have to shift rapidly between predicting the desired and actual sensory inputs, without being able to predict both simultaneously. As such, it's not clear how the system could keep track of the desired trajectory while engaged in revision-PEM aimed at anticipating the body's actual current state, since the very same nodes in the proprioceptive hierarchy that predict the desired trajectory will have to be recruited to predict the body's actual state (and vice versa).

6 Support for this characterization of the view comes from a number of passages. For example, they write "As the prior precision increases in relation to the sensory precision, prior beliefs are gradually able to incite more confident movement" (p. 421). Additionally, in describing akinesia (failure to move) they write, "Here, the sensory attenuation leaves the sensory precision higher than the precision of the prior beliefs about internal hidden causes" (p. 420). Again describing akinesia, Brown et al. add, "In this case, bottom-up prediction errors retain a higher precision than descending predictions during movement" (p. 420). The balance is clearly between prediction precisions and precision estimates of the error signal.

7 Clark and Hohwy offer a different characterization of the precision balance view. On their characterization, the relevant balance is not between precision expectations for proprioceptive error signals and the precision of proprioceptive predictions but, rather, between precisions accorded to different aspects of the proprioceptive bottom-up error signal. For example, Clark writes, "Such a system...is able to generate a bodily movement when (but only when) the balance between reliance upon current sensory input and reliance upon higher level proprioceptive predictions is correct. At the limit, errors associated with the higher level proprioceptive predictions (specifying the desired trajectory) would be accorded a very high weighting, while those associated with current proprioceptive input (specifying the current position of the limb or effector) would be low-weighted" (Clark 2016, p. 216). And Hohwy writes, "[A]ction ensues if the counterfactual proprioceptive input is expected to be more precise than actual proprioceptive input, that is, if the precision weighted gain is turned down on the actual input. This attenuates the current state and throws the system into active inference" (Hohwy 2013, p. 83). In the main text, I focus only on Brown et al.'s characterization, since that it the official version of the view, and the one which Clark and Hohwy aim to summarize.

## 4 Deafferentation and Proprioceptive Experience

In this section, I raise a second problem for active inference. The problem appeals to patients who have large fiber neuropathy in their limbs and entirely lack proprioceptive experience of the affected limbs. Despite their lack of proprioceptive experience, these deafferented patients can move their affected limbs (Forget and Lamarre 1987; Ghez and Sainburg 1995; Messier et al. 2003; Rothwell et al. 1982). Deafferented patients tend to exhibit significant motor deficits when compared to normal subjects on prescribed tasks. Yet, motor performance significantly improves when these patients can utilize exteroceptive feedback (e.g. from vision) to help guide action (Sainburg et al. 1993; Sainburg et al. 1995). The challenge for proponents of active inference is to explain how the proprioceptive channels allegedly responsible for action-PEM and revision-PEM remain functional while such patients entirely lack proprioceptive experience of the affected limbs.

Here we must distinguish lack of proprioceptive experience from lack of proprioceptive reliability. Proprioceptive experience concerns phenomenology. Reliability concerns how accurately the relevant processing responds to inputs at the proprioceptive periphery over a range of circumstances. The reliability of proprioceptive processing in deafferented subjects is greatly reduced compared to controls. But reduction of reliability implies nothing about whether such processing generates phenomenology or not. In principle, a perfectly reliable process could be entirely lacking in phenomenology, and a phenomenally rich experience could be entirely unreliable, as would be the case in some forms of hallucination. Deafferented patients under discussion lack both proprioceptive phenomenology and reliability with respect to affected areas. My focus in this section concerns their lack of proprioceptive phenomenology.

The central neuropathic difference between the deafferented patients in question and normal controls is that deafferented patients lack Ia afferent feedback at the proprioceptive periphery, while Ia feedback remains intact for controls. Active inference could use this difference to explain the radical difference in proprioceptive reliability between deafferented patients and controls. However, the presence or absence of Ia afferent feedback cannot on its own explain deafferented patients' total lack of proprioceptive phenomenology in affected areas on the active inference approach. In PP, phenomenology partly supervenes on properties of top-down and lateral neural activation—i.e. features of predictions and precision estimates. Input signals at the sensory and proprioceptive periphery, help drive and shape the system's processing. But these bottom-up signals are not part of the supervenience base of perceptual or proprioceptive phenomenology. So the presence or absence of Ia afferent feedback cannot on its own explain the lack of proprioceptive phenomenology in deafferented patients.

Besides the lack of Ia feedback, there are some other differences between deafferented and normal subjects. But none of these differences can explain the the total lack of proprioceptive phenomenology with respect to deafferented limbs, given that deafferented patients can nevertheless move the affected limbs and can improve their movement with exteroceptive feedback. For example, proprioceptive predictions in deafferented subjects may be quite different in content from control subjects. A difference in the content of proprioceptive predictions during action can explain some difference in the phenomenal character of proprioceptive experience during action in deafferented subjects. But it cannot explain the total lack of proprioceptive experience for such subjects. After all, in order to explain how voluntary movement is possible in the affected limbs of deafferented patients, proponents of active inference must accept that proprioceptive predictions occur with respect to the affected limbs for deafferented patients.

Proponents of active inference could respond to my argument so far as follows. During movement in normal subjects, it is well known that sensory attenuation occurs: that is, normal subjects are less reliable and feel proprioceptive experience less robustly in their active limbs. On the precision balance approach, this is because action ensues when the precision balance favors proprioceptive predictions rather than proprioceptive prediction error. The precision balance is used to explain proprioceptive

attenuation in normal subjects during action: by attenuating the precision expectation for error signals from the relevant limbs, proprioceptive sensitivity and the robustness of phenomenology in those limbs is reduced. If deafferented patients lack relevant proprioceptive prediction error, then the balance might be such as to fully attenuate proprioceptive experience in deafferented patients; that is, lack of relevant prediction error during action in deafferented limbs might explain why deafferented patients wholly lack proprioceptive phenomenology.

I reply as follows: contrary to the above response, deafferented subjects do not entirely lack the relevant proprioceptive error signal in all cases. As Adams et al. 2013 note, parts of the somatomotor hierarchy dealing with affected limbs of deafferented patients are not entirely without relevant feedback in all cases. When such patients can gain visual information about their body and the environment, their accuracy on prescribed tasks significantly improves (Sainburg et al. 1993; Sainburg et al. 1995). In active inference, this means that the somatomotor system in deafferented patients can utilize visual and other exteroceptive information to help generate relevant prediction errors in the somatomotor hierarchy, which can help improve the accuracy of somatomotor predictions during movement. When visual inputs play a role at low levels of the proprioceptive hierarchy, active inference entails that deafferented patients can approximate proprioceptive processes of normal subjects. Visual feedback, say, can provide an approximate substitute for Ia afferent feedback. In such cases, if active inference were true (including its claim that implicit motor commands are realized in proprioceptive pathways), we would expect deafferented patients to have some proprioceptive phenomenology with respect to the affected areas. For, there is considerable proprioceptive phenomenology in normal subjects under such conditions: both during movement where the phenomenology is present but less robust and while the limbs in question are at rest. Deafferented patients' proprioceptive systems engage in the relevant processes of top-down predictions, precision estimates (with respect to error signals in the somatomotor hierarchy that are generated not by Ia feedback but indirectly through visual input), and error correction (again with indirect visual origin). And proponents of active inference accept that all these processes occur to explain why deafferented subjects can engage in prescribed movement and improve accuracy of movement in part through revision-PEM during movement. Yet once this concession is made, there remains no relevant difference to explain why some of these deafferented subjects entirely lack proprioceptive experience during visually guided movement or at rest while visually attending to affected limbs. The only difference in these cases between deafferented patients and controls is that, for deafferented patients, the relevant proprioceptive prediction error comes entirely via visual input, whereas in controls the relevant error comes both from Ia feedback and from vision. This difference can account for some qualitative difference in the proprioceptive phenomenology between deafferented and normal subjects, but it cannot explain the total lack of proprioceptive phenomenology in deafferented subjects in affected areas.

## 5 Deafferenation and Movement

In this section, I raise a final problem for active inference. Like the previous section, my objection in this section appeals to deafferented patients with large fiber neuropathy. Unlike the previous section, here I am concerned with accuracy of movement rather than proprioceptive phenomenology. In some deafferented patients, there is a complete lack of Ia afferent feedback from muscle spindles. Yet these patients are still able to move their affected limbs, sometimes with surprising accuracy. My objection in this section is that active inference cannot account for the accuracy of some deafferented patients' movements.

To be clear, my objection is not that active inference lacks the resources to explain how deafferented patients can move their affected limbs at all. Friston and colleagues are aware of that objection. They put it as follows:

> One important question for active inference is: if movement depends on spinal reflex arcs, then why can neuropathic patients—who lack Ia afferent feedback from muscle spindles—still move? Surely, in the absence of anything to predict there can be no prediction error and no movement (Adams et al. 2013, p. 636).
>
> And they respond as follows:
>
> In fact, the absence of primary afferents does not mean there is no prediction error—top-down predictions can still elicit alpha motor neuron activity. Under active inference, a forward model in the brain converts visuospatial predictions in extrinsic coordinates (low dimensional extrapersonal space) to proprioceptive predictions in intrinsic coordinates (high dimensional proprioceptive space). These predictions then leave the brain and are converted to motor commands by a simple inverse mapping in the spinal cord (see "Discussion"). This spinal inverse mapping is effectively driven by proprioceptive prediction errors and corresponds to the classical reflex arc.
>
> A loss of proprioceptive feedback, therefore, will severely impact upon the spinal inverse mapping, while the cortical forward model can compensate using visual feedback (Bernier et al. 2006). Descending proprioceptive predictions should still be able to activate motor neurons, but they can no longer be compared with precise proprioceptive information and cannot be modified by proprioceptive feedback. (p. 636)

According to Friston and colleagues, even without Ia feedback, the motor systems of deafferented patients can generate the prediction error encoded by alpha motor neurons used to engage reflex arcs. Recall that the relevant error arises when alpha motoneurons compare downward predictions from motor cortex and upward Ia signals. They encode the difference as error and send that error signal outward to enervate the muscles. This error message encodes information for how the muscle should contract. When Ia feedback is lacking, there is still a mismatch between the efferent prediction received by the alpha motoneuron and the (nonexistent) afferent signal. In such a case, the mismatch just is the prediction: when a prediction is compared to no bottom-up signal, the entire prediction is mismatched. Hence, in such a case, alpha motoneurons send the entire prediction as an error signal outward to enervate the muscle. Movement ensues. The worry that active inference entails deafferented subjects cannot move at all is misguided. It is not the worry I raise here. My worry grants that active inference can account for the fact that deafferented patients can move at all.

My objection is that active inference cannot account for the accuracy of these movements in some cases and inaccuracy in others. In active inference, the precise control of movement depends critically on the details of the error signal encoded by alpha motoneurons. A large error signal engages the muscles differently than a small one does. The generative model cannot send just any prediction to the alpha motoneurons and expect to get the desired action trajectory. The predictions must be calibrated such that, when combined with Ia afferent signals, the resulting error signal encodes the right instructions sent to the muscles.

However, in deafferented patients, the error signals encoded by relevant alpha motoneurons will be radically different from the error signals encoded by normal subjects given the same proprioceptive predictions. Given some downward predictions in the motor hierarchy with content P, the error signal for normal subjects will encode the mismatch between the prediction of P and Ia afferent feedback. But given a prediction of P in deafferented patients, the error signal generated by affected alpha motoneurons will be very different, since there's no afferent feedback to compare the prediction with. Since, according to active inference, the specifics of the muscle activation depend critically on the information encoded in the outward error signal generated by alpha motoneurons, active inference predicts that, given the same proprioceptive predictions delivered to the alpha motoneurons of deafferented

patients and controls, the movements by deafferented patients will be radically different from controls — at least when no supplementary feedback is available, e.g. from exteroception.

Many movements by deafferented subjects are indeed radically different from (and less accurate than) movements by normal subjects with the same intentions (Gordon et al. 1995). However, a number of studies show that deafferented patients without Ia feedback can perform some movements without motor control deficit compared to normal subjects. Some such movements are single-joint. For example, Rothwell et al. 1982 found of a deafferented man that "Although he was grossly disabled, it was remarkable to find that he could execute a large repertoire of learned manual motor tasks with both speed and accuracy, despite lacking any useful feedback from his hands" (Rothwell et al. 1982, p. 516; Cf. Forget and Lamarre 1987). Importantly, some patients in these studies could engage in movements with accuracy comparable to controls even when performing prescribed tasks without feedback from exteroceptive modalities such as vision.

