The self-evidencing brain[1]

Jakob Hohwy
Cognition & Philosophy Lab
Monash University

Abstract:
An exciting theory in neuroscience is that the brain is an organ for prediction error minimization (PEM). This theory is rapidly gaining influence and is set to dominate the science of mind and brain in the years to come. PEM has extreme explanatory ambition, and profound philosophical implications. Here, I assume the theory, briefly explain it, and then I argue that PEM implies that the brain is essentially *self-evidencing*. This means it is imperative to identify an evidentiary boundary between the brain and its environment. This boundary defines the mind-world relation, opens the door to skepticism, and makes the mind transpire as more inferentially secluded and neurocentrically skull-bound than many would nowadays think. Therefore, PEM somewhat deflates contemporary hypotheses that cognition is extended, embodied and enactive; however, it can nevertheless accommodate the kinds of cases that fuel these hypotheses.

1. Introduction

How does a system such as the brain manage to use its sensory input to represent the states of affairs in the world? This question is at the heart of central philosophical endeavors. It is also central to neuroscience, and non-biological versions of the question are discussed in artificial intelligence and machine learning. In these disciplines an answer (with philosophical progenitors) is rapidly gaining momentum: the brain minimizes its prediction error and thereby infers the states of the world (Friston 2010 and references therein).

When properly unpacked, this answer has profound philosophical implications for understanding the nature of perception, action and attention (Clark 2013, Hohwy 2013). Here, the prediction error minimization (PEM) theory is explored further in an attempt to understand the mind-world relation. PEM should make us resist conceptions of this relation on which the mind is in some fundamental way open or porous to the world, or on which it is in some strong sense embodied, extended or enactive. Instead, PEM reveals the mind to be inferentially secluded from the world, it seems to be more neurocentrically skull-bound than embodied or extended, and action itself is more an inferential process on sensory input than an enactive coupling with the body and environment. I base the claims on the idea that PEM implies that the brain is *self-evidencing*, where this enforces an evidentiary boundary between it and the external causes of sensory input harbored in the environment and in the rest of the body.

Once we adopt PEM, it is thus difficult to maintain some of the stronger formulations of embodied, extended, enactive cognition, which have come to dominate recent research in cognitive science. Instead a more old-fashioned, skepticism-prone view of the mind-world relation imposes itself.

The paper begins with a brief explanation of PEM and how this simple idea about the brain extends to account for everything the mind is and does, and leads to the idea of the self-evidencing brain. The next sections demonstrate how seclusion and neurocentricity follows, and explain those notions in more detail. The following sections then go on to consider how key examples from the embodied/extended/enactive cognition literature can be dealt with.

2. Prediction error minimization

PEM says that prediction error minimization is the only principle for the activity of the brain. This means that the brain is an organ that on average and over time continually minimizes the error between the sensory input it predicts on the basis of its model of the world and the actual sensory input. (Since the sum of prediction error over time is also known as free-energy, PEM is also known as the free-energy principle; Friston and Stephan 2007).

This is a very ambitious theory. If this is all the brain does, then perception, action, attention, and all other mental processes, must come down to prediction error minimization. This paper is not concerned with arguing for the truth of PEM, instead it will assume it and then explore its consequences. Suffice to say that there is mounting evidence from computational, neuroimaging, and psychophysical studies that PEM is true and has this explanatory reach (for reviews, see Bubic, Von Cramon et al. 2010, Huang and Rao 2011, Den Ouden, Kok et al. 2012). Mathematically and conceptually, the theory also has immense unifying power, it can subsume information theoretical, reinforcement learning, filtering approaches, and many more theoretical approaches (Friston 2010); likewise, it can explain many puzzling aspects of the phenomenology of perception and action (Hohwy 2013).

There are various ways in which PEM can be introduced and explained. For the purposes of the argument in this paper, a brief, statistical heuristic is used, which captures its application to perception, action, and attention (much more detail is given in the texts cited above).

In statistical inference, samples are assumed to be generated by a hidden, underlying cause, and a statistical model is then fitted. For example, the mean of the samples could be chosen as the statistical model. It is easy to calculate the error of such a model: square and sum all the individual errors between the sample data points and the corresponding point on the model (e.g., where error is the distance from the mean to each corresponding data point). This total error should be small. If not then it will be difficult to use the model to predict without much error where the next sample data point will be. For example, if there is much error between the mean-model and the samples, then the model will generate a lot of prediction error—it will be rather useless for knowing what happens next. In contrast, a better fitting model (a higher-order polynomial) will

have a better chance at predicting the next data point. This can be turned around such that the aim is to have models that fit the samples so well that we can minimize the error in the predictions on the basis of the model—minimize the prediction error.

Intuitively, if a model has little prediction error, then it will carry more information about the states of affairs in the world causing the samples. The parameters of the model will come to predict the states of the world, and the states of the world will predict the parameters of the model. In contrast, a model whose predictions are as likely as not to be successful carries no information about the world. The success of a model's representation of the world is then scalable between chance and perfect prediction, as filtered through the sample.

This description of statistical inference appeals to the idea of the magnitude of (prediction) error. This is a context-dependent notion since what is a large error in one context may not be so in another. The size of the error should be evaluated relative to the expected levels of noise and uncertainty in the manner the samples are caused and measured. If the signal received through the sample is very noisy, then a large error is not necessarily a symptom of a bad model – indeed it would be bad to fit a model closely to relatively noisy samples (*cf*. overfitting). If the noise and uncertainty is low, the signal can be trusted, and then the same, large amount of error would be detrimental to keeping prediction error low. It is therefore essential to statistically assess the error itself. In other words, there should be statistical models of noise and uncertainty too, so that the levels of error can be factored into statistical inference in a reasonable manner. This reflects the need to measure and evaluate the variance in the distributions in ordinary statistical inference. The inverse of variance is precision, so in statistical inference one needs to build up expectations for the precisions of the sample data. This seemingly technical point will be important in what follows.

Statistical inference is the tool of science, that is, scientists are in the business of minimizing the error in predictions generated by their hypotheses. But scientists can do various things in response to prediction error. They can adapt the parameters of the model, to make it better reflect the samples drawn from the hidden cause. Or, they can change the samples they draw from the hidden cause, for example, by controlling better for confounds, or by intervening in other ways to check for causal relationships. Prediction error can thus be minimized under two directions of fit.

As it is for statistical inference, so it is for the brain in perception, attention, and action. The brain *perceives* by minimizing prediction error between its hypotheses about the world through updating the parameters of those hypotheses. The brain *attends* by minimizing prediction error efficiently through reliance on prediction errors that are expected to be precise. The brain *acts* by changing its sensory input so it is less erroneous (more accurate) with respect to its hypotheses about the world.

Mechanistically, prediction error minimization happens by signaling within cortical layers and message passing between levels of the cortical hierarchy

(Kiebel, Daunizeau et al. 2008, Kiebel, Daunizeau et al. 2010, Bastos, Usrey et al. 2012). Each level of the cortical hierarchy tends to treat sensory input reflecting regularities operating at different time scales: fast time scales low in the sensory stream, and slower time scales higher in the hierarchy, towards the front of the brain. Each level is only concerned with attenuating as best possible the input (*i.e.*, explaining away the prediction error) from the level below and passing any unpredicted parts of this input up to the next level for it to explain away. Bottom-up sensory input is then conceived as nothing but prediction error—feedback for the brain's top-down hypothesizing.

Computationally, perception can then be described as empirical Bayesian inference, where priors are shaped through experience, development and evolution, and harnessed in the parameters of hierarchical statistical models of the causes of the sensory input. The best models are those with the best predictions passed down to lower levels, they have the highest posterior probability and thus come to dominate perceptual inference. Error is minimized through some minimization scheme such as gradient descent, expectation maximization, or variational Bayes (for informal introductions, see Clark 2013, Hohwy 2013).

Attention occurs through precision weighting of the prediction errors, both in terms of their intrinsic signal-noise ratio as well as top-down modulated expected precisions (Feldman and Friston 2010, Hohwy 2012). Mechanistically, this happens through weighted gain control. This ensures resources are allocated to signals that will facilitate long-term prediction error minimization.

Action arises not through computation of motor commands but instead by the sensorimotor system passing predictions of proprioceptive input to the classic reflex arcs, which fulfill them and thereby cause action; this is known as *active inference* (Friston, Adams et al. 2012, Friston, Samothrakis et al. 2012).

Attention and action will be discussed in more detail below: their inferential nature contributes strongly to the secluded and neurocentric nature of the mind.

In an ideal but impossible design, perception, attention and action would require the brain to simultaneously access both the internal estimates and the true states of affairs in the world. This would allow it to compare the representation and the represented, the attended and what is worth attending to, action planning and what is acted upon. Philosophers have long recognized that there is no such access since we never have unfettered knowledge of states of affairs in the world. PEM delivers the tools for circumventing this problem: it gives the brain access to two things it can compare, namely the predicted and the actual input. Moreover, the divergence between these two is harnessed in a bound that can be minimized in inference, in a way guaranteed (modulo malfunction and skepticism) to approximate the true state of affairs.[2]

As mentioned, there is mounting empirical and modeling evidence in favor of PEM; it is also attractive because of its ability to unify very diverse approaches to cognition. Philosophically speaking it is attractive, as I mentioned, because it

provides a clear way to circumvent the problem that we represent, attend to, and act on the world, yet have no unfettered, independent access that would guide these processes. No other account of cognition comes close to deliver as compellingly and comprehensively as PEM. This alone is enough to take PEM seriously.

According to PEM, then, perception, attention, and action are nothing but aspects of statistical inference. This gives an austere picture of the mind as purely a statistical inference machine. The allowable questions about the mind should then all boil down to questions of the type: what is inferring what? What is the relation between the inference and the inferred? What kind of inference is used? What is the quality of the inference? How can mere inference explain this or that mental characteristic? Why is there inference?

In the most general terms, PEM leads us to expect that just as there is a schism between a statistical model and the modeled cause in statistical inference, there is a schism between the prediction-generating models of the brain and the modeled states of affairs in the world. Views of mind and cognition that emphasize openness, extension, embodiment, and enactive coupling into the environment seem to ignore this inferential conception of the mind. The next section describes the schism in more systematic terms.

3. Self-evidencing

PEM is essentially inference to the best explanation, cast in (empirical, variational) Bayesian terms. The winning hypothesis about the world is the one with the highest posterior probability, that is, the hypothesis that best explains away the sensory input, in a context-dependent fashion, under expectations of precision, and with long run, average fit taken into account. This maps on quite well to the best-makers of inference to the best explanation: the best explanation actually explains away the evidence, it integrates with beliefs about other pieces of evidence as well as considerations about the wider context, it explains in detail, and is fecund and unificatory in the longer run.[3]

In inference to the best explanation, an interesting situation may arise. A given hypothesis, h_i, best explains away the occurrence of some evidence, e_i, and by doing so it provides evidence for itself. That is, e_i becomes evidence for h_i to the extent h_i explains e_i. In these cases, the fact that the evidence occurs is an indispensable part of the evidential basis for h_i. In his seminal work on scientific explanation, Carl Hempel (1965: 372–4) thus described h_i as a *self-evidencing* explanation when the "information or assumption that [e_i occurs] forms an indispensable part of the only available evidential support for [h_i]."[4]

When h_i is self-evidencing, there is an explanatory-evidentiary circle (EE-circle) where h_i explains e_i and e_i in turn is evidence for h_i. This may look suspiciously circular but is an acceptable, common epistemic pattern. Adapting one of Lipton's examples (2004: 24), say you look out the window one morning, see a footprint in the snow and consider various hypotheses about what might explain the occurrence of this surprising evidence (notice that here the footprint works as prediction error, something not predicted by your current model). It turns out

the best explanation of the footprint evidence is that a burglar is afoot, so the probability of the burglar hypothesis will then increase, and you may then come to believe this hypothesis too. Simultaneously, if someone were to ask you what evidence you have for the burglar hypothesis you would be justified in quoting the occurrence of the footprint. As Hempel says, "an acceptable self-evidencing explanation benefits, as it were, by the wisdom of hindsight derived from the information that the explanandum event [i.e. the footprint] has occurred, but does not misuse that information so as to produce a circular explanation" (Hempel 1965: 373).[5]

Hempel and Lipton however both note that there is one situation where the EE-circle turns vicious. If someone raises novel doubts about the occurrence of e_i, then the higher posterior of h_i, acquired through its explanatory prowess, cannot be used to allay this doubt. If someone suggests that the footprint evidence was really produced by hoaxers, tricking you into believing burglars are afoot, then you cannot dismiss them by appeal to your newly formed burglar belief, even though this belief is now well justified. Hempel stresses that the grounds for believing that e_i occurred must be established independently of the information or assumption about the occurrence of e_i, and h_i is not independent in this sense. This is not a problem, if independent evidence can be procured against the hoax-hypothesis. But if no such evidence is to hand there is nothing that can be done to address this kind of doubt. This illustrates that an EE-circle creates an *evidentiary boundary* between e_i and h_i on the one side, and the hidden causes of e_i beyond the circle on the other side. It is *evidentiary* because it is defined by the occurrence of the evidence, and it is a *boundary* because causes beyond it must be inferred—they can only be represented vicariously, under the threat of ignorance just described.

The EE-circle characterizes Bayesian inference in the shape of PEM too. The internal model that generates hypotheses that over time makes the evidence most likely, and does so most precisely and simply, will have its own evidence maximized. That is, as a model generates hypotheses that explain away occurring surprising evidence (i.e., minimize prediction error) it maximizes the evidence for itself. Prediction error minimization thus constitutes self-evidencing. This is then the doctrine of the *self-evidencing brain.* The brain is an organ that approximates optimal Bayesian inference, through prediction error minimization. It does this by being an active model of the world, an agent, which continually either updates its expectations or changes its sensory input to fit the expectations. The sensory input it encounters determines the parameters of the model and the shape of the agent as it is successfully explained away, and thus the occurrence of the sensory input becomes indispensable evidence for the existence of the agent. This is self-evidencing because the model or agent would not exist unless the sensory input it explains away occurred. Accordingly, treating sensory input as surprise, and the minimizing of surprise as minimizing prediction error, Friston can state that "surprise is called the (negative) model evidence. This means that minimizing surprise is the same as maximizing the sensory evidence for an agent's existence, if we regard the agent as a model of its world" (Friston 2010: 2).[6]

Initially, one might worry that self-evidencing presents the brain as a very passive organ, which merely waits for surprising evidence to come its way and then adjusts its parameters to explain it away (gaining evidence for its existence in this manner). This would seem to leave out the notion of active inference touted above. But self-evidencing describes a PEM system in such abstract terms that it encompasses active inference as well as perceptual inference. Active inference is the process of actively sampling the world for evidence that would support one's hypotheses. For example, expecting that a burglar has been about, I might walk over to the window and look out at the snow and acquire the expected footprint evidence for this hypothesis (a similar example can be made for how eye movement is guided by hypotheses). In such a case, information that the sensory input occurs is still an indispensable part of the evidential support for the hypothesis. For the brain, the better it is at actively seeking out evidence it can explain, the more evidence it gets for its own existence.[7]

The notion of the self-evidencing brain allows a non-trivial description of the possibilities and limits of reasonable probabilistic inference under PEM. This puts the focus on the evidentiary boundary and the way it forces a clear distinction between internal states, where the prediction error minimization occurs, and hidden causes on the other side of the boundary, which must be inferred. It is thus a requirement on any PEM-based account that it allows, and is in principle able to describe, the boundary relative to which prediction error is being minimized, from behind which the mind tries to infer the hidden causes on the other side. Failing that, there will be no evidence for the existence of the agent in question.