In another set of studies, Messier et al. 2003 found that a deafferented subject could achieve accuracy equal to controls when engaging in reaching movements that required multi-joint coordination, even when performing the task with eyes closed, so that there was no perceptual feedback to help guide the movement. Interestingly, Messier et al. found that the speed of action was a significant factor. They write, "Surprisingly, however, he [the deafferented patient] made much larger errors than control subjects at slow and natural speeds, but not at fast speed" (Messier et al. 2003, p. 399). That is, when deafferented patients performed actions at natural and slower-than-natural speeds, they exhibited significant motor deficits compared to normal subjects performing the same tasks at the same speeds. By contrast, deafferented subjects performed as well as normal subjects when they performed the tasks rapidly.

These results are difficult to explain on the active inference approach. Assume for the moment that the alpha motoneuron error signals in areas relevant to movement are very different for deafferented and normal subjects. If that's the case, it's not clear how active inference can explain any of the cases where deafferented and normal subjects perform equally well, since radical differences in the error signals enervating the muscles should result in radical differences in muscle activation. Moreover, it's not clear why speed of movement should matter so much on the active inference approach. Concerning Messier et al.'s results, active inference can appeal to a significant difference in alpha motoneuron error signals in affected areas to explain why deafferented patients perform poorly when performing tasks slowly. But if the error signals are significantly different for deafferented patients (when compared with normal subjects) during slow movements, they will also be significantly different (compared with controls) for fast movements. Thus, active inference predicts that these deafferented patients will exhibit motor deficits compared to controls on both fast and slow movements. Since deafferented and normal subjects perform roughly equally well on fast movements, active inference fails to account for the finding.

Proponents of active inference could offer a number of replies. I now consider several. First, active inference proponents could argue that deafferented patients' motor systems learn a new mapping from alpha motoneuron activation to muscle activation. This proposal is a non-starter because the mapping from alpha motoneuron activity to muscle activation cannot be revised in the PP framework. It engages a classical reflex arc and does not benefit from error correction. A second reply could be that deafferented patients use visual feedback to help guide activation of alpha motor neurons, without Ia afferent proprioceptive feedback. However, as noted above, in some of the relevant cases, deafferented patients' moved without deficit compared to controls, even without exteroceptive feedback. Third, perhaps deafferented patients use visual imagery to help guide activation of alpha motor neurons, without Ia afferent proprioceptive feedback. Unfortunately, visual imagery could not explain improved accuracy of movement. Imagination does not provide relevant feedback correction.

One further reply is worth considering in more detail: perhaps in deafferented patients who exhibit accurate prescribed movement, the proprioceptive generative model has been revised to accommodate deafferentation. For example, suppose that in the deafferented patients in question, the generative

model encoded in the motor hierarchy has been changed so that a high-level action intention (e.g. reaching for a cup) is unpacked into different predictions sent to alpha motoneurons. By altering the predictions sent to alpha motoneurons, the motor hierarchy could generate the same error signal from alpha motoneurons in deafferented patients as in controls, using different downward predictions, even without online proprioceptive or exteroceptive feedback.

The last reply is the most promising. However, it too fails to account for the full range of empirical findings. For one thing, the suggestion entails that there will be a time immediately after deafferentation during which the motor system must relearn the relevant hierarchical relationships. The proposal predicts that at onset of neuropathy the hierarchy will still encode the pre-pathology top-down generative model and will thus issue incorrect motor commands. However, as far as I know, there is no evidence that patients experience any such relearning period. In addition, the proposal fails to account for deafferented patients' retention of some motor abilities while losing others. For example, one deafferented patient, G.O., retained the ability to accurately draw figure 8s in the air but lost the ability to grasp a pen (Rothwell et al. 1982). If the generative model could be revised to allow the subject to draw figure 8s, one would expect it could be revised to allow for grasping a pen. Finally, the proposal fails to account for deafferented patients' ability to retain some skills while being unable to acquire very similar skills after deafferentation. For example, Rothwell et al. (Rothwell et al. 1982, p. 252) write that G.O. "was able to continue driving his old car [used prior to deafferentation] even at night, but found it impossible to learn to drive his new car". If the correct explanation for G.O.'s continued ability to drive the old car were that the generative model in G.O.'s motor hierarchy was radically revised, he should have been able to learn to drive a new car as well. For the revision required to adapt to the new car could not have been sufficiently different from the revision required to retain his old driving abilities. That he could not learn to drive the new car counts strongly against the proposal that the generative model is radically revised to accommodate deafferentation.

## 6 Action Prevents Error

Active inference is not the only way to model action-PEM in a PP framework. On active inference, proprioceptive predictions lead to prediction error, which is then quashed when classical reflex arcs are engaged. However, to account for prediction error minimization through action, one need not assume that error arises first, only to be quashed. Instead, it might be the case that action ensues to prevent prediction error. Call this the *action prevents error* thesis.

The action prevents error (APE) thesis is consistent with widely-accepted approaches to motor control, such as optimal feedback control theory. Recall that in optimal control theory, motor commands project downward to the motor periphery driven by a hierarchical inverse model (or controller), and that each time a motor command is issued, an efference copy is sent to the Bayesian estimator. The estimator then predicts the likely effects of the motor command on the environmental state, which allows the system to predict the likely sensory consequences of the intended action. In a PP framework, the Bayesian estimator can work via downward predictions in a hierarchical generative model, just as downward predictions operate in exteroceptive hierarchies. As a result, adopting an optimal control approach in a broadly PP framework allows us to model efference copy (corollary discharge) using the familiar hierarchical message passing process.

On APE, as in active inference, the motor system initiates downward proprioceptive and exteroceptive predictions of sensory inputs associated with an expected trajectory. Unlike active inference, in APE, downward predictions associated with expected trajectories do not serve as implicit motor commands. There is a separate channel driven by the controller for that. In APE, as in optimal control, downward predictions resulting from efference copy help the motor system make online adjustments to the motor command sequence by providing feedback. However, unlike optimal control theory, we can emphasize in APE that the predictions carried in efference copy also provide a crucial resource for

prediction error minimization in action. Because the predictions are of the sensory inputs associated with intended action, the occurrence of the intended action minimizes error with respect to these predictions. If the intended action did not occur, prediction error would be more pronounced. In APE, action prevents the prediction error that would occur as the result of corollary discharge predictions if the action did not occur. As a result, the APE thesis shows how one could develop an optimal control approach that differs from active inference, but is consistent with the core claim of PP frameworks; i.e. that everything the mind does aims at minimizing prediction error.[8]

The action prevents error approach solves all the problems raised against active inference in the previous sections. It solves the action/revision problem. It can account for the complete lack of proprioceptive experience in mobile deafferented patients. And it can account for deafferented patients' ability to sometimes move in prescribed ways with surprising accuracy without exteroceptive feedback. I now explain each point in turn.

APE solves the action/revision problem. In APE, proprioceptive predictions and motor commands do not travel along the same neural pathways. They are realized by distinct but interactive subsystems. The explanation for why action-PEM or revision-PEM occurs appeals to this architecture. Action-PEM ensues when motor commands initiate action, thereby minimizing prediction error that would arise from the predictions that are made as part of corollary discharge. On APE, there is no competition between action-PEM and revision-PEM in motor pathways. Relatedly, APE explains how extensive proprioceptive revision can occur during actions requiring online adjustment during movement. APE entails that proprioception occurs via distinct pathways from motor control. So proprioceptive predictions can be made constantly during an action, simultaneously with the occurrence of motor commands. The result is that according to APE — but not active inference — proprioception and motor control can proceed simultaneously and interactively, rather than alternating in competition for use of the same pathways.

APE also allows for the complete lack of proprioceptive experience in deafferented patients who can move their limbs. On APE, the motor commands that entrain limb movement are realized by neural pathways distinct from those that are crucial for proprioceptive experience. APE accounts for limb movement in deafferented patients by appeal to the fact that, in such patients, motor pathways remain intact. Moreover, in deafferented patients, the proprioceptive predictions (distinct from motor commands) that would give rise to proprioceptive experience cease to occur. As a result of deafferentation, the afferent signal in the affected areas is eliminated. With no incoming signal, the relevant proprioceptive pathways have nothing to predict, so predictions cease, and with their cessation proprioceptive experience is eliminated with respect to the affected areas.

In addition, APE accounts for the cases in which deafferented patients retain considerable accuracy in prescribed actions, even without feedback from exteroceptive modalities. On active inference, alpha motoneurons encode error as the mismatch between Ia afferent signals and downward predictions from motor cortex. In deafferented patients, there is no Ia afferent signal. As Friston and colleagues note, in such patients, the error encoded by alpha motoneurons is equivalent to the downward prediction, since there's no Ia afferent signal at all. But, as I argued above, this means that the error signal used to engage reflex arcs are very different for deafferented patients compared with normal subjects. It is implausible that the somatomotor hierarchy compensates for such a radical change concerning which error signals occur at the periphery given the same downward predictions, especially given other limitations experienced by deafferented patients in modifying motor routines. The APE approach faces none of these difficulties. On APE, downward motor commands engage muscle fibers directly, without prediction error arising first. So, on APE, the motor commands sent to affected limbs in deafferented patients will be similar to the commands sent in normal subjects. The main difference

8 One could also combine APE with other Bayesian alternatives to active inference, such as paired forward-inverse model approaches (Wolpert and Kawato 1998; Haruno et al. 2003).

will be in the lack of proprioceptive feedback in deafferented patients. The role of feedback on APE does well to explain the results surveyed above. For example, it explains why some deafferented patients perform relatively poorly compared to controls when reaching at slow speeds with their eyes closed. In such cases, the motor systems of normal subjects utilize significant proprioceptive feedback that is not available in deafferented systems. However, at high speeds, feedback plays less of a role in normal subjects. As a result, normal subjects' performance is closer to that of deafferented subjects at high speeds.

In addition to solving the above problems for active inference, APE gains additional support from core commitments in PP frameworks. A familiar challenge for PP is to explain why organisms engage in playful and explorative activity. For example, why do humans climb mountains, explore caves, or attend parties with strangers? Such activity involves lots of novel stimuli and often increases short-term prediction error compared with less playful and explorative alternatives. A plausible and promising reply to this worry is that organisms aim to minimize 'global' prediction error (Clark 2016; Lupyan 2015) or 'long-term' prediction error (Hohwy 2013) under a counterfactually rich set of scenarios the organism might encounter (Seth 2015). That is, predictive minds accept—and even seek out—prediction error in the short term or in some restricted domain in the service of minimizing prediction error over the longer term and over a wider range of scenarios. Giving a central explanatory role to the minimization of long-term and wide ranging prediction error makes sense only if we characterize at least some prediction error minimization in terms of preventing prediction error proactively rather than waiting for it to arise and then quashing it. So a version of APE seems required in PP. According to the version of APE I have developed here, action is thoroughly proactive in its prevention of proprioceptive prediction error: preventing prediction error during action is the primary way in which action-PEM occurs. But it is worth remembering that APE fits well with the general emphasis in PP on proactive global and long-term prediction error minimization.

## 7 Conclusion

In this paper, I have argued for the action prevents error thesis as an alternative to active inference, as developed by Friston and collaborators. I raised three problems for active inference, and I argued that APE solves all three problems. In addition, since APE is compatible with optimal control theories, it can claim any further advantages that such theories might have. Finally, APE fits well with the need to account for long-term and indirect prediction error minimization in PP.

Although I have criticized active inference, I have not argued against the predictive processing framework more broadly. On the contrary, my proposals are fully consistent with the core commitment of PP to the centrality of prediction error minimization in everything the mind does. APE fits well with what Clark 2013; Clark 2016 call 'action-oriented' predictive processing in a broad sense. Despite its recent rise in prominence, predictive processing remains highly controversial in many quarters. One of my goals in this paper has been to show how the core claims of PP are consistent with highly successful approaches to sensorimotor psychology, such as optimal feedback control. Distinguishing the core commitment to prediction error minimization in perception and action from a commitment to active inference as Friston and colleagues have developed it is useful in an intellectual climate where PP remains controversial.

## 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.

Azim, E., Fink, A. J. & Jessell, T. M. (2014). Internal and external feedback circuits for skilled forelimb movement. *Cold Spring Harbor Symposia on Quantitative Biology, 79*, 81–92. https://dx.doi.org/10.1101/sqb.2014.79.024786.

Brown, H., Adams, R. A., Parees, I., Edwards, M. & Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing, 14* (4), 411–427.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 181–204.

——— (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.

Forget, R. & Lamarre, Y. (1987). Rapid elbow flexion in the absence of proprioceptive and cutaneous feedback. *Hum Neurobiol, 6* (1), 27–37.

Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13* (7), 293–301.