4. Inferential seclusion and scepticism

As we saw, PEM induces an evidentiary boundary. All perceptual and active inference happens in an interplay between the evidence to the system, that is, activity at the sensory epithelia, and the predictions generated under the overall model in the brain. This creates a sensory blanket—the evidentiary boundary—that is permeable only in the sense that inferences can be made about the causes of sensory input hidden beyond the boundary. The brain doing the inference is secluded at least in the sense that certain kinds of doubt about the occurrence of the evidence are unanswerable without further, independent evidence. Of course, once we average over the entire sensory input, there is no possibility of independent evidence, which would require us to crawl outside of our own brains.

This becomes an affirmation of simple Cartesian skepticism. Since we cannot obtain an independent view of our position in the world, we cannot exclude the skeptical hypothesis that the sensory input we receive is caused by an evil, hoaxing scientist rather than the external states of affairs we normally believe in. The Bayesian framework thus entails skepticism. Consequently, rejecting skepticism entails rejecting Bayesian inference, and hence PEM (a point acknowledged also by others who adopt Bayesian or information theoretical schemes, e.g., Eliasmith 2000, Usher 2001, Grush 2003).

This general kind of point is familiar. For example, classic Marr-like feature detecting computation happens on sensory input, which means it is a process that occurs all in the brain. Given computation determines perception and cognition, perception and cognition happen in the brain. The mind can then be understood in internalist, solipsistic terms, throwing away the body, the world and other people; as Shapiro describes this kind of position in his review, "give [brains] their input, and the rest of the world makes no difference to them" (Shapiro 2011: 26). Similarly, the Cartesian element is captured by Grush who convincingly defends the idea that "the mind is essentially a thinking thing (where thinking = having and manipulating representations), and the mind is autonomous, in that it is not essentially bound up with the body or environment" (Grush 2003: 54).

Of course, few take the possibility of Cartesian evil scientists as a serious, real threat. But scepticism is often viewed as a symptom of a misguided, unproductive and impoverished view of mind and cognition. Accordingly, orthodox, internalist accounts have come under heavy attack. They are viewed as examples of the "traditional view of content [that] generates skeptical worries about how knowledge of the world is possible: why couldn't our beliefs about the external world be rampant delusions produced by mad scientists manipulating our brains?" (Hurley 1998: 8). Opposed to this orthodox, skepticism-inducing view we do not find epistemological views that are directly aimed at putting knowledge on a safe footing. Instead, we find versions of the extended, embodied, and enactive view of mind and cognition on which "[t]he self does not lurk hidden somewhere between perceptual input and behavioral output, but reappears out in the open, embodied and embedded in the world" (Hurley 1998: 3).[8] When taken literally, such views abolish the evidentiary boundary and therefore they simply cannot lead to skepticism. Scepticism becomes the canary in the coal mine, indicating whether an account of mind and cognition is orthodox and internalist, or not.

Given the epistemic description of PEM as self-evidencing it is not surprising that PEM is seen as a relic of the bad old times of internalist representationalism. This is captured vividly in criticisms of Andy Clark's recent treatment and defence of PEM: "Clark's adoption of a thoroughgoing inferential model of perception…threaten to return us to the bad old days of epistemic internalism that the field…rightly left behind"; "[w]estern philosophy has been down this lonely and unproductive road many times" (Anderson and Chemero 2013). Contrary to this sentiment, I think it is in fact the road best taken, and I believe it will be a productive one at that. However, it takes some care to manage this debate, and guide it away from simple dichotomies.

The mind is secluded in the inferential sense described by the evidentiary boundary. Of course, it is not total, hermetic insulation, which would completely disable PEM as an account of perception, attention and action. If the hypotheses generated in our brains were completely divorced from the world then there would be no prediction error minimization and thus no meaningful perception, attention or action. Indeed, a major attraction of PEM is that it allows that our minds are *open* to and in tune with the world at multiple, interlocked time-scales

of representation, in both perception and action and as modulated throughout by attention. A system that minimizes prediction error is constantly tuning its hypotheses in response to the prediction error, where the prediction error essentially is a feedback signal received from the world in response to the brain's hypothesis-testing efforts. A PEM system is thus supervised directly by the truth.

The challenge is then to balance seclusion and openness in our understanding of the mind-world relation. If seclusion is weighted too much, then the very notion of prediction error is deflated. If openness is weighted too much then the inferential core of prediction error minimization is lost.

Clark, in his take on PEM, weighs openness more than seclusion. He builds up to the claim that his "account is, as intended, walking a delicate but important line" between seclusion and openness (Clark 2013: 242). In response to my earlier statements concerning seclusion (in my 2007), Clark acknowledges that "there is something right about…the challenging vision…that our expectations are in some important sense the primary source of all the contents of our perceptions, even though such contents are constantly being checked, nuanced, and selected by the prediction error signals consequent upon the driving sensory input" (2013: 199). But he insists that "it remains correct to say that what we perceive is not some internal representation or hypothesis but (precisely) the world…We do so courtesy of the brain's ability to latch on to how the world is by means of a complex flow of sub-personal processes" and adds that "[i]t is by [inferential] means that biological beings are able to establish a truly tight mind-world linkage", and thus "[t]he intervening [subpersonal, inferential] mechanisms…introduce no worrisome barrier between mind and world". (2013: 199; see also Clark 2012; Ward 2012).

Though it will be difficult to adjudicate conclusively a discussion as abstract as this, it seems to me that there are several aspects of PEM that should make us consider it a challenging and perhaps worrisome account of a secluded mind rather than a comforting account of an open mind, porous to the world.

First, given PEM entails skepticism, which severs the mind-world linkage, there is a non-trivial task in establishing the precise sense in which there can be, in Clark's words, a "truly tight mind-world linkage". Adding fast time scales to this picture, such that there is a "constant flow" of statistical processes, does nothing to change the inferential seclusion; nor does adding hierarchical levels. There is no principled difference between a highly hierarchical, spatiotemporally fine-grained system like our brains and a rudimentary system with only few levels, whose model only barely carries information about some environmental cause. Both such systems are self-evidencing and therefore inferentially secluded in exactly the same sense. Adding more of the same will not create a tighter mind-world linkage or make the mind more porous to the world.

Second, any account that ties perceptual content to a statistical model within an evidentiary boundary will wedge apart the statistical model and the hidden causes it models. To illustrate, having access to rain samples and the mean of the rainfall is a very different thing from having access to the actual rainfall, even if

the mean carries information about the rain. An explanation of this difference in the case of perceptual inference cannot soften the characterization of the hidden causes so they come to appear somehow unhidden, since this undermines the EE-circle and thereby PEM. It is therefore difficult to say that what we perceive is in some sense directly or precisely the world. On the other hand, seclusion need not take an untenable, naïve form of indirectness, where some homunculus perceives the internal model. The seclusion stems from the inferential component such that the upshot of the sub-personal processes is a probabilistically favoured statistical model. This model is personal-level in the sense that it determines perceptual content—but it is still a model. In this way, though PEM mandates seclusion, it is not a non-starter that is automatically condemned to a homunculus fallacy.

Third, the division between inner and outer for any evidentiary boundary is strict, with only inference being done by the mind, and being done only on the inside; inference and mind cannot bleed into the world of hidden causes (even though the brain is itself a hidden cause, as I shall explain). Indeed, as we saw, we only have evidence for the existence of the agent insofar as it engages in inference, so softening this boundary undermines the evidence for the existence of the agent. Any kind of tight linkage to the world has to happen in the context of a strict boundary between agent and world.

In sum, PEM seems an unlikely candidate to "provide an antidote to the inward retreat of the mind in modern philosophy" (Hurley 1998: 3). Instead, PEM will indeed take us down the worrying "lonely and unproductive road" with its challenges to the truly tight mind-world linkage. I will now substantiate this claim by using PEM to consider a series of contemporary trends in cognitive science and philosophy of mind, which are thought to help us avoid this road. Section 5 discusses extended cognition, Section 6 embodied cognition, and Section 7 looks at some enactive aspects of cognition. I will argue that PEM can accommodate the cases fuelling these trends rather well while staying firmly internalist and inferentialist.

5. Extended cognition

The location of the evidentiary boundary determines the relata of the mind-world relation: what is on the 'mind'-side and what is on the 'world'-side. This immediately speaks to the wide-ranging debate about extended cognition: the boundary should determine what is part of the representing mind and what is part of the represented world.

PEM as functionalism. To find out how widely cognition is extended, the first question to ask is what *could* be part of the EE-circle determined by a prediction error minimizing system. The answer is that many things, including non-neural ones, could be part of such a system. PEM is at heart a functionalist theory: on sensory input, and given a prior state, the system settles into another state, and produces a certain kind of output (we can add active states to the system in order to accommodate active inference). These functional state transitions are described in terms of hierarchical message passing for predictions, prediction error, and precision optimization. In principle, anything could realize such a

system—it doesn't have to be a brain, and it could in principle be performed by a brain in combination with other equipment (and, appropriately, the key formal concepts of PEM arose in machine learning and statistical physics). There is nothing novel in this claim; multiple realizability is a fundamental facet of functionalism.[9]

What does matter to PEM is that an evidentiary boundary is defined, behind which hidden causes lurk. If a given object is on the inside of the evidentiary boundary, and that object plays a role realizing the PEM role, it will be a part-realizer of the mental state in question. If the object is outside the boundary then it is not a realizer of the mental state. Instead, it is represented by the inner states to the extent the system can infer its properties.

Causes beyond the evidentiary boundary are modelled. With this we can go to the question about what is *actually* part of the EE-circle determined by our prediction error minimizing systems?

The answer to this question is that it is brains and sensory states (*i.e.,* the states of the sensory organs) that form the EE-circle. There is no good reason to include anything else in the EE-circle. The parts of our own bodies that are not functionally sensory organs are beyond the boundary, so cognitive states are not extended into the body—there is no embodied extension. Likewise, things in the environment are outside the evidentiary boundary, as are other people and their mental states. So the mind is not extended to things around us or to other people.

The initial reason for asserting this answer is that the states of the body, the states of things in the environment, of other people and their mental states all tend to interact with each other and thereby produce non-linear effects on our senses. As such they are just the kinds of states that should be modeled in internal, hierarchical models of a PEM system. An agent can grasp and use her phone only because she has a more or less precise and accurate internal representation of the phone, the things in her drawer that may occlude it, and the causal interactions between her fingers, eyes and ears, voice and the states of the phone. Only on the basis of such a model can she convolve the sensory input from these hypothesized causes and thereby be in a position to predict how it will hit her senses. In so far as we can interact with things in the environment, including our own and other's bodies, we must be modeling them, forming hypotheses about them and their interactions, predicting the next sensory input, assessing the prediction error generated and updating the hypotheses accordingly.[10] In other words, there is reason to think that these states are all hidden causes, situated beyond the evidentiary boundary.

Against extended cognition. The discussion up to this point puts pressure on claims that objects beyond the brain are part of mental states. It is not clear in what sense PEM could or should allow objects beyond the evidentiary boundary to serve as anything but causes of sensory input and thus of prediction error. I now turn to some explicit arguments for the extended cognition hypothesis. The upshot will be that though there is no knock-down argument against that hypothesis it is not a particularly good fit for PEM.

It is possible to imagine items such as, famously, notebooks or smartphones, which in some way realize part of the functional role associated with a given mental state, such as remembering something. Clark and Chalmers (1998) and Clark (2008) influentially propose that parity of reasoning compels us to say that

> "if, as we confront some task, a part of the world functions as a process which, were it to go on in the head, we would have no hesitation in accepting as part of the cognitive process, then that part of the world is (for that time) part of the cognitive process" (Clark 2008: 76).

Very much has been said about this kind of argument (for a comprehensive discussion, see Adams and Aizawa 2008). Here, the argument is viewed only through the prism of PEM. Three PEM-inspired points militate against the parity-of-reasoning argument for the extended mind.

First, it is far from clear that notebooks and smartphones actually play any part of the functional role set out by PEM. There does not seem to be the right kind of hierarchical message passing between the notebook or phone and the rest of the neural system to implement variational Bayes with respect to hidden causes. 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.

Second, the PEM-believer who defends extended cognition (perhaps such as if Clark 2008 were combined with Clark 2013) is required to define a plausible evidentiary boundary. 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. It is crucial that this minimization happens on average and in the long run because the surprise that is sought minimized is defined in terms of the states the creature tends to occupy in the long run (states which in this way define its phenotype). 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.

Third, a complaint about explanatory virtues. 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. This is not an inconceivable state of affairs but it is something that sets these supposedly external parts of mental states apart from mental states that are harbored wholly within the brain. It requires that we posit two overlapping yet intimately linked EE-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.[11] The opponent of the extended mind may then opt to divide and conquer: allow that there are extended mental states in some weak extra-agentual sense but

maintain the neurocentric existence of the mind, properly speaking (this can be combined with the second point above, namely that the neurocentric mind is the one that best minimizes prediction error in the long run).

Inference, and differences in trust. I have presented three reasons to doubt that the mind is extended. I believe they are difficult to completely rebut while retaining PEM, but they are not conclusive reasons. An important move in the dialectic against the parity-of-reasoning argument is therefore to provide a better, PEM-based alternative than the extended mind hypothesis for accommodating the characteristic role of certain external objects such as notebooks and smart phones. Such an alternative is available, and it echoes the most often heard objection to the extended mind proposal, namely that the agent has *beliefs* about the external object but that the external object is not itself part of a belief (Clark 2008: 80). From the perspective of PEM, the agent's causal interaction with external objects such as notebooks and smart phones are, as I said at the beginning, modeled by the agent's brain. As such, the brain holds beliefs about these external objects (and about itself), and these beliefs exhaust the role of those objects in the mental states of the agent.

The extended mind hypothesis draws some of its inspiration from the observation that those external objects seem to play a *special* functional role, akin for example to part of a memory role (Clark 2008: 79). If the extended mind hypothesis is rejected in favor of the view that agents merely hold beliefs about these objects, then the way in which this role is special should instead be explained in terms of differences in belief.

To discharge this explanatory burden, it is possible to appeal to the expectations with regard to the prediction error generated from the external object. For example, the prediction error may be expected to be more precise than that of many other external objects, which do not play this special role. This would be justified, given the nature of notebooks and smartphones, and our interactions with them. It would help explain why sensory input from these objects is preferentially trusted, and trusted to much the same degree as internal memory is trusted by many people. However, this does not entail that the external object is itself part of the mental state in question. In contrast to the internal memory, it is still a (precisely) inferred hidden cause.

Such preferential trust would play a role in the active inferences involving the external object. Recall, active inference accomplishes prediction error minimization by changing the input to fit with the hypothesis. Here I extrapolate this simple idea to the kind of example typical of the extended mind debate. For an agent, having the intention of going to the museum is to predict sensory input as of being at the museum—that state has been assigned a higher probability than other, actually occurring states. This may sound slightly odd but is consistent with the idea (*i.e.*, the free-energy principle) that organisms act to maintain themselves within certain expected states in the long run. This expectation is conditional on action, that is, on the sensory input ensuing causal interaction between the agent's body and the environment. This allows the agent to expect a flow of sensory input defining a trajectory to the expected state. This

flow specifies our expected proprioceptive, kinesthetic, and other sensory input on the assumption that the predicted action actually unfolds. These expectations are processed as predictions, which are fulfilled by the body plant. (Later I will discuss this mechanism further, and unpack how action is triggered; here it is the notion of expected flows of sensory input that matters).