——— (2011). What is optimal about motor control? *Neuron, 72*, 488-98.

Friston, K., Daunizeau, J., Kilner, J. & Kiebel, S. J. (2010). Action and behavior: A free-energy formulation. *Biol Cybern, 102*(3), 227-260.

Ghez, C. & Sainburg, R. (1995). Proprioceptive control of interjoint coordination. *Canadian Journal of Physiology and Pharmacology, 73* (2), 273–284.

Gordon, J., Ghilardi, M. F. & Ghez, C. (1995). Impairments of reaching movements in patients without proprioception. *I. Spatial Errors. Journal of Neurophysiology, 73* (1), 347–360.

Haruno, M., Wolpert, D. M. & Kawato, M. (2003). Hierarchical MOSAIC for movement generation. *International Congress Series, 1250,* 575–590. https://dx.doi.org/10.1016/S0531-5131(03)00190-0.

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

Knierim, J. (1997). Motor systems. In J. Byrne (Ed.) *Neuroscience online*. http://neuroscience.uth.tmc.edu/s3/index.htm.

Liu, D. & Todorov, E. (2007). Evidence for the flexible sensorimotor strategies predicted by optimal feedback control. *Journal of Neuroscience, 27* (35), 9354–9368.

Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: Predictive systems are penetrable systems. *Review of Philosophy and Psychology, 6* (4), 547–569.

Messier, J., Adamovich, S., Berkinblit, M., Tunik, E. & Poizner, H. (2003). Influence of movement speed on accuracy and coordination of reaching movements to memorized targets in three-dimensional space in a deafferented subject. *Experimental Brain Research, 150* (4), 399–416.

Metzinger, T. (2017). The problem of mental action. Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Rescorla, M. (2016). Bayesian sensorimotor psychology. *Mind & Language, 31* (1), 3–36.

Rothwell, J. C., Traub, M. M., Day, B. L., Obeso, J. A., Thomas, P. K. & Marsden, C. D. (1982). Manual motor performance in a deafferented man. *Brain, 105* (3), 515–542.

Sainburg, R. L., Poizner, H. & Ghez, C. (1993). Loss of proprioception produces deficits in interjoint coordination. *Journal of Neurophysiology, 70* (5), 2136–2147.

Sainburg, R. L., Ghilardi, M. F., Poizner, H. & Ghez, C. (1995). Control of limb dynamics in normal subjects and patients without proprioception. *Journal of Neurophysiology, 73* (2), 820–835.

Seth, A. K. (2015). The cybernetic Bayesian brain. In T. K. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt am Main: MIND Group. https://dx.doi.org/10.15502/9783958570108.

Shea, N. (2013). Perception versus action: The computations may be the same but the direction of fit differs. *Behavioral and Brain Sciences*, 228–229.

Shipp, S., Adams, R. A. & Friston, K. J. (2013). Reflections on agranular architecture: Predictive coding in the motor cortex. *Trends in Neurosciences, 36* (12), 706–716.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience, 7* (9), 907–915.

——— (2009). Parallels between sensory and motor information processing. In M. S. Gazzaniga (Ed.) *The Cognitive Neurosciences,* 613–623. Cambridge, MA / London, UK: MIT Press.

Todorov, E. & Jordan, M. I. (2002). Optimal feedback control as a theory of motor coordination. *Nature Neuroscience, 5* (11), 1226–1235.

Vance, J. & Stokes, D. (2017). Noise, uncertainty, and interest: Predictive coding and cognitive penetration. *Consciousness and Cognition, 47*, 86–98. http://dx.doi.org/10.1016/j.concog.2016.06.007.

Wolpert, D. M. & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks, 11* (7), 1317–1329.

# Predictive Processing and the Phenomenology of Time Consciousness

## A Hierarchical Extension of Rick Grush's Trajectory Estimation Model

Wanja Wiese

This chapter explores to what extent some core ideas of predictive processing can be applied to the phenomenology of time consciousness. The focus is on the experienced continuity of consciously perceived, temporally extended phenomena (such as enduring processes and successions of events). The main claim is that the hierarchy of representations posited by hierarchical predictive processing models can contribute to a deepened understanding of the continuity of consciousness. Computationally, such models show that sequences of events can be represented as states of a hierarchy of dynamical systems. Phenomenologically, they suggest a more fine-grained analysis of the perceptual contents of the *specious present*, in terms of a hierarchy of temporal wholes. Visual perception of static scenes not only contains perceived objects and regions but also spatial *gist*; similarly, auditory perception of temporal sequences, such as melodies, involves not only perceiving individual notes but also slightly more abstract features (*temporal* gist), which have longer temporal durations (e.g., emotional character or rhythm). Further investigations into these elusive contents of conscious perception may be facilitated by findings regarding its neural underpinnings. Predictive processing models suggest that sensorimotor areas may influence these contents.[1]

**Keywords**

Auditory perception | Bayesian brain | Consciousness | Event segmentation theory | Ideomotor principle | Phenomenology | Predictive processing | Specious present | Time consciousness | Trajectory estimation model

The aim of this chapter is to try to connect research on predictive processing (PP) with research on the phenomenology of time consciousness. The motivation for this comes, on the one hand, from Grush's work on temporal perception,[2] and, on the other, from Hohwy's work on prediction error minimization.

On the first page of his monograph *The Predictive Mind*, Hohwy suggests that "the idea that the brain minimizes its prediction error […] explains not just that we perceive but *how* we perceive: the idea applies directly to key aspects of the phenomenology of perception." (Hohwy 2013, p. 1). Here, I will attempt to apply this idea to key aspects of the phenomenology of *temporal* perception. Luckily, we can build on existing work by Grush (Grush 2005), who has developed a model of temporal perception he calls the trajectory estimation model (TEM). This draws on control and filtering models, but is pitched at a level of abstraction which makes it compatible with specific PP models. A central point for the purposes of this chapter is that TEM does not posit a hierarchy of representations, whereas the types of PP models considered here do. As I shall argue, extending TEM by drawing on features of hierarchical PP models can help account for key aspects of the phenomenology of temporal perception (which are not addressed by Grush's TEM). I shall call the resulting extension of TEM the hierarchical trajectory estimation model (HiTEM).

The paper is structured as follows. In section 1, I briefly review models of temporal consciousness and highlight two features of conscious temporal perception, which I shall call **endurance** and **continuity**. In section 2, I explain the basic aspects of Grush's TEM and formulate a question, which I call the **interface question** and which is not addressed by TEM. Crucially, providing an answer to this

1 I am highly grateful to Martin Butz, Jakob Hohwy, Marius Jung, Thomas Metzinger, Mark Miller, Iuliia Pliushch, and Lisa Quadt for providing a number of very useful comments on drafts of this paper. Thanks to Robin Wilson for excellent editorial help.

2 "Temporal perception" should here be understood as a shorthand for "perception of temporally extended processes or events".

question would be necessary to account for **endurance** and **continuity**. To extend TEM, I then explain core features of a computational model by Kiebel and colleagues (Kiebel et al. 2008a) of how the brain represents temporal sequences (section 3). Generalizing from this model, I develop an extension of TEM: HiTEM (section 4). HiTEM provides an answer to the interface question (as I show in section 5). It also suggests how to account for **endurance** and **continuity** (section 6). In section 7, I offer some tentative remarks on what HiTEM says about the contents of consciousness, considering empirical findings on the neural underpinnings of auditory perception.

## 1 The Phenomenology of Time Consciousness

Experiencing successions of events, as with a series of notes comprising a melody, poses a puzzle. It seems that neither experiencing the different notes simultaneously nor experiencing them in sequence can give rise to the experience of succession. If we experience all notes simultaneously, we experience not a melody but a chord. If we experience first one note, then another, this is a succession of experiences, not an experience of succession (cf. James 1890). So how can we conceive of the experience of successions of events, and of temporally extended processes in general? The two dominant approaches are what Dainton (Dainton 2014) calls *extensional* and *retentional* models, respectively. Interestingly, although these models entail different metaphysical[3] claims about temporal consciousness, they are not necessarily committed to different phenomenological assertions (cf. Dainton 2014, § 3).

According to extensional models, an experience of a succession of events involves a temporally extended experience with proper temporal parts. These correspond to the different temporal parts of the experienced succession of events. For instance, experiencing a succession of two notes involves a *single* experience corresponding to the entire experienced temporal whole (the succession of notes), but the global content of this experience has two temporal parts – one for the first note and one for the second. The notes are experienced as successive, not simultaneous, because the corresponding temporal parts of the total experience are not simultaneous, but successive. In other words, the temporal structure of conscious experience matches the apparent[4] temporal structure of the experienced events (Watzl 2013 calls this the *structural matching thesis*).

According to retentional models, experiencing a succession of events does not always involve a succession of experiences. At least on short timescales, conscious experiences are atomic (cf. Lehmann 2013). Here, "atomic" does not mean that the neural underpinnings are static: This type of atomicity is compatible with the assumption that the neural underpinnings of conscious experiences are always temporally extended (cf. Lee 2014). It just means that the proper temporal parts of a conscious experience cannot be mapped onto different temporal parts of an experienced temporal whole (such as a succession of events). So retentional models reject the assumption that the temporal structure of conscious experience always matches the apparent temporal structure of the experienced events. Still, an experience of a succession has *synchronous* parts which can be mapped onto the different elements of the succession. The parts of the succession are not experienced *as* simultaneous (although they are simultaneously experienced) because the different parts of the experience do not all represent their targets in the same way. As a result, the different events in the succession are represented *as temporally related*. In Husserl's words, events which are just past are represented by *retentions* (cf. Husserl 1991; hence Dainton's label "retentional models").

Disagreements between extensional and retentional models thus mainly concern the metaphysics of our momentary conscious experience (what we are experiencing now, "as present"). According to

3 Metaphysical claims about consciousness deal, for instance, with the relationship between conscious processes and neural activity, or with properties conscious experiences are deemed to have. Phenomenological claims, by contrast, deal with how consciousness appears from the first-person perspective and with the contents of consciousness.

4 This qualification is important to allow for temporal illusions in which, for instance, the actual order of a succession of events is misperceived. The apparent temporal structure would then be the temporal structure *as it is experienced*.

extensional models, momentary conscious experiences have different experiences as proper temporal parts. According to retentional models, they don't (see figure 1 for an illustration).


**Figure 1:** Two conceptions of the specious present: retentional versus extensional models. According to retentional models (the retentional specious present is highlighted in red), different stages of an experienced temporal process are present in consciousness at the same time. According to extensional models (the extensional specious present is highlighted in blue), the conscious experience of a temporal process is itself a temporal process, with proper temporal parts corresponding to the temporal parts of the experienced process. For further details, see the discussion in (Dainton 2014).

What these models agree on is the phenomenological claim that momentary conscious experience constitutes a *specious present* (cf. James 1890). The contents of the specious present comprise an interval extended in time but with parts that are all present (so they are experienced "at the same time", but not *as simultaneous*). James famously affirmed:

> The unit of composition of our perception of time is a duration, with […] a rearward- and a forward-looking end. It is only as parts of this duration-block that the relation of succession of one end to the other is perceived. We do not first feel one end and then feel the other after it, and from the perception of the succession infer an interval of time between, but we seem to feel the interval of time as a whole, with its two ends embedded in it. (James 1890, pp. 609-610)[5]

What the models disagree about is whether the temporal extension of the specious present itself is explanatorily relevant for the phenomenological claim. The extensionalist claims that the specious present can contain an experience of enduring processes or successions of events, because the specious present itself is a succession of conscious experiences, or an enduring conscious experience. The retentionalist, on the other hand, claims that the specious present can comprise an experience of enduring processes or successions of events, because it has synchronous proper parts which are directed at different times (or represent events which are occurring at different times).

An example of such a retentional model is Grush's trajectory estimation model (TEM) – at least it shares the central intuition of this class of models. According to TEM, the content of the specious present can be described as a trajectory estimate, which contains estimates of what is happening at different times.

Combining this basic idea with theoretical research on predictive processing, I argue that a more fine-grained phenomenological analysis of temporal consciousness can be provided: The content of the specious present is not best conceived as a linear stream of events, but rather as a hierarchy of temporal wholes.[6] I try to show that this view can account for two features of temporal consciousness,

5 Some people disagree with this description and claim that the contents of consciousness are more like dynamic snapshots (cf. Prosser 2016; this is compatible with the assumption that the neural underpinnings of such snapshots are always temporally extended). I restrict the treatment in this chapter to accounts which are compatible with the specious present view, to avoid making the discussion unnecessarily complicated.