There are indefinitely many different possible trajectories that could lead to the same goal (here, of ending at the museum). For example, the agent can walk through the park or along the main road to get to the museum. These different possible actions are associated with different flows of expected sensory input, which can be called policies. On the policy on which the input is assumed to be caused by the agent walking through the park, the expected flow of sensory input will include many trees and joggers, whereas on the policy on which the input is caused by the agent walking down the main road, the sensory input will include many cars and shoppers. The task for the agent is to rank policies such that prediction error is minimized most efficiently, in the long run, and taking the wider context in to account.

This means that ranking policies is nothing but extracting another type of regularity from the world, a type of regularity that involves one's own body, and mental states, as causes in the world. These regularities are in principle no different from the regularities one might extract about other things in the world (e.g., that light normally comes pretty much from above, unless something in the context such as artificial lighting suggests otherwise). In other words, policies are prior beliefs, ranked according to their probabilities. The agent might then have assigned higher prior to the policy that takes him or her through the park rather than the policy taking the agent down the main road.

Now the role of special external objects such as notebooks and smart phones can be described. An agent relying on a notebook to get to the museum has high confidence assigned to the policy that his or her actions involve the notebook. That is, there is preferential trust in the notebook input such that the favoured predicted sensory flow includes the sensory input arising from the agent's causal interactions with the notebook. A person with mild Alzheimer's, as in Clark and Chalmers' example, might be fully justified in assigning this policy high confidence; for example, not using the notebook will make it much harder for action to reliably fulfill the predictions of being at the museum, resulting in a prediction error increase and, in the long run, a difficulty with remaining within the expected states—such as literally getting lost. In contrast, a healthy agent who does not rely on the notebook does not favor this policy, and does not then predict the flow of sensory input that involves the notebook. The special functional role of the notebook is then fully accounted for in terms of expected precisions and ranking of policies for active inference.

Spurious complexity? As mentioned, this account of the special role of some external objects basically says agents have beliefs about those objects, where belief is cashed out in terms of policies and precisions of expected sensory flow. Clark and Chalmers object to this type of belief-based objection because it adds "spurious complexity" (Clark 2008: 80): we are not inclined to say of the healthy

agent who does not rely on the notebook that he or she had a *belief* about their own memory, and acts on that belief to extract the memory. Given the automaticity of memory this seems a spuriously complex conception of memory processes. Therefore, given the automaticity with which the other agent relies on a notebook, we should not attribute belief to that agent either.

The chief worry about this response is that PEM's notion of policies imposes a difference between the two agents, in spite of both processes being automatic. The agent who relies only on his or her memory does not treat memory as a cause of sensory input. The memory itself is not modeled as something that contributes to the expected sensory flow. The memory of the location of the museum is a prior that works top-down to reduce uncertainty about the expected sensory flow, and thereby makes it more likely that the agent ends up at the museum (a vague, imprecise memory of the location will be less likely to make the action succeed in producing the expected input). In this sense, the memory shapes the expected sensory input without itself occurring also as a causal parameter in the model. As such, there is no belief about the memory, and thus no spurious complexity. In contrast, the agent who relies on the notebook does predict it to contribute causally to the flow of sensory input and this is what motivates the belief-response.

The observation that reliance on the notebook happens automatically is dealt with by appealing to efficiency gains for trusted prediction error, where such errors are preferentially weighted in perceptual and active inference. The observation that the reliance on memory happens automatically is dealt with, at least in part, by the idea that active inference is unconscious, on a par with unconscious perceptual inference.

From the point of view of PEM there is then no good reason to accept that external objects are parts of mental states. PEM can accommodate some external objects as especially trusted environmental causes of expected sensory flows, it can explain the reliance on memory in different terms, and it does not have to impute spurious complexity to the reliance on memory.

Shrinked cognition? A more thorny issue is whether PEM entails "shrinked" cognition. Prediction error minimization is ordered hierarchically in the cortex. Low levels predict statistical regularities at fast time scales and sends prediction errors to the level above, which processes regularities at slower time scales and sends its error onwards as input to the next level up, and so on up through the hierarchy. In principle, if we focus on the entire hierarchy minus the bottom layer, we would still have a fully formed EE-circle with its own evidentiary boundary, modeling the external world, including the states of the bottom layer. The input to this system then becomes the evidence for a shrinked model, and thus for the existence of a new "agent". This process of nesting agents within each other could continue throughout the cortical hierarchy. The consequence is that each of us contains a proliferation of nested agents.

This may be deemed as unattractive as the outwards proliferation of agents, which I saddled the extended cognition hypothesis with earlier. I think the other

points I raised against extended cognition still makes that hypothesis especially unattractive. However, one response to this is to acquiesce in the multitude of agents and simply apply the contrastive method to establish which agent is the focus of explanation: ask "why did this behavior rather than that behavior occur?" and the agent responsible for the contrast is the agent of interest. The advantage of going this way is that it allows an ecumenical attitude with respect to hypotheses like the extended mind, which are judged then on their overall explanatory prowess.

A disadvantage with the ecumenical attitude, however, is that the explanatory focus will be on different agents each time another EE-circle is picked. For example, we would equivocate on 'agent' if we try to attribute properties to the extended agent, such as Otto with the notebook, and thinking this carries over to the non-extended agent, such as Inga without the notebook. Another, somewhat more principled response to the problem of the shrinking agent is to rank agents according to their overall, long-term prediction error minimization (or free-energy minimization): the agent worthy of explanatory focus is the system that in the long run is best at revisiting a limited (but not too small) set of states. It is most plausible to think that such a minimal entropy system is constituted by the nervous system of what we normally identify as a biological organism: shrinked agents are not able to actively visit enough states, and extended agents do not maintain low entropy in the long run (*cf*. Friston 2013 for an argument along these lines, where EE-circles and self-evidencing are cast in terms of Markov blankets).

6. Embodied cognition

The extended cognition hypothesis focuses on objects beyond the body. A related question is whether cognition is *embodied*. Unfortunately, the notion of embodied cognition is rather more heterogeneous than the notion of the extended mind (for discussion and review, see Gallagher 2005, Kiverstein and Clark 2009, Shapiro 2011, Alsmith and Vignemont 2012, Metzinger 2013). In some respects, it is the important observation that cognitive science will have trouble understanding mind and cognition if it fails to properly recognize the role of the body in shaping and constraining perception and action. In other respects, it is a series of more radical theses to the effect that somehow internal representations, studied in traditional cognitive science, are not needed at all for understanding mind and cognition, as somehow the body and the world furnishes its own representation (Brooks 1991, Noë and O'Regan 2001, Noë 2004, Hutto and Myin 2013).

The various notions of embodied cognition have sparked a very large volume of discussion. Here I again confine the discussion to exploring what PEM would say about the role of the body in unconscious perceptual and active inference. I will also indicate what to say about the claim that internal representation is obviated by the body and the world.

There is no doubt PEM must assign a key role to the body in perception. This follows from the notion of active inference where hypotheses are tested in selective sensory sampling. Such inference depends on causal interactions

between the body and the world. For example, to figure out how the sensory input will change as an object is manipulated it is necessary to factor in the way the hands interact with the object, relative to the agent's first-person perspective. Inference is then conditionalized on a dynamic body representation, harboured in the cortical hierarchy, which treats the body itself as a hidden cause of sensory input. It follows that the way perception unfolds will differ depending on the body's interactions with the world and how these interactions are modelled. Different bodies in different environments will be associated with different expected flows of sensory inputs, which are fulfilled differently in active inference. In other words, perception cannot be understood without factoring in the body and its interactions with the environment. In this sense, embodied cognition is inevitable, according to PEM.

It is crucial to acknowledge that accommodating embodied cognition in this way happens within the strictures of the self-evidencing brain. Active inference procures evidence for the hypotheses generated under the model by putting the brain in a condition where the sensory input matches the expected input—it is a very internal process, wholly in line with the statistical take on prediction error minimization. The role of the body is real and substantial, but only in the sense that the body is represented in the model, as a parameter useful for minimizing prediction error.[12] In this sense, there is no fundamental difference between types of inference that rely on the body and types that don't: they all consist just of inner representations of interacting, hidden causes and generation of expected sensory inference on the basis of these representations. Embodied cognition boils down to the fact that one of these modeled causes is the agent's own body. This approach to embodiment is put nicely by Friston:

> Not only does the agent embody the environment but the environment embodies the agent. This is true in the sense that the physical states of the agent (its internal milieu) are part of the environment. In other words, *the statistical model entailed by each agent includes a model of itself as part of that environment*. This model rests upon prior expectations about how environmental states unfold over time (Friston 2011: 89; my emphasis).

When introducing embodied cognition above, I noticed that it comes in a more radical version, which not only focuses on the role of the body in cognition, but which also says that given the role of the body and the environment in cognition, there is no need for the brain to represent its environment. This version of the embodiment idea is inconsistent with PEM. This is because PEM must necessarily rely on internal representations of hidden causes in the world (including the body itself) in order to predict the sensory input that they give rise to (for discussion, see Clark 2012, Hohwy 2013: Ch. 3). The ground for this point was laid above, when I explained how interacting external causes must be explicitly modeled in the brain in order allow prediction of the non-linear evolution of sensory input from interacting causes.[13]

How neurocentric should PEM be? Proponents of embodied cognition argue in various ways that we should be careful about assuming neurocentrism, the idea that cognition is exclusively to be understood in terms of the central nervous system and the neurons of the brain in particular (see Hurley 1998, Clark 2008:

105, Anderson, Richardson et al. 2012). I have suggested, in contrast, that PEM is a good companion for neurocentrism. It is possible to be a little more specific about where PEM would likely place the mind-world boundary.

First, there is reason to think that proprioception and interoception are just as much subject to PEM as exteroception: states of the body beyond the neural system are treated as hidden causes, which are modeled by the brain. These internal states of muscles in our limbs, of heart rate, states of the digestive system and so on, are all on the inferred side, beyond the evidentiary boundary. These states of the body can therefore not be part of the mind. In principle, they are as hidden to us as very distal causes of sensory input such as the receding galaxies causing evidence of red shift in our scientific equipment (for development of these views and how they relate to emotion, self and pathology, see Hohwy 2011, Seth, Suzuki et al. 2012, Seth 2013, Hohwy 2013: Ch. 12).

Second, there is a quite specific account of what happens in active inference, which puts part of the boundary at the dorsal horn of the spinal cord. This is where descending proprioceptive predictions from the brain are compared with ascending prediction errors from stretch receptors in the muscles such that the error can be suppressed by enslaving movement (Friston, Daunizeau et al. 2010). In other words, action arises when the brain predicts what its bodily state should be and hands these predictions over to the body at the dorsal horn, after which the body as it were goes away and does its own thing until the predictions come true.

This tells us how neurocentric we should be: the mind begins where sensory input is delivered through exteroceptive, proprioceptive, and interoceptive receptors and it ends where proprioceptive predictions are delivered, mainly in the spinal cord.[14]

The important caveat to this account of the sensory boundary is that in principle, there can be a proliferation of nested agents, some larger than this boundary, some smaller. This is what has emerged in the discussion so far. There will be ways of seizing upon this possibility, for various explanatory purposes (see for example Anderson, Richardson et al. 2012, Di Paolo and De Jaegher 2012). Giving explanatory priority to the neurocentric boundary I highlight here will thus require some further argument, for example, in the mentioned terms of minimum entropy and long term average prediction error minimization.

Decoupling the mind. We are now in a position to say a little more about the role of the body in cognition. The body is a system that complements neurocentric cognition because it enables prediction error minimization in active inference. Without the body, the system would only be able to minimize prediction error via passive perceptual inference and complexity reduction. This point is in some respects friendly to the 'second wave' extended/embodied cognition hypothesis, which focuses on complementing rather than extending the mind (Menary 2010, Orlandi 2013). However it counters the spirit of such proposals because it relies on explicit internal representations of body and world, and because it comes with the consequence that all behavior of the system can be understood in terms

of the prediction error minimizing interplay between sensory input and the hypotheses of internal models. So, one might say that the mind and the body are "coupled" in this specific complementation sense of the body plant being what enables active inference. But it does not follow that the systems are "coupled" in the sense that states of the body are referred to directly by variables in the differential equations that define prediction error minimization, even as specified for active inference. (See again Grush (2003) for further discussion of coupling, and a defense of decoupling in terms of what parts of a system can be "plugged" in and out).

7. Enactive and coupling approaches to cognition

There is clearly an element of explanatory pragmatics in the considerations so far. I am advocating decoupling the brain from the body and the environment in an epistemic sense, because PEM comes with an inferential boundary that gives clear epistemic roles to each side. But this should not be paired with the claim that no kind of coupling of brain and body/environment whatsoever can pull any explanatory weight. It very obviously can, and this is described in parts of the literature on *enactive* cognition, which is the view that cognition is essentially action-based in a way that couples or blends mind and world closely together. Indeed, the notion of self-evidencing appears to be the epistemic cousin to the dynamic systems theory notions of self-organization and self-enabling, which are often used to explain enactivism (see Di Paolo and Thompson forthcoming). In the light of the stark mind-world schism implied by self-evidencing, it seems to me the explanatory worth that comes from an insistence on coupling is in laying bare how the self-evidencing brain vicariously enslaves an external body plant to fulfill its predictions of sensory input, given its internal model of the world and itself. Nothing in such an explanation would however suggest that the evidentiary boundary between mind and body/environment has somehow been blurred—and it is this boundary that has been driving the argument so far. Put as a basically friendly challenge, it would be useful if enactivism could show how a notion such as self-enabling, for example, can avoid an epistemic, inferential reading in terms of the self-evidencing that entails an evidentiary boundary and thus decoupling.[15]

The issue of skepticism, discussed at the beginning, can be used to illustrate how PEM approaches enactivism. PEM entails skepticism because it cannot help allay skeptical doubts of the variety where an evil scientist operates a brain-in-a-vat. The enactivists Cosmelli and Thompson (2011) point out that in order to actually operate a brain-in-a-vat it seems (nomologically) necessary that it is coupled with a body in an environment conducive to its autonomous functioning. This nomologically driven perspective deviates from the standard epistemologist's concerns with skepticism, but it is nevertheless compelling when the issue concerns what skepticism tells us about the nature of the mind: our very conception of the brain as self-organized does not easily allow a skeptical scenario. However, even though the evil scientist needs to have amazing machinery and needs to be extremely obsessive in order to maintain a PEM brain-in-a-vat, it still holds that the evil scientist's main objective just is to maintain the activity at the evidentiary boundary, such that the envatted brain can engage in prediction error minimization, and turn the sensations proffered

by the evil scientist into evidence for its existence. In other words, the brain may need to be solidly embedded in the real world, but an evidentiary boundary must exist.[16]

Coupling with action, affordances and sparse representations. PEM presents a picture of the brain as a secluded inference-machine populated with massive, hierarchical representations of the world external to the brain and its sensory organs. Perception and action are harboured in top-down predictions, and bottom-up sensory input to the brain is mere corrective feedback to the machinations of the internal models. It is easy to read into this picture an overly laborious, passive style of inference, which sits poorly with the fluid and active way actual brains exploit the signals from its environment. Using a famous example (Clark 1999), one might suspect that if baseball outfielders rely on such internal inference to catch a ball, they would have to compute the arc, acceleration and distance of the ball much like one would do by following a physics textbook; this would be very difficult and resource consuming computationally. In fact, outfielders seem to rely on a much simpler heuristic, or higher-order invariant, where they continually adjust their movement such that they maintain a linear optical trajectory (or perhaps optical acceleration cancellation) of the ball in their visual field. This obviates the need for resource heavy, fine-grained computation; as Clark puts it:

> Such co-ordination dynamics constitute something of a challenge to traditional ideas about perception and action: they replace the notion of rich internal representations and computations with the notion of less expensive strategies whose task is not first to represent the world and then reason on the basis of the representation, but instead to maintain a kind of adaptively potent equilibrium that couples the agent and the world together (Clark 1999: 346).