6 A hint at a similar idea can be found in Thomas Metzinger's *Being No One*: "[C]onvolved holism also reappears in the phenomenology of time experience: Our conscious life emerges from integrated psychological moments, which, however, are themselves integrated into the flow of subjective time." (Metzinger 2004[2003], p. 151).

which are treated as primitive by existing models or left unaddressed. This view can be construed as an extension of TEM, but the central phenomenological analysis (according to which the contents of the specious present consist of a hierarchy of wholes) is compatible with all models that embrace the view that the content of momentary conscious experience comprises an interval (which is common ground between retentional and extensional models). I focus on TEM and not on other retentional, or even extensional, models because TEM is formulated in computational terms, which makes a connection with PP models and further development relatively straightforward.

Conceiving of the contents of the specious present as a hierarchy of temporal wholes can help clarify the following two features of temporal experience:

**Continuity** $=_{Df}$ At least sometimes, we experience smooth successions of events (or smooth changes). An example is a series of notes played *legato* by a single instrument (contrast this with a series played *staccato*). Such sequences are experienced as temporal continua (which, strictly speaking, would involve an infinite number of events).

**Endurance** $=_{Df}$ At least sometimes, we experience temporally extended events as enduring. An example is an opera singer holding a single note for an extended period (this example is taken from Kelly 2005, p. 208). By contrast, when one is surprised by a sudden bright flash, this punctual event is not experienced as part of an enduring event.

These features are not mutually independent. **Continuity** implies **endurance**: When we experience a temporal continuum, we experience a dynamic event, in which a higher-order event is experienced as enduring through change. This idea is not completely new (see Prosser 2016, especially p. 172) and Zacks' event segmentation theory (EST) is related (cf. Zacks et al. 2007), although there are important differences. I explain it in more detail below, having illustrated the two features and shown that they pose a challenge to Grush's TEM. I draw on a computational PP model by Kiebel et al. to provide a theoretical sketch of how the features can be accounted for and I review some empirical results which enrich the proposal.

## 2 The Trajectory Estimation Model (TEM)

TEM is an abstract[7] model of how the brain represents consciously experienced, temporally extended sequences at small timescales (on the order of 200 ms,[8] see Grush 2006, p. 444). A core assumption is that, at such timescales, consciously experienced events are represented as related by temporal relationships such as "earlier than" or "simultaneous with", not as events that are occurring "now" or will occur in the future (cf. Grush 2016, p. 8). Consequently, when one event is represented as occurring earlier than another, this does not entail that one is experienced as less real. Furthermore, the content of perception at a time comprises a trajectory – an ordered tuple of events, not just events occurring

7 The model is abstract in the sense that it does not specify which exact process models are computed by the brain and which exact computational strategies are employed to generate trajectory estimates (cf. Grush 2005, p. S218).

8 Why does Grush assume that this interval has a length of around 200 ms? The assumption is motivated by research on temporal illusions, especially *postdictive phenomena* (cf. Shimojo 2014), in which a percept of a stimulus is influenced by input received around 100-200 ms after the first stimulus presentation. A classic example is a type of apparent motion in which two stimuli of different colors are used (often called "colored phi", see Kolers and von Grunau 1975). When, say, the brief presentation of a green spot is followed by the presentation of a red spot, this can lead to the percept of a moving spot which abruptly changes its color (from green to red). Clearly, this percept cannot be formed before the second stimulus has been processed. This means that the percept is a function of sensory signals obtained over an interval of time. Since apparent motion is perceived when stimuli are separated by an interval of around 100 ms, this suggests momentary conscious perception reaches "into the past". Similarly, research on an effect called "representational momentum" (see Thornton and Hubbard 2002; Hubbard 2014) suggests that momentary conscious perception also reaches "into the future", i.e., it comprises representations of anticipated events (just about to happen, in the very near future). From such results, Grush concludes that conscious perception presents us with events which are currently happening, events which have just happened, and events which are expected to happen, so conscious perception has "a lag and reach on the order of 100 ms each, for a total temporal magnitude on the order of 200 ms." (Grush 2006, p. 444). Note that this is a claim about conscious perception, not about activity in the brain as such. For the purposes of this paper, nothing hinges on the exact temporal extension of the interval. However, Grush's considerations do make it plausible that the interval covers only a fraction of a second.

simultaneously. In particular, it involves *smoothed*[9] and *predicted* estimates respectively of future and past events, to capture the intuition underlying the posit of *retentions* and *protentions* in Husserl's account of time consciousness (see Grush 2006). Representations of perceived events also comprise *filtered* estimates. These combine current sensory information with prior knowledge about the target (see Grush 2008, p. 152). Formally, Grush describes the trajectory estimate as follows:

> With such tools in place, it is possible to describe a system that combines smoothing, filtering and prediction to maintain an estimate of the trajectory of the modeled domain over the temporal interval *[t - j, t + k]*, by determining, at each time *t*, the following ordered $j + k + 1$–tuple: $(\tilde{p}(t-j), \tilde{p}(t-j+1), ..., \hat{p}(t), \bar{p}(t+1), ..., \bar{p}(t+k))$. (Grush 2005, p. S211)

Here, $\tilde{p}$ denotes a smoothed estimate, $\hat{p}$ a filtered estimate, and $\bar{p}$ a predicted (prospective) estimate. As Grush shows, TEM can account for a variety of perceptual illusions (see Grush 2005; Grush 2006; Grush 2008). Since TEM is a model of conscious *perception*, its scope is explicitly restricted to perceptual representations and, more specifically, to perceptual representations of what is currently happening (within an interval of approximately 200 ms).[10] See figure 2 for an illustration.

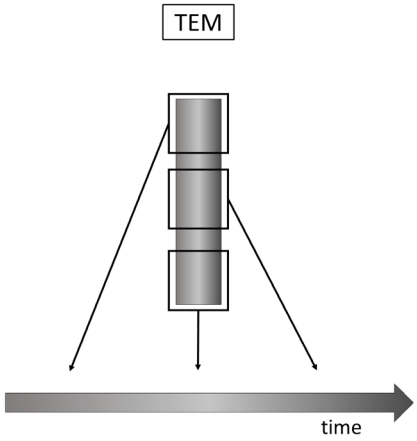


**Figure 2:** The trajectory estimation model (TEM). Estimates of what is happening at different times are computed simultaneously. Hence, what is experienced at a time is an interval (a succession of events, or a temporally extended process).

Intuitively, it should be plausible that there is difference between *perceiving* events that are currently happening and vividly *remembering* events that happenend in the past, or *imagining* events that may happen in the future. According to Grush, these different conscious experiences involve different types of representation. Remembering and imagining involve *conceptual* representations, while experiencing events that are currently happening involves *perceptual* representations:

9 "Smoothing" is the technical term for methods in which an estimate at a given time step is generated by taking measurements obtained after that time step into account. An example is the moving average method, in which an estimate at a time is the average of a set of data (obtained before and after that time).

10 TEM bears an apparent similarity to event segmentation theory (EST). *Event segmentation* refers to the capacity of dividing perceptual streams into meaningful chunks (see Zacks 2008 for a brief introduction). EST posits *event models* to account for this capacity. The apparent similarity to TEM is that an "event model is a representation of 'what is happening now,' which is robust to transient variability in the sensory input." (Zacks et al. 2007, p. 274). Similarly, trajectory estimates in TEM code the contents of the specious present, which is our conscious experience of "what is happening now". There is a huge difference, however, between the temporal grain of events that are relevant to TEM and EST respectively. EST deals with events that have a relatively long duration (several seconds, see Zacks et al. 2001, p. 653; Zacks et al. 2007, p. 274), whereas TEM applies only to events that occur within a fraction of a second. However, some aspects of the information-processing strategy implied by EST may have fruitful connections to the account sketched in this paper. For instance, Zacks et al. "hypothesize that the architecture in EST is implemented simultaneously on a range of timescales, spanning from a few seconds to tens of minutes." (Zacks et al. 2007, p. 276). Similarly, the hierarchical extension of TEM sketched here, HiTEM, posits multiple timescales over which perceptual estimates are computed. Again, however, the relevant temporal grain is much smaller than that referred to by Zacks et al. (and there are some more subtle differences, see section 4 for details).

> There are two ways in which it could be plausibly maintained that contents characterizable only in temporal interval terms play a role in experience. One, which potentially spans a larger interval, might be described as *conceptual* in the sense that it is a matter of interpreting present experience in terms of concepts of processes that span potentially large intervals. Music appreciation would fall into this category. When I recognize something as part of a larger whole (a spatial whole or a temporal whole), then my concept of that whole influences the content grasped via the part. Something along these lines is what appears to be happening with music. On the other hand, there is what might be called a perceptual or phenomenal phenomenon of much brief[er] magnitude. In the music case, the listener is quite able to draw a distinction between some things she is perceiving and some she is not, and notes from a bar that sounded three seconds ago will not typically be misapprehended by the subject as being currently perceived, even though their presence is felt in another, contextual or conceptual sense. (Grush 2006, p. 447)

When I am listening to a piece of music, I can be aware of the temporal context in which the currently sounding notes occur, but, as Grush points out, I do not have the impression that those parts are occurring at the same time as the notes that are currently sounding. It may be debatable whether the term "conceptual" is apt for such representations, but it should at least be plausible that there is a phenomenal difference between perceiving the notes which are sounding *now* and being aware of the notes which sounded a few seconds ago (or that are about to sound). So, for the purposes of this paper, let us stick to Grush's label, and call all non-perceptual conscious representations conceptual. What matters here is that Grush seems to draw a sharp distinction between two types of conscious representation (perceptual versus non-perceptual), and we can use the labels *perceptual* and *conceptual*, respectively, for these types.

I shall argue that a more useful distinction can be drawn by focusing on the timescale at which a representation operates. By this, I mean the temporal extension of the process or event that is represented by a representation. As we shall see in section 3, some PP models posit estimates which track features at different timescales, i.e., features which change more or less quickly (or, conversely, remain invariant for shorter or longer times). In the passage quoted above, Grush already hints at this, when he writes that a representation can be "conceptual in the sense that it is a matter of interpreting present experience in terms of concepts of processes that span potentially large intervals" (Grush 2006, p. 447). A suggestion inspired by work on PP is that events which are currently happening are *always* represented in terms of processes that span intervals of different lengths. Crucially, some of these intervals are shorter than the interval of Grush's TEM, and some are only slightly longer. So when conscious representations are categorized according to the timescale at which they operate, there is no sharp distinction between two types of representation, because there are not only representations operating at very short timescales (Grush's perceptual representations) and representations operating at very long timescales (Grush's conceptual represenations); but there are also *intermediate* representations, which can only arbitrarily be classified as either perceptual or conceptual.

Assuming a sharp distinction between perceptual and conceptual representations would lead to a puzzle when we try to account for **endurance** (and **continuity**). Recall that, according to **endurance**, we sometimes experience temporally extended processes as enduring. We are aware that they have just been present and we are aware that they are still present. If we assume a sharp distinction between conceptual and perceptual representations, some enduring processes would have to be represented by two conscious representations of different types – a conceptual and a perceptual representation. Since these representations are qualitatively different, and since the represented processes are still experi-

enced as identical (it is the *same* process that has occurred and that is still occurring), this raises what I shall call the **interface question**:[11]

> **Interface question:** $=_{Df}$ How are perceptual representations of sequences integrated with conceptual representations of sequences?

This question is not addressed by TEM (because it is only concerned with perceptual trajectory estimates). Before showing how PP can inspire an extension of TEM, i.e., HiTEM, which avoids the **interface question**, let me illustrate how the question is related to **endurance** and **continuity**, to emphasize its relevance. Hopefully, this will also make the explanatory potential of HiTEM more salient. To a first approximation, a phenomenological formulation of the interface question is: How can I experience past and present events as parts of a single temporal horizon (cf. Husserl 1991, p. 29)? How can I experience recent events as being seamlessly connected to present events?[12] In particular, how can a sound I am perceiving right now (as part of the present) be experienced as the same sound I heard in the recent past?[13] When I perceive an enduring sound, I don't simply experience part of it as present and part of it as past. Noë puts it thus:

> What you experience, rather, is, to a first approximation, the rising of the current sounds out of the past; you hear the current sounds as surging forth from the past. You hear them as a continuation. This is to say, moving on to a better approximation, you hear them as having a certain trajectory or arc, as unfolding in accordance with a definite law or pattern. It is not the past that is present in the current experience; rather, it is the trajectory or arc that is present now, and of course the arc describes the relation of what is now to what has already happened (and to what may still happen). In this way, what is present, strictly speaking, refers to or is directed toward what has happened and what will happen. (Noë 2006, p. 29)[14]

Such phenomenological descriptions cannot be accounted for by TEM, since TEM is just a model of the perceived present (which Grush assumes to have a temporal extension of about 200 ms, at least usually). By contrast, Noë refers to perceived processes which have significantly longer extensions, on the order of seconds.[15] These processes are still experienced as seamlessly connected, as enduring. This is why the **interface question** arises in this context, and why it is beyond the scope of TEM.

Let us now consider some core assumptions underlying PP models. In particular, we shall focus on models of temporal sequence generation and recognition.

## 3 Hierarchical Models of Sequence Recognition

Stefan Kiebel and colleagues have recently developed computational models of phenomena involving the representation of sequences, including recognition of bird songs (cf. Kiebel et al. 2008a; Yildiz and

11 Another question is what one could call *the flow question*: What accounts for the experienced temporal flow of events, and for differences in the speed of the flow? A PP-inspired answer to the flow question has been proposed by Hohwy and colleagues (Hohwy et al. 2016).

12 Again, an example of an event not experienced as seamlessly connected to past events is a sudden, surprising flash.

13 Note that "experiencing as the same as" is different from "experiencing as being seamlessly connected to". The first description refers to what I am calling **endurance** here, the second to the feature of **continuity**. When an event is experienced as enduring, distinct temporal parts are experienced as belonging to a single event (such as notes experienced as part of a melody). A continuous sequence is experienced when, in addition, no temporal gaps or boundaries between the individual parts are experienced (e.g., when a melody is played *legato*, as opposed to *staccato*). Thanks to Jakob Hohwy for pressing me to clarify this.

14 This idea, that experienced events have something like a continuous tail which extends into the past, can also be found in Husserl's work: "During the time that a motion is being perceived, a grasping-as-now takes place moment by moment; and in this grasping, the actually present phase of the motion itself becomes constituted. But this now-apprehension is, as it were, the head attached to the comet's tail of retentions relating to the earlier now-points of the motion." (Husserl 1991, p. 32).

15 The claim that processes on the order of a few seconds are also consciously perceived as integrated wholes is empirically supported by a variety of findings, including speech segmentation, short-term memory tasks, and sensorimotor tasks (for a review, see Pöppel 1997, Pöppel 2009).

Kiebel 2011) and of artificial speech (cf. Kiebel et al. 2009). In these models, sequences (trajectories) are not modeled as successions of events, but as states of "a collection of hierarchical, dynamical systems, where slower environmental changes provide the context for faster changes" (Kiebel et al. 2008a, p. 2). An interesting aspect of such models is that they are compatible with TEM but are more specific — specifying that trajectories are represented hierarchically. Furthermore, they are still neurally plausible, because the cortical hierarchy seems to match the temporal hierarchy entailed by these models (for a review of neuroscientific evidence, see Kiebel et al. 2008b).

These models presuppose the first six features and the ninth of predictive processing (as defined in Wiese and Metzinger 2017). Of these, the **ideomotor principle** and **hierarchical processing** are particularly important. Both are combined with **prediction error minimization**. Let me explain each in turn.

## 3.1 The Ideomotor Principle

An assumption underlying the models of Kiebel and colleagues is that the perception (recognition) of a sequence is enabled by a model of its generation. The idea applies not only to sequences the subject herself can generate (like the movement of an arm), but also to others (like a falling snowflake perceived by a subject). For sequences that can actually be generated by the subject, like bodily movements, this entails that at least some of the representations which are active when the subject is perceiving the sequence are also active when the subject herself is performing such movements. Following William James (James 1890), we can call this the *ideomotor principle* (see Wiese and Metzinger 2017 and Limanowski 2017, for more details, and Wiese 2016a for a discussion in the context of active inference). Regarding the neural underpinnings of perception, this suggests that areas not ordinally regarded as sensory may influence the contents of conscious perception (more on this in section 7).

## 3.2 Hierarchical Processing

A further assumption is that a given sequence can be modeled as a hierarchy of dynamical systems, where the output of a dynamical system at a given level functions as a control parameter[16] for the system at the level below. This principle is illustrated in figure 3. The output at the lowest level corresponds to the sensory consequences of the sequence (those signals are received by the perceiving subject); all other states are hidden and have to be inferred.

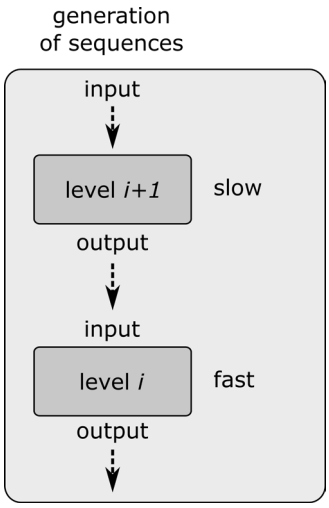


**Figure 3:** Hierarchical processing: sequences are construed as the states of a hierarchy of coupled dynamical systems. The states of the systems change at different speeds, i.e., they operate on different timescales.

16 A control parameter is a parameter which can change the phase space of a dynamical system continuously or discontinuously. For instance, it can determine whether the system has a fixed-point or a chaotic attractor, and continuous changes in control parameters can lead to discontinuous changes in phase space (these are called *bifurcations*, cf. Arrowsmith and Place 1998[1992], p. 224).

More formally, the essential aspects can be captured as follows (here, the hierarchy has only two levels; the equations are simplified versions of the ones found in Kiebel et al. 2008a, p. 3):

$\dot{x}^{(2)} = f^{(2)}(x^{(2)}$, "input from level above", "slow") + noise[17] (1)

$\dot{x}^{(1)} = f^{(1)}(x^{(1)}$, "input from level above", "fast") + noise (2)

The variables $x^{(1)}$ and $x^{(2)}$ describe the states of two dynamical systems, which unfold according to the differential equations (1) and (2), respectively. These equations each have a parameter governing how quickly the respective system changes ("slow" versus "fast"). Furthermore, how $x^{(1)}$ and $x^{(2)}$ evolve depends on input from the level above (here, the input to the second level is a constant). In the example in (Kiebel et al. 2008a), both dynamical systems are *Lorenz systems*.[18] Lorenz systems can have different types of attractors, depending on their *Rayleigh number*:

> We coupled the fast to the slow system by making the output of the slow system [...] the Rayleigh number of the fast. The Rayleigh number is effectively a control parameter that determines whether the autonomous dynamics supported by the attractor are fixed point, quasi-periodic or chaotic (the famous butterfly shaped attractor). (Kiebel et al. 2008a, p. 3)

The Rayleigh numbers are denoted by "input from the level above" in equations 1 and 2. Coupling two dynamical systems in this way already enables very complex dynamics. For instance, the authors use these equations to simulate birdsongs. Crucially, they simulate not only the generation of a birdsong but also the recognition of the song, exploiting the first principle mentioned above, the ideomotor principle. The principle entails that a recognizing system uses a model of how the song has been generated. Ideally, this model contains the same differential equations that describe how the song has actually been generated and can thus be used as a representation of the song (see Wiese 2016b, sections 3 and 4, for a general description of how such models can be used as representations). Recognition is also based on a third computational principle: prediction error minimization.

## 3.3 Prediction Error Minimization

Prediction error minimization is here used as a generic term for computational methods in which prediction error terms are minimized. One such method is predictive coding, originally a strategy to compress data (cf. Shi and Sun 1999). The idea is that if we want to transmit data $d_1$ and $d_2$ from $A$ (the sender) to $B$ (the receiver), we can reduce the amount of data if we exploit informational relations between $d_1$ and $d_2$. For instance, if $d_2$ is highly predictive of $d_1$, we can just transmit $d_2$, and let the receiver infer $d_1$ (based on $d_2$). But what does it mean that $d_2$ is predictive of $d_1$? A general answer is that $d_1$ is a mathematical function of $d_2$. So if we know $d_2$ and the functional relation $d_1 = f(d_2)$, we can compute $d_1$. Hence, the amount of data needed to transmit $d_1$ and $d_2$ from $A$ to $B$ can be reduced.

In a slightly more realistic setting, there would be more than two pieces of data (e.g., the pixels of an image), so there would be, say, data $d_1$ and $d_2$ for which the function relating $d_1$ and $d_2$ would not yield a completely accurate estimate of $d_1$ when applied to $d_2$. For instance, instead of transmitting the values of all pixels of an image, the sender could transmit only a subset of the pixels, as well as a prediction error which tells the receiver how to correct any errors. Clark (Clark 2013, p. 182) attests:

> In most images, the value of one pixel regularly predicts the value of its nearest neighbors, with differences marking important features such as the boundaries between objects. That means that the code for a rich image can be compressed (for a properly informed receiver) by encoding only the "unexpected" variation: the cases where the actual value departs from the predicted one. What

17 The "noise" terms capture any unpredictable influences on $x^{(1)}$ and $x^{(2)}$, i.e., they reflect the uncertainty about the respective estimated variables.

18 A Lorenz system is a set of ordinary differential equations which can have the famous butterfly-shaped attractor.

needs to be transmitted is therefore just the difference (a.k.a. the "prediction error") between the actual current signal and the predicted one.

In other settings, the mapping between different variables will be non-deterministic. This means that there is some uncertainty about our computation of $d_1$; we can only compute an estimate that is more or less reliable. More formally, we can describe this as follows (where, again, the "noise" terms capture all unpredictable influences):

$$d_1 = f(d_2) + \text{noise} \qquad (3)$$

Given, $d_2$ and $f$, we can thus compute an estimate $\hat{d}_1 := f(d_2)$. Depending on the level of noise (or, in general, uncertainty), there will again be a prediction error (because $d_1$ is not equal to $f(d_2)$). If the sender knows the exact values of $d_2$ and $d_1$, the sender can again transmit the prediction error, which will allow the receiver to compute the exact value of $d_1$.[19]

When it comes to the problem of recognition (perception), things are even worse, because the recognizing system only receives sensory signals. To simplify, say the sensory signals are given by the value of $x^{(1)}$ (this corresponds to $d_1$). There are two important differences from the situation above. The second value, coded by $x^{(2)}$, cannot be computed from $x^{(1)}$ (at least not given equations (1) and (2)). Furthermore, the recognizing system does not receive a prediction error, but only sensory signals. The solution to this problem is to give the idea a twist. The recognizer does not simply compute an estimate of the value of $x^{(1)}$, but first estimates $x^{(2)}$; this estimate is then used to compute a prediction of $x^{(1)}$ (using equation (2)), and this prediction is compared to the actual signal coded by $x^{(1)}$. A prediction error can then be used to update the estimate of $x^{(2)}$.

This third feature thus exploits the other two features mentioned above, i.e., the ideomotor principle and hierarchical processing. The ideomotor principle entails that recognition of a sequence is based on a model of how the sequence is generated, which enables a prediction of sensory signals. In the simple example given here, we only have two layers, but the same principle can be applied to systems with a more complex hierarchy. Figure 4 illustrates the basic idea (with just two layers).

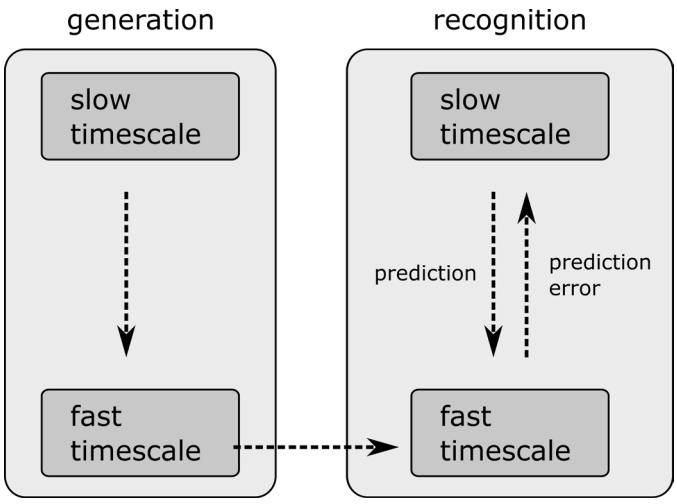


**Figure 4:** Recognition of a sequence is based on a model of its generation and implemented using prediction error minimization. Processes at higher levels of the hierarchy operate at slower timescales than processes at lower levels. The arrow at the bottom represents sensory signals received by the recognizing system. Processes in the box on the left-hand side are hierarchically coupled dynamical systems (cf. equations (1) and (2)). Processes in the box on the right-hand side model

19 Here, $d_1$ is a random variable, because it is a non-deterministic function of $d_2$. In the example given, it is assumed that the sender knows the exact value of $d_2$ (which may be a deterministic variable) and has access to a sample, which is modelled as a particular outcome of $d_1$. This is why, at least in this toy example, the sender can compute the prediction error, although it requires knowledge about the "noise" term, which is by definition unpredictable.

these dynamical systems and therefore enable hierarchical prediction error minimization, which ideally helps keep the model accurate.