This coupling is used by Clark to argue in favour of a less secluded, more open mind-world relation than I have advocated on behalf of PEM:

> Instead of using sensing to get enough information inside, past the visual bottleneck, so as to allow the reasoning system to "throw away the world" and solve the problem wholly internally, they use the sensor as *an open conduit allowing environmental magnitudes to exert a constant influence on behavior*. Sensing is here depicted as the opening of a channel, with successful whole-system behavior emerging when activity in this channel is kept within a certain range. What is created is thus a kind of new, task-specific agent-world circuit (Clark 2008: 16).

The question here is how this kind of case looks from the perspective of PEM? Can PEM accommodate such "quick and dirty" processing, and would that liberate the mind from seclusion?

It is a mistake to think that just because the brain only does inference, it must build up its internal model like it was following a sober physics textbook. As long as prediction error is minimized on average and over the long run, it doesn't matter which model is doing it. For this reason a model that predicts linear optical trajectories is entirely feasible and can easily be preferable to a more cumbersome series of computations. This is particularly so if it is a less complex model, with fewer parameters, since prediction error minimization in the long

run is helped by minimal complexity. Furthermore, models that rely on expected precision in prediction error minimization will be preferable to models that rely on less expected precision, and it is reasonable to think that expecting linear optical trajectories is associated with high precisions over the required time scales for catching an object like a ball. Finally, catching a ball requires active inference, that is, a policy that leads to the desired state of holding the ball. As such it is likely that a high precision policy is one that makes the body be tightly controlled by the prediction errors generated by deviance from the expected linear optical trajectory, giving rise to just the kind of fluid coordination between visual image and movement that allows the ball to be caught.

For these reasons, PEM can accommodate these kinds of cases: processing is still a matter of internal inference on sensory input, which relies on explicitly representing the interaction between causes in the environment, namely the ball and the outfielder, and the consequences of this interaction on sensory input. Prediction error with respect to the expectation of linear optical trajectory can arise through any number of non-linear interactions between the path of the running outfielder, and other environmental factors such as the shape and spin of the ball, and wind and weather, so knowing how to minimize prediction error requires modeling of these factors. In fact, we can see that the processing is not particularly quick and dirty. Once we are told that there is a fairly limited, selective use of the sensory input it may *appear* that the brain is not engaged in very substantial, explicit internal representation, and that instead the representational work is somehow fluidly taken over by the world itself. But this overlooks the fact that the kind of selective, sparse sampling in play here requires heavy, explicit modeling of external causes, including complexity reduction and updating of expected precisions. In other words, the sampling can be simple and efficient in the way highlighted by Clark only because countless aspects of the causal order of the world are already being modeled internally.

Clark remarks that in these types of cases "[s]ensing…acts as a constantly available channel that productively couples agent and environment rather than as a kind of 'veil of transduction' whereby world-originating signals must be converted into a persisting inner model of the external scene (Clark 2008: 15). Here Clark is right insofar as the incoming visual signal drives action but wrong insofar as this driving in fact does rely on a veil of transduction, namely the evidentiary boundary within which there is ample inference, and beyond which lies nothing but inferred causes.

Triggering action with inference. The upshot so far is that the kinds of cases fueling enactive and coupled cognition can be accommodated within PEM, even though PEM operates with notions of inferential seclusion and neurocentric internal representation that are anathema to those hypotheses. In this section I want to demonstrate how PEM pushes even the mental event of *acting* into the realm of the brain's statistical inference. This will help underscore that acting in the world is far from a direct, seamless, coupled engagement with the environment—in fact, PEM's take on acting will amplify "the inward retreat of the mind in modern philosophy" which we saw Hurley admonishing.

The account of acting begins with an explanatory problem for PEM's notion of active inference. In active inference, a prediction of sensorimotor input is generated, conditional on movement. For example, a prediction is generated of proprioceptive input conditional on performance of a certain arm movement. This prediction is sent to classic reflex arcs, which initiate movement until the proprioceptive prediction is fulfilled and the arm is in the predicted position (Friston, Daunizeau et al. 2010, Friston, Samothrakis et al. 2012). This is what action is: minimization of prediction error through changing the body's configuration and position.

The problem is that it seems prediction error minimization could go either way here: instead of making the prediction come true through movement, the hypothesis generating the proprioceptive prediction could just as well be revised so as to fit with the actual proprioceptive input, which prior to movement of course suggests that there is no movement. This is a direct outcome of the opposite directions of fit for perceptual and active inference. Perceptual inference makes the hypothesis fit the input, active inference makes the sensory input fit the hypothesis. In the case of action, there is a hypothesis that one's proprioceptive input is different than it actually is, making it an option just to focus on the *actual* proprioceptive input and revise that hypothesis to fit. As a result, action would not occur.

The solution to this problem has to do with longer-term regularities in maintaining a balance between perceptual and active inference (Brown, Adams et al. 2013). The idea is that whereas relatively short-term prediction error minimization can be obtained by sticking with perceptual inference, in the longer term refraining from acting will tend to increase overall prediction error (or free-energy). In other words, by staying in the same situation for longer periods, and thus continuously sampling the same signal, one should expect signal strength in the sensory input to deteriorate. This claim is based on the presumption that the world is a changing place, such that a sensory signal that is strong now will not remain so in perpetuity. As signal strength decreases, prediction error cost creeps up, making it less attractive to remain in the chosen state. This implies that agents expect fluctuations in the precisions of the current input, specifically that the current input can be trusted less and less as time goes by. This creates an imperative to act and at the same time creates an expectation that current input is losing precision. Loss of expected precision of the actual proprioceptive input is then what triggers action because it means the competition between the actual and the expected proprioceptive input becomes skewed: the system begins to down-weight the current input in its inferences. Since prediction error minimization is facilitated by weighting precise prediction error, the hypothesis that one is moving will not be updated in the light of the actual (imprecise) input and instead it will be made true by the next (precise) input generated by movement.

In short, action is triggered as a result of prediction error minimization in the light of the expected precision of prediction error. Of course, this is nothing but application of the central statistical idea about precisions to the case of active inference. The brain gets better evidence for its existence by being *precisely* self-

evidencing. Acting in the world is therefore nothing to do with a direct, seamless engagement with the world—something that is substantially different from representing the world. To act is just to engage in more statis3tical inference.[17]

This is a testable hypothesis: whenever there is action there should be a down-weighting of actual proprioceptive input. There is evidence in favor of this from studies of self-generated and other-generated sensory input. Famously, for example, whereas others can tickle us, we cannot so easily tickle ourselves. This is consistent with the idea that self-generated input (self-tickle) is attenuated, whereas other-generated input is not. Crucially, this attenuation should be global and not specific to just the sensory effects of the movement itself, and it should occur whenever there is movement (for healthy individuals). Impressive computational work of sensory attenuation has been done, which supports the precision-related solution to the problem (Brown, Adams et al. 2013), and we provide support for this hypothesis in a study that showed that people cannot tickle themselves even when they are in the highly surprising context of experiencing that they are swapped into other people's bodies (Van Doorn, Hohwy et al. 2014).