# 4 The Hierarchical Trajectory Estimation Model (HiTEM)

## 4.1 From TEM to HiTEM

Having described the main aspects of a predictive processing model of sequence perception, we can generalize the model and combine it with Grush's TEM. Recall that the essential part of TEM is a trajectory estimate (which combines smoothing, filtering, and prediction) over the temporal interval $[t - j, t + k]$:

$$T := (\tilde{p}(t-j), \tilde{p}(t-j+1), \ldots, \hat{p}(t), \tilde{p}(t+1), \ldots, \tilde{p}(t+k)). \text{ (cf. Grush 2005, p. S211)} \quad (4)$$

The challenge now is how to combine TEM with hierarchical models. Two general options are the following:

**Localized** $=_{Df}$ The trajectory estimate $T$ coding the perceptual contents of the specious present corresponds to the state of a dynamical system represented at a specific (single) level of the hierarchy.

**Distributed** $=_{Df}$ The trajectory estimate $T$ is distributed across at least two levels of the hierarchy.

The simplest version of **localized** would be a two-layer hierarchy in which sensory signals are found at the bottom layer, and the trajectory estimate is located at the second layer. A slightly more complex version would involve more than two layers, but the trajectory estimate coding the perceptual contents of the specious present would still be found at a single level. Note that the neural activity coding the value of $T$ could still be parallel distributed processing, but not over different levels of the processing hierarchy. What **localized** entails is that, given a hierarchical model like the one described in (Kiebel et al. 2008a), which specifies a hierarchy of dynamical systems, there is exactly one level of the hierarchy such that $T$ corresponds to the state of the dynamical system at that level.

As an illustration, consider the following statement by Andy Clark (without implying that Clark would endorse **localized**): "Just as the higher levels in a shape-recognition network respond preferentially to invariant shape properties (such as squareness or circularity), so we should expect to find higher-level networks that model driving sensory inputs (as filtered via all the intervening levels of prediction) in terms of tomatoes, cats, and so forth." (Clark 2012, p. 762). One (though not the only) way to interpret this is that most levels of the PP hierarchy process information unconsciously but at one level *it all comes together* (as in the Cartesian Theater, cf. Dennett and Kinsbourne 1992, p. 183), and this is where information is processed consciously. Again, I would not interpret Clark in this way, but the quotation is at least suggestive, and it is not obviously incoherent to claim that the contents of consciousness are coded at a single level of the hierarchy. This means that localized cannot be dismissed without further argument.

**Distributed**, by contrast, entails that the description of the trajectory estimate in equation (4) may not map neatly to the estimates over which computations are carried out in the predictive processing hierarchy. So if the states of hierarchically nested dynamical systems can be described by variables $x_1$, $x_2$, $x_3$, …, it is not the case that $T$ corresponds to the value of exactly one $x_i$. Instead, $T$ corresponds to

the states of at least two dynamical systems in the hierarchy. This means that a more detailed description of $T$ could look like this:

$$\begin{pmatrix} T^{(2)} \\ T^{(1)} \end{pmatrix} := \begin{pmatrix} \tilde{p}^{(2)}(t-j) & \tilde{p}^{(2)}(t-j+1) & \ldots & \hat{p}^{(2)}(t) & \bar{p}^{(2)}(t+1) \ldots \bar{p}^{(2)}(t+k) \\ \tilde{p}^{(1)}(t-j) & \tilde{p}^{(1)}(t-j+1) & \ldots & \hat{p}^{(1)}(t) & \bar{p}^{(1)}(t+1) \ldots \bar{p}^{(1)}(t+k) \end{pmatrix} \quad (5)$$

Note that this hierarchical trajectory estimate is just a "doubled" version of Grush's trajectory estimate and hence does not differ significantly from it. In particular, it does not yet capture the essential part of the hierarchical architecture – that the timescales on which the different levels operate are different. To make this formally explicit, let me adopt a notational convention proposed by Grush:

> I will let $\hat{p}$ stand for a perceptual representation ($p$, without a hat, will stand for the domain that is being represented), and will indicate the time that the representation represents and the time that the representation is produced by two subscripts separated by a slash, so that $\hat{p}_{a/d}$ is notation for a perceptual representation produced at time $d$ that represents what is (/was/will be) happening at time $a$. This notation can be generalized to intervals: $\hat{p}_{[a,c]/[d,f]}$ will stand for a perceptual representation of what happened over the interval $[a,c]$ that is produced over the interval $[d, f]$. (Grush 2008, p. 151)

For the discussion at hand, the represented time is more important than the time of representing (note that in TEM, the different elements of the trajectory estimate are produced at the same time). For this reason, I will drop the reference to the latter. As in equation (5), any reference to times will be to the represented time. So $\hat{p}_{[a,c]}$ is an estimate of what is happening over the interval $[a, c]$. If we let $t_{-(j+1)} < t_{-j} < t_{1-j} < t_{2-j} < \cdots < t_{-1} < t_0 < t_1 < t_2 < \cdots < t_{k-1} < t_k < t_{k+1}$, we can define a more interesting distributed version as follows ($T_d$ for "distributed trajectory estimate"):

$$T_d := \begin{pmatrix} T_d^{(2)} \\ T_d^{(1)} \end{pmatrix} := \begin{pmatrix} \tilde{p}^{(2)}_{[t_{-(j+1)},t_{2-j}]} & \tilde{p}^{(2)}_{[t_{2-j},t_{5-j}]} & \cdots & \hat{p}^{(2)}_{[t_{-1},t_2]} & \bar{p}^{(2)}_{[t_2,t_5]} & \cdots \bar{p}^{(2)}_{[t_{k-2},t_{k+1}]} \\ \tilde{p}^{(1)}_{[t_{-j},t_{1-j}]} & \tilde{p}^{(1)}_{[t_{1-j},t_{2-j}]} & \cdots & \hat{p}^{(1)}_{[t_0,t_1]} & \bar{p}^{(1)}_{[t_1,t_2]} & \cdots \bar{p}^{(1)}_{[t_{k-1},t_k]} \end{pmatrix} \quad (6)$$

The essential difference between this estimate and the estimate in equation (5) is that the represented times are different on the two levels: In equation (6), the intervals on level two are longer than the intervals on level one. So the events represented at the second level have a longer duration than the events on the first level. An informal illustration of this idea can be found in figure 5:

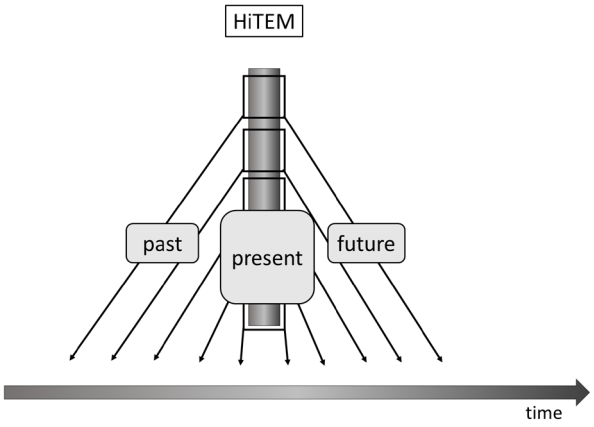


**Figure 5:** A hierarchical extension of TEM (HiTEM). The core feature of HiTEM is that it posits a hierarchy of temporal wholes. A phenomenological prediction is that slowly-changing features (which remain invariant for more than 200 ms) can also contribute to the perceptual contents of the specious present. Most of these representations do not represent

individual events in time, but represent features which remain invariant over shorter or longer timescales. Some of these representations only refer to the present, but there are at least some which represent features we experience as past, present, *and* future – or, rather, we experience some of them as rising from the recent past, and as continuing into the near future, because they remain invariant over an interval which is slightly longer than the interval of what we experience as happening *now*.

## 4.2 HiTEM and Event Segmentation Theory

Here, we encounter a similarity to event segmentation theory (EST), which has been developed by Jeffrey Zacks and colleagues. According to EST, the brain constructs *event models* at different temporal scales. Crucially, a given event model can remain active even in the presence of changing perceptual input:

> For example, an event model for a tooth-brushing might include information about the water cup and toothbrush and their locations, the goal of cleaning teeth, and the person doing the brushing. For event models to be useful, they need to be stable in the face of moment-to-moment fluctuations in sensory input – the cup needs to remain in the model even if it is temporarily occluded. (Tversky and Zacks 2013, pp. 89 f.)

Similarly, since the elements of $T_d^{(2)}$ represent features which remain invariant over longer intervals than those represented by $T_d^{(1)}$, the estimate comprised by $T_d^{(2)}$ will often be stable in the face of changes in the estimates contained in $T_d^{(1)}$. Furthermore, $T_d$ contains a hierarchy of representations (restricted to two levels here for the sake of simplicity). This is in line with EST's assumption that events are segmented hierarchically. The represented temporal boundary of an event often corresponds to the represented achievement of a goal; since many tasks can be divided into subtasks (with sub-goals), a given event is typically represented as consisting of a sequence of smaller events (cf. Tversky and Zacks 2013, p. 87).

At this point, it will be helpful to point out two marked differences between the hierarchy of event models posited by EST and the hierarchy of trajectory estimates posited by HiTEM:

1. Event models operate at significantly longer timescales than the estimates posited by TEM and HiTEM.
2. The notion of an event used by Zacks et al. is more restricted than the notion that is relevant to the contents of the specious present. This has some subtle implications which become salient when it comes to accounting for **endurance** and **continuity**. In particular, events need not be represented as having determinate boundaries, according to HiTEM.

Let us call this second feature **fuzzy boundary**:

> **Fuzzy boundary:** $=_{Df}$ At least at very short timescales (on the order of 200 ms), some events are not represented as having determinate boundaries, and some processes are not represented as having a determinate beginning or ending.

To clarify the second point, let us consider in more detail what the elements of $T_d$ represent. In which sense are they representations of events? Arguably, the most useful and conservative interpretation is as follows. The elements of $T_d$ do not represent events (in the sense of EST) but property instantiations. At least according to certain conceptions of events, an event is just a property instantiation at a given time and place (or the exemplification of a property by an object at a time, see Kim 1966, p. 231). The

crucial point is that a property represented at the second level and a property represented at the first level can correspond to a single event. So the estimate

$$\begin{pmatrix} \hat{P}^{(2)}_{[t_{-1},t_2]} \\ \hat{P}^{(1)}_{[t_0,t_1]} \end{pmatrix}$$

can represent a single event. Let us call such an event representation a dynamic event representation (DER).

Usually, dynamic events are posited to accommodate the intuition that events can involve change (in fact, some authors would claim that something which does not contain any change *cannot* be an event, see Casati and Varzi 2015, § 2.2). A DER represents change in a minimal form: There is at least one property which is not instantiated during the entire interval in which the event unfolds. A stronger form of change would be a succession of property instantiations. One could object that such a succession would correspond to a succession of events (in which case a DER would just be a representation of a succession of events – just as a higher-order event model in EST models an event which is just a succession of lower-order events). But a DER does not necessarily involve such successions. The key theoretical advantage of the concept is that the times during which the different properties are represented as being instantiated can overlap. This opens the interesting possibility that represented events can overlap in time as well.[20]

Consider the following two estimates:

$$\hat{e}_1 := \begin{pmatrix} \hat{P}^{(2)}_{[t_{-1},t_2]} \\ \hat{P}^{(1)}_{[t_0,t_1]} \end{pmatrix}, \hat{e}_2 := \begin{pmatrix} \hat{P}^{(2)}_{[t_{-1},t_2]} \\ \hat{P}^{(1)}_{[t_1,t_2]} \end{pmatrix}.$$

If these estimates are elements of the same trajectory estimate, they can share a part, $\hat{p}^{(2)}_{[t_{-1},t_2]}$. This does not make sense in all cases: If two red balls bounce at the same time at two different locations, there is a sense in which these events overlap (because they share the property of redness), but the two bouncing events are clearly distinct (the property of redness is exemplified by two distinct objects). But there are cases in which it can make sense to represent distinct events as sharing a single property instantiation.

Take the example of music. When a single instrument like a flute plays a sequence of notes *legato* (as opposed to *staccato*), some properties remain invariant (at least during short intervals), for instance the timbre of the notes. Hence, we can describe the melody as a sequence of overlapping property instantiations: Properties like pitch may change more quickly than properties like loudness or timbre. This means we can have an interval $[t_1, t_2]$ during which, say, two different pitches are instantiated (e.g., the first during the first half of the interval, the second during the second half), and a single timbre is instantiated (during the entire interval). Since the timbre is in fact the same timbre during the interval (after all, the notes are produced by the same instrument), it makes sense to use a single representation for this property instantiation. The key difference with the case of the two bouncing balls is that there is a common cause underlying the sensory signals.[21] Furthermore, using a DER not only makes computational sense (because it is more efficient than representing the same property twice); it may also account for the experienced continuity (e.g., in music perception). Let us explore this by first considering how this proposal can deal with the interface question.

20 A perhaps controversial implication is that some events are not represented as having a determinate beginning and ending. But wouldn't it be important to know the exact time at which an event starts (and ends, respectively)? If we think of *intentional binding* (cf. Haggard et al. 2002), in which the interval between two causally connected events is systematically underestimated, it seems there are cases in which the exact timing of events does not matter. Instead, it seems more important to represent temporally distinct events as parts of a temporal whole (perhaps as having a common cause). So representing events as overlapping (with indeterminate temporal boundaries) may be a way of prioritizing the temporal whole over its parts.

21 At a deeper level, there can of course still be a common course, for instance, if both balls are thrown by a juggler (thanks to Jakob Hohwy for this example). This would be comparable to a case in which two instruments (e.g., a flute and a drum) were being played by the same person.

### 4.3 Answering the Interface Question

Recall that we are still considering how to combine TEM with hierarchical predictive processing models, and the crucial challenge is whether we should favor a **localized** or a **distributed** option. In this section, I argue that a hierarchically distributed extension of TEM (involving something like $T_d$, i.e., an estimate with at least two levels, in which properties at different levels of temporal granularity are represented) provides an answer to the **interface question**. Recall the formulation of the latter:

> **Interface question** $=_{Df}$ How are perceptual representations of trajectories integrated with conceptual representations of trajectories?

HiTEM answers the question thus:

- There is no compelling reason to assume a sharp boundary between perceptual (concrete, perspective-dependent) and conceptual (abstract, perspective-invariant) representations.
- In PP, the continuum between perceptual and conceptual representations is typically assumed to be distributed over the hierarchy.

This suggests that clearly perceptual and clearly conceptual representations are found at different levels of the hierarchy. Furthermore, if there are neural representations that are neither purely perceptual nor purely conceptual, these could function as mediators between (conceptual) representations of remembered events and (perceptual) representations of currently occurring events.[22] A theoretical advantage of the distributed option is that the neural vehicles of perceptual and conceptual trajectory estimates can overlap spatially (i.e., by sharing parts), so mediating estimates would not have to be posited as additional representations, but would partly determine both perceptual and conceptual experiences of temporal wholes (I explore this idea in a much wider context in Wiese 2017).

Let me make more explicit what a mediating representation would be in this context. According to TEM, all elements of the trajectory estimates represent events that are occurring within the interval of the specious present (which has, according to Grush, a duration of about 200 ms). All these events are experienced as currently happening; they have features which are represented as being instantiated during this interval. By contrast, a mediating representation in HiTEM represents features as being instantiated during an interval which is longer than that identified by Grush: It represents features as having been present in the recent past, as being present now, and as continuing into the near future. Crucially, such features can be bound to features which are represented as changing more quickly. The result is a dynamic event representation (DER), which corresponds to the experience of an event as present, but also as having been present in the recent past. On the other hand, such mediating features can also be bound to features which are represented as changing more slowly (or features pertaining to events which are represented as past). The result is again a DER, but more akin to what Grush describes as a conceptual (as opposed to a perceptual) representation. Since there is no sharp boundary between purely perceptual and purely conceptual representations, instances of these two types of representation can be integrated by mediating representations.

## 5 How Can We Account for Continuity and Endurance?

Let us next consider to what extent mediating representations (operating at intermediate timescales, between clearly perceptual and clearly conceptual repsentations) can help account for **continuity** and **endurance**. Let me repeat the definitions of these two features of temporal consciousness:

22 Just as there can be representations that mediate between purely perceptual (descriptive) representations and (prescriptive) goal representations (cf. Wiese 2014).

**Continuity** $=_{Df}$ At least sometimes, we experience smooth successions of events (or smooth changes). An example is a series of notes played *legato* by a single instrument (in contrast with a series played *staccato*). Such sequences are experienced as temporal continua (which, strictly speaking, would involve an infinite number of events).

**Endurance** $=_{Df}$ At least sometimes, we experience temporally extended events as enduring. An example is an opera singer holding a single note for an extended period (this example is taken from Kelly 2005, p. 208). By contrast, when one is surprised by a sudden bright flash, this punctual event is not experienced as part of an enduring event.

The answer given to the **interface question** already suggests how to account for **endurance**: when an event is represented by a conscious DER, some of its properties are represented as remaining the same while others are changing, which corresponds to the experience of an event as present, but also as having been present in the recent past; in other words, a conscious DER represents an event as *enduring*. Not all its features are however experienced as having been present in the past, and this is why it can be so difficult to describe our experience of enduring events phenomenologically. An example from Kelly provides an excellent illustration:

> There you are at the opera house. The soprano has just hit her high note – a glass-shattering high C that fills the hall – and she holds it. She holds it. She holds it. She holds it. She holds it. She holds the note for such a long time that after a while a funny thing happens: You no longer seem only to hear it, the note as it is currently sounding, that glass-shattering high C that is loud and high and pure. In addition, you also seem to hear [...] something about its temporal extent. (Kelly 2005, p. 208)

This "something" is, according to HiTEM, a slightly more abstract feature of the note, which is represented as being invariant for more than 200 ms (perhaps even a few seconds). One's conscious experience will certainly have other, additional aspects which characterize what it is like to hear such an enduring high C (for instance, a feeling of tension or stress). But at least some aspects correspond to perceptual (or quasi-perceptual) features of the note which change slowly.[23]

Such features are, according to HiTEM, always experienced, but they are not always very salient. For instance, to most people hearing a melody, it seems obvious that what they are perceiving is not just one note after the other; but to describe what exactly it is that makes the difference might seem more difficult. HiTEM suggests that the additional experienced features are slightly more abstract (more gist-like) than features such as pitch or loudness (and hence more difficult to describe). Crucially, the additional features contribute to the perception of each individual note; since these features are shared by all of them, temporally separated notes can be experienced as a temporal whole, as flowing into each other. This accounts for **continuity**.

Let us compare the proposal again with EST. A hierarchy of event representations in EST would not necessarily involve a representation of a continuous flow (the event models in EST seem to be more abstract, purely conceptual representations of events). The temporal boundaries of events are assumed to be determinate in EST, so even if a tooth-brushing event is represented as a succession of shorter events (brushing the first tooth, brushing the second tooth, …), this would still only be a succession of events: First is A, then B, and both jointly constitute an event C.

23 As Kiebel et al. 2008a point out (with respect to their model of birdsong), such features can also provide information about the creature which generated the temporal sequence: "Birdsong contains information that other birds use for decoding information about the singing (usually male) bird. It is unclear which features birds use to extract this information; however, whatever these features are, they are embedded in the song, at different time-scales. For example, at a long time-scale, another bird might simply measure how long a bird has been singing, which might belie the bird's fitness. At short time-scales, the amplitude and frequency spectrum of the song might reflect the bird's strength and size." (Kiebel et al. 2008a, p. 2). Thanks to Jakob Hohwy for suggesting this citation.

By contrast, a DER would represent a succession of events that do not have determinate temporal boundaries as follows.

$$\begin{pmatrix} & \hat{p}^{(2)}_{[t_{-2},t_2]} \\ \hat{p}^{(1)}_{[t_{-1},t_0]} & \hat{p}^{(1)}_{[t_0,t_1]} \end{pmatrix}$$

Here, the entire matrix represents, say, a succession of notes, but $\hat{p}^{(1)}_{[t_{-1},t_0]}$ & $\hat{p}^{(2)}_{[t_{-2},t_2]}$ jointly constitute a single representation of a note (the first note in the succession), and $\hat{p}^{(1)}_{[t_0,t_1]}$ & $\hat{p}^{(2)}_{[t_{-2},t_2]}$ likewise (the second note in the succession). On the one hand, the first note is represented as occurring before the second, because $\hat{p}^{(1)}_{[t_{-1},t_0]}$ and $\hat{p}^{(1)}_{[t_0,t_1]}$ represent properties (say, pitch) as being instantiated during distinct intervals ($[t_{-1},t_0]$ and $[t_0,t_1]$, respectively). It is not true, however, that the first note is represented as occurring completely before the second, because the other property associated with the notes (say, timbre), which is represented by $\hat{p}^{(2)}_{[t_{-2},t_2]}$, is represented as being instantiated during a longer interval. Hence, the notes are represented as being distinct, but overlapping (where the overlapping part is not just a further note). This is why the entire representation is not just a representation of two events, or of a succession of events, but of a continuous succession, where one event flows smoothly[24] into the next.

## 6 What Are the Contents of Mediating Representations?

Recall that Grush draws a rather sharp boundary between perceptual and conceptual representations. By constrast, assuming that the contents of the specious present are coded by a hierarchy of representations, it is already suggestive to believe that there are mediating representations. If they contribute to the contents of consciousness, however, it will be relevant to determine their contents. I alluded to the example of auditory perception and suggested that examples of mediating representations could include representations of timbre or rhythm. To explore more options, and to make first steps towards finding neural evidence for such representations, let us consider results from empirical research on auditory processing in the brain.

As Lima et al. (Lima et al. 2016) point out in a recent review, neural processing of auditory information is distributed over anatomically and functionally different streams, which can broadly be divided into an anteroventral "what" pathway and a posterodorsal "how/where" pathway (cf. Lima et al. 2016, p. 530). Interestingly, whereas the hierarchy in the "what" pathway seems to provide more and more abstract re-representations of semantic information, the "how/where" pathway seems to provide sensorimotor representations, involving also supplementary motor areas (SMA) and pre-supplementary motor areas (pre-SMA). Furthermore, SMA and pre-SMA not only play a role in speech perception, but also in music perception and auditory imagery (cf. Lima et al. 2016, p. 532). The authors hypothesize that these "regions mediate spontaneous motor responses to sound, and support a more controlled generation of sensory predictions based on previous sensorimotor experience, predictions that can be flexibly exploited to enable imagery and optimize a variety of perceptual processes." (p. 539). This hypothesis suggests that SMA and pre-SMA contain the kind of sensorimotor representations which are posited by the ideomotor principle and which are required if indeed the perception of the (auditory) sequence is based on a model of its generation.

So, given that activity in these regions correlates not only with motor or cognitive processes (for evidence, see the references cited in Lima et al. 2016, p. 534), we can speculate that these regions harbor mediating representations, which are not purely perceptual (becaue they are also relevant for motor tasks) but still correlate with consciously experienced perceptual contents (which may be gist-like). The evidence presented by Lima et al. seems to be consistent with this hypothesis, but more work will

24 Note that this also involves computationally smoothed estimates, but this computational technique does not account for the smoothness of the flow (because trajectory estimates in TEM involve smoothed estimates as well, without thereby accounting for the experienced smoothness).

have to be done to find stronger support for it (for instance, it is not clear whether activity in SMA and pre-SMA correlates with *conscious* perception; cf. Repp 2001). Bearing this in mind, let us briefly consider with which perceptual contents activity in these regions has been associated. This will at least illustrate what the contents of representations at higher levels in a hierarchical trajectory estimate could be.

According to Lima et al.'s review, SMA and pre-SMA become specifically activated by non-verbal vocal emotional cues (cf. Lima et al. 2016, Box 2 on p. 532, and the evidence cited therein). It is plausible that such contents are part of what determines the perceptual character of conscious music perception, and at the same time part of what makes such conscious experiences difficult to describe. In general, the way in which these areas contribute to auditory perception is complex, as Lima et al. point out:

> There is no consensus position on the roles of SMA and pre-SMA responses in auditory processing and imagery. When such responses are discussed, they have been linked to a variety of processes. Timing functions have been suggested for perceptual tasks requiring evaluations of temporal aspects of auditory stimuli […], or for stimuli varying in the sequential predictability and rhythmic regularity that they afford […]. SMA and pre-SMA, together with the cerebellum and the basal ganglia, have in fact been considered to form the substrates for a 'temporal processing' network […]. (Lima et al. 2016, p. 535)

Rhythmic regularities are among the features which are especially relevant in this context, because they change more slowly than such features as loudness or pitch. But the general picture is even more complex. In one study cited by Lima et al. (Raij and Riekki 2012), activity in pre-SMA was stronger for voluntarily generated imagery than for auditory hallucinations, suggesting a role in coding voluntary imagery (Lima et al. 2016, p. 532). Futhermore, activity in SMA seems to be correlated with perceived vividness of auditory imagery (p. 534).

Such findings are consistent with the claim that the perceptual contents of the specious present involve more than just successions of events. Instead, individual events (such as the sounding of a single note) are experienced in the context of larger temporal wholes, which may be marked by an affective character, a rhythmic regularity, volitional aspects (like an "urge to move", cf. Lima et al. 2016, p. 537; see also Grahn and McAuley 2009), or the experienced vividness of imagery.

We can distinguish between two types of hierarchy of temporal wholes here, *nested* and *non-nested*. The elements of a nested hierarchy stand to each other in part-whole relations (just as brushing the first tooth may be part of a larger tooth-brushing event, which has a longer temporal duration). The elements of a non-nested temporal hierarchy are only hierarchically ordered by the relation "has a longer duration than". For instance, the emotional response accompanying hearing a short melody could have a longer temporal extension than the melody itself, but it is not experienced as part of the melody (and neither is the melody experienced as part of the emotional response). A functional difference between these two types of hierarchy might be that one could selectively attend to the elements of a non-nested temporal hierarchy (only to the melody, or only to the emotional response), yet not always be able to do so for a nested temporal hierarchy (e.g., it may be impossible to attend only to the rhythm of a melody, without thereby also attending to the sounds of which the melody is composed).

## 7 To What Extent Are Mediating Representations Predictive of Perceptual Contents?

So far, I have only suggested that regularities tracked at different temporal (and spatial) grains may determine the contents of our conscious perception of temporal processes and successions of events. This idea sits well with hierarchical PP models, and I gave examples in the previous section, but I have not yet addressed the question as to whether features tracked at different levels of the hierarchy

can plausibly be assumed to be predictive of each other. Despite the differences between TEM (and HiTEM) and EST to which I alluded, we here encounter an interesting parallel: EST entails that predictions are derived from event models, and when there is an increase in prediction error an event boundary is inferred, and the event model is updated (cf. Reynolds et al. 2007, p. 616). The fact that a single event model can be predictive of a stream of perceptual input is exploited here, and this idea can of course be generalized to hierarchical models (cf. Butz 2016).

Applying this to conscious auditory perception of melodies, can we identify predictive relationships between the contents mentioned in the previous section? More specifically, to what extent are mediating representations (neither purely perceptual nor purely conceptual) predictive of perceptual representations? First of all, representations of rhythm or meter are predictive of the *timing* of individual notes. Furthermore, emotional responses can be predictive of the *key* in which a melody is played, and the key can be predictive of intervals in a melody. An urge to move may be an even higher-level representation, which is not predictive of a particular rhythm but perhaps of a certain class, e.g., rhythms familiar to the subject or with a clearly perceivable meter or beat. So it is at least plausible to assume that the contents experienced in temporal perception (like music perception), are not only ordered (or nested) in a temporal hierarchy but are also predictive of each other. Therefore, it should be possible to model them in the way suggested by the hierarchical predictive processing models mentioned above (section 3).[25]

## 8 Conclusion

This chapter has focused on two features of temporal consciousness, which I called **endurance** and **continuity**:

**Continuity** $=_{Df}$ At least sometimes, we experience smooth successions of events (or smooth changes).

**Endurance** $=_{Df}$ At least sometimes, we experience temporally extended events as enduring.

Rick Grush's trajectory estimation model (TEM), a compelling model of conscious temporal perception, cannot account for these features, but I have tried to show that the model can be extended by drawing on features of hierarchical predictive processing models. Such models posit representations operating at various timescales. As a result, sequences are not just represented as successions of events but as hierarchical wholes. This accounts for **endurance** if the proposal in this chapter is on the right track. A key feature, which I call **fuzzy boundary**, is that events need not be represented as having determinate temporal boundaries. This may account for **continuity**.

Since this extension of Grush's TEM, which I call HiTEM (hierarchical trajectory estimation model), draws on features of existing computational PP models, it is at least theoretically supported. Empirically, more work needs to be done to find direct support for the model, but current evidence on neural underpinnings of auditory perception is at least consistent with HiTEM. In particular, empirical results may also enrich phenomenological descriptions of temporal consciousness: They will allow us to say in more detail what exactly we experience when we consciously perceive temporally extended processes or successions of events.

25 With respect to auditory perception, an excellent overview and a model can be found in (Winkler and Schröger 2015).

## References


Arrowsmith, D. & Place, C. M. (1998[1992]). *Dynamical systems: Differential equations, maps, and chaotic behaviour.* London: Chapman & Hall / CRC Press.

Butz, M. V. (2016). Toward a unified sub-symbolic computational theory of cognition. *Frontiers in Psychology, 7*. https://dx.doi.org/10.3389/fpsyg.2016.00925.

Casati, R. & Varzi, A. (2015). Events. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/win2015/entries/events/.

Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind, 121* (482), 753–771. https://dx.doi.org/10.1093/mind/fzs106.

——— (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36* (3), 181–204. https://dx.doi.org/10.1017/S0140525X12000477.

Dainton, B. (2014). Temporal consciousness. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/spr2014/entries/consciousness-temporal/.

Dennett, D. C. & Kinsbourne, M. (1992). Time and the observer. *Behavioral and Brain Sciences, 15* (2), 183–201.

Grahn, J. A. & McAuley, J. D. (2009). Neural bases of individual differences in beat perception. *Neuroimage, 47* (4), 1894–1903. https://dx.doi.org/10.1016/j.neuroimage.2009.04.039.

Grush, R. (2005). Internal models and the construction of time: Generalizing from *state* estimation to *trajectory* estimation to address temporal features of perception, including temporal illusions. *Journal of Neural Engineering, 2* (3), S209–S218. https://dx.doi.org/10.1088/1741-2560/2/3/S05.

———(2006). How to, and how not to, bridge computational cognitive neuroscience and Husserlian phenomenology of time consciousness. *Synthese, 153* (3), 417–450.

———(2008). Temporal representation and dynamics. *New Ideas in Psychology, 26* (2), 146–157. https://dx.doi.org/10.1016/j.newideapsych.2007.07.017.

———(2016). On the temporal character of temporal experience, its scale non-invariance, and its small scale structure. *Manuscript.* https://dx.doi.org/10.21224/P4WC73.

Haggard, P., Clark, S. & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience, 5* (4), 382–385.

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

Hohwy, J., Paton, B. & Palmer, C. (2016). Distrusting the present. *Phenomenology and the Cognitive Sciences, 15* (3), 315–335. https://dx.doi.org/10.1007/s11097-015-9439-6.

Hubbard, T. L. (2014). Forms of momentum across space: Representational, operational, and attentional. *Psychonomic Bulletin & Review, 21* (6), 1371–1403. https://dx.doi.org/10.3758/s13423-014-0624-3.

Husserl, E. (1991). *On the phenomenology of the consciousness of internal time (1893-1917).* Dordrecht, Boston, London: Kluwer Academic Publishers.

James, W. (1890). *The principles of psychology.* New York: Henry Holt.

Kelly, S. D. (2005). Temporal awareness. In D. W. Smith & A. L. Thomasson (Eds.) *Phenomenology and philosophy of mind* (pp. 222–234). Oxford: Oxford University Press.

Kiebel, S. J., Daunizeau, J., Friston, K. J. & Sporns, O. (2008a). A hierarchy of time-scales and the brain. *PLoS Computational Biology, 4* (11), e1000209. https://dx.doi.org/10.1371/journal.pcbi.1000209.

———(2008b). Supporting information. *PLoS Computational Biology, 4* (11), e1000209. https://dx.doi.org/10.1371/journal.pcbi.1000209.s001.

Kiebel, S. J., von Kriegstein, K., Daunizeau, J. & Friston, K. J. (2009). Recognizing sequences of sequences. *PLoS Comput Biol, 5* (8), e1000464. https://dx.doi.org/10.1371/journal.pcbi.1000464.

Kim, J. (1966). On the psycho-physical identity theory. *American Philosophical Quarterly, 3* (3), 227–235.

Kolers, P. A. & von Grunau, M. (1975). Visual construction of color is digital. *Science, 187* (4178), 757-759. https://dx.doi.org/10.1126/science.1114322.

Lee, G. (2014). Temporal experience and the temporal structure of experience. *Philosopher's Imprint, 14* (3), 1–21. www.philosophersimprint.org/014003/.

Lehmann, D. (2013). Consciousness: Microstates of the brain's electric field as atoms of thought and emotion. In A. Pereira Jr. & D. Lehmann (Eds.) *The unity of mind, brain and world: Current perspectives on a science of consciousness.* (pp. 191–218). Cambridge: Cambridge University Press.

Lima, C. F., Krishnan, S. & Scott, S. K. (2016). Roles of supplementary motor areas in auditory processing and auditory imagery. *Trends in Neurosciences, 39* (8), 527-542. https://dx.doi.org/10.1016/j.tins.2016.06.003.

Limanowski, J. (2017). (Dis-)attending to the body. Action and self-experience in the active inference framework.

In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Metzinger, T. (2004[2003]). *Being no one: The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

———(2017). The problem of mental action. Predictive control without sensory sheets. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Noë, A. (2006). Experience of the world in time. *Analysis, 66* (289), 26–32. https://dx.doi.org/10.1111/j.1467-8284.2006.00584.x.

Prosser, S. (2016). *Experiencing time.* Oxford: Oxford University Press.

Pöppel, E. (1997). A hierarchical model of temporal perception. *Trends in Cognitive Sciences, 1* (2), 56–61. https://dx.doi.org/10.1016/S1364-6613(97)01008-5.

——— (2009). Pre-semantically defined temporal windows for cognitive processing. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364* (1525), 1887–1896. The Royal Society. https://dx.doi.org/10.1098/rstb.2009.0015.

Raij, T. T. & Riekki, T. J. J. (2012). Poor supplementary motor area activation differentiates auditory verbal hallucination from imagining the hallucination. *NeuroImage: Clinical, 1* (1), 75–80. https://dx.doi.org/10.1016/j.nicl.2012.09.007.

Repp, B. H. (2001). Phase correction, phase resetting, and phase shifts after subliminal timing perturbations in sensorimotor synchronization. *Journal of Experimental Psychology: Human Perception and Performance, 27* (3), 600–621.

Reynolds, J. R., Zacks, J. M. & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science, 31* (4), 613–643. https://dx.doi.org/10.1080/15326900701399913.

Shi, Y. Q. & Sun, H. (1999). *Image and video compression for multimedia engineering: Fundamentals, algorithms, and standards.* Boca Raton, FL: CRC Press.

Shimojo, S. (2014). Postdiction: Its implications on visual awareness, hindsight, and sense of agency. *Frontiers in Psychology, 5* (196). https://dx.doi.org/10.3389/fpsyg.2014.00196.

Thornton, I. M. & Hubbard, T. L. (2002). Representational momentum: New findings, new directions. *Visual Cognition, 9* (1-2), 1–7. https://dx.doi.org/10.1080/13506280143000430.

Tversky, B. & Zacks, J. M. (2013). Event perception. In D. Reisberg (Ed.) *Oxford handbook of cognitive psychology* (pp. 83–94). New York: Oxford University Press.

Watzl, S. (2013). Silencing the experience of change. *Philosophical Studies, 165*, 1009–1032. https://dx.doi.org/10.1007/s11098-012-0005-6.

Wiese, W. (2014). Jakob Hohwy: The predictive mind. *Minds and Machines, 24* (2), 233–237. https://dx.doi.org/10.1007/s11023-014-9338-6.

——— (2016a). Action is enabled by systematic misrepresentations. *Erkenntnis*. https://dx.doi.org/10.1007/s10670-016-9867-x. http://rdcu.be/nZs0.

———(2016b). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 1-22. https://dx.doi.org/10.1007/s11097-016-9472-0.

———(in press). *Experienced wholeness. Integrating insights from Gestalt theory, cognitive neuroscience, and predictive processing.* Cambridge, MA: MIT Press.

Wiese, W. & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Winkler, I. & Schröger, E. (2015). Auditory perceptual objects as generative models: Setting the stage for communication by sound. *Brain and Language, 148*, 1–22. https://dx.doi.org/10.1016/j.bandl.2015.05.003.

Yildiz, I. B. & Kiebel, S. J. (2011). A hierarchical neuronal model for generation and online recognition of birdsongs. *PLoS Computational Biology, 7* (12), e1002303. https://dx.doi.org/10.1371/journal.pcbi.1002303.

Zacks, J. M. (2008). Event perception. *Scholarpedia, 3* (10), 3837.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L. & Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience, 4* (6), 651–655.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S. & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin, 133* (2), 273. https://dx.doi.org/10.1037/0033-2909.133.2.273.