We have seen so far that perception and action in the self-evidencing brain is best described as secluded and neurocentric inference. We can now see that this extends even to the triggering of action. On the basis of this picture, we can strengthen the inferential seclusion. Various proponents of enactive cognition hopes to couple us more directly to the world by conceiving cognitive processes in terms of action and the body's role in the environment (this is done with a variety of different tools and agendas, see Clark 1997, Hurley 1998, Noë 2004, Thompson 2007). The overall idea is that this will enable us to set aside skepticism and the internalist, secluded picture of our access to, and place in, the world that goes with it. But now we have reduced action to mere inference on statistical patterns in sensory input, and it thus transpires that action cannot easily be used to prevent the skepticism and the seclusion that comes with the self-evidencing brain.

8. Concluding remarks

I have brought the prediction error minimization (PEM) framework to bear on influential contemporary ideas of extended, embodied and enactive cognition. The upshot is that the kinds of examples and cases upon which these ideas rest can be reasonably accommodated by PEM without any of the more radical consequences thought to come with extended and embodied cognition: mental states do not extend into the environment, and the involvement of the body and of action in cognition can be described in wholly neuronal, internal, inferential terms.

This argument began with establishing the notion of the self-evidencing brain, the brain that gains evidence for its existence as it seeks out and minimizes its prediction error. This is essential to PEM and induces an evidentiary boundary with the brain on one side and the worldly and bodily hidden causes on the other side. This boundary immediately makes it very difficult to see how the more radical aspects of extended, embodied and enactive cognition could be true. On

this basis, I then demonstrated in more detail how the self-evidencing brain's boundary does indeed retain all cognition within inner, statistical models of the world. In fact, I strengthened the internalist aspect by showing how even acting itself is nothing but optimization of statistical inference.

Once PEM is adopted the mind transpires as a much more neurocentrically secluded statistical mechanism than many would nowadays believe. Giving up PEM in order to sustain the radical spirit of extended, embodied and enactive cognition is ill-advised. PEM is set to dominate neuroscience, it provides the best theoretical account of perception, and can unify under one framework a wide range of phenomena, including, crucially, the phenomena that sparked the notions of embodied, extended and enactive cognition in the first place.

References

Adams, F. and K. Aizawa (2008). *The Bounds of Cognition*. Oxford, Blackwell.
Alsmith, A. J. T. and F. Vignemont (2012). "Embodying the Mind and Representing the Body." *Review of Philosophy and Psychology* 3(1): 1–13.
Anderson, M. and A. Chemero (2013). "The problem with brain GUTs: Conflation of different senses of 'prediction' threatens metaphysical disaster." *Behavioral & Brain Sciences* 36: 204–205.
Anderson, M. L., M. J. Richardson and A. Chemero (2012). "Eroding the boundaries of cognition: Implications of embodiment." *Topics in Cognitive Science*: 717–730.
Bastos, Andre M., W. M. Usrey, Rick A. Adams, George R. Mangun, P. Fries and Karl J. Friston (2012). "Canonical microcircuits for predictive coding." *Neuron* 76(4): 695–711.
Brooks, R. A. (1991). "Intelligence without representation." *Artificial Intelligence* 47(1–3): 139–159.
Brown, H., R. Adams, I. Parees, M. Edwards and K. Friston (2013). "Active inference, sensory attenuation, and illusions." *Cognitive Processing* 14(4): 411–427.
Bubic, A., D. Y. Von Cramon and R. I. Schubotz (2010). "Prediction, cognition and the brain." *Frontiers in Human Neuroscience* 4. doi: 10.3389/fnhum.2010.00025
Clark, A. (1997). *Being There*. Cambridge, Mass., MIT Press.
——— (1999). "An embodied cognitive science?" *Trends in Cognitive Sciences* 3(9): 345–351.
——— (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press, USA.
——— (2009). "Spreading the Joy? Why the Machinery of Consciousness is (Probably) Still in the Head." *Mind* 118: 963-993.
——— (2012). "Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience." *Mind* 121(483): 753–771.
——— (2013). "Whatever next? Predictive brains, situated agents, and the future of cognitive science." *Behavioral & Brain Sciences* 36(3): 181–204.
Clark, A. and D. Chalmers (1998). "The extended mind." *Analysis* 58(1): 7–19.

Conant, R. C. and W. R. Ashby (1970). "Every good regulator of a system must be a model of that system." *International Journal of Systems Science* 1(2): 89–97.
Den Ouden, H. E., P. Kok and F. P. De Lange (2012). "How prediction errors shape perception, attention and motivation." *Frontiers in Psychology* 3. doi: 10.3389/fpsyg.2012.00548.
Di Paolo, E. and E. Thompson (forthcoming). "The enactive approach." *The Routledge Handbook of Embodied Cognition*. L. Shapiro. Oxford, Routledge.
Di Paolo, E. A. and H. De Jaegher (2012). "The interactive brain hypothesis." *Frontiers in Human Neuroscience* 6. doi:10.3389/fnhum.2012.00163.
Eliasmith, C. (2000). *How Neurons Mean: A Neurocomputational Theory of Representational Content.* Ph.D. dissertation. Washington University in St. Louis.
Feldman, H. and K. Friston (2010). "Attention, uncertainty and free–energy." *Frontiers in Human Neuroscience* 4(215). doi: 10.3389/fnhum.2010.00215.
Friston, K. (2010). "The free–energy principle: a unified brain theory?" *Nature Reviews. Neuroscience* 11(2): 127–138.
——— (2011). "Embodied Inference: or 'I think therefore I am, if I am what I think'". In *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): 1–12
Friston, K., R. Adams, L. Perrinet and M. Breakspear (2012). "Perceptions as hypotheses: saccades as experiments." *Frontiers in Psychology* 3. doi: 10.3389/fpsyg.2012.00151.
Friston, K., J. Daunizeau, J. Kilner and S. Kiebel (2010). "Action and behavior: a free–energy formulation." *Biological Cybernetics* 102(3): 227–260.
Friston, K., S. Samothrakis and R. Montague (2012). "Active inference and agency: optimal control without cost functions." *Biological Cybernetics* 106(8): 523–541.
Friston, K. and K. Stephan (2007). "Free energy and the brain." *Synthese* 159(3): 417–458.
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.
Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York, Free Press.
Hohwy, J. (2007). "Functional integration and the mind." *Synthese* 159(3): 315–328.
——— (2011). "Phenomenal variability and introspective reliability." *Mind & Language* 26(3): 261–286.
——— (2012). "Attention and conscious perception in the hypothesis testing brain." *Frontiers in Psychology* 3. doi: 10.3389/fpsyg.2012.00096.
——— (2013). *The Predictive Mind*. Oxford, Oxford University Press.
Huang, Y. and R. P. N. Rao (2011). "Predictive coding." *Wiley Interdisciplinary Reviews: Cognitive Science* 2(5): 580–593.
Hurley, S. L. (1998). *Consciousness in Action*. Harvard: Harvard University Press.
Hutto, D. and E. Myin (2013). *Radicalizing enactivism: Basic minds without content*. Cambridge, Mass., MIT Press.

Kiebel, S. J., J. Daunizeau and K. J. Friston (2008). "A Hierarchy of Time–Scales and the Brain." *PLoS Computational Biology* 4(11): e1000209.
Kiebel, S. J., J. Daunizeau and K. J. Friston (2010). "Perception and hierarchical dynamics." *Frontiers in Neuroinformatics* 4: 12. doi: 10.3389/neuro.11.020.2009.
Kiverstein, J. and A. Clark (2009). "Introduction: Mind embodied, embedded, enacted: One church or many?" *Topoi* 28(1): 1–7.
Lipton, P. (2004). *Inference to the Best Explanation*. London, Routledge.
McDowell, J. H. (1994). *Mind and World*. Harvard, Harvard University Press.
Menary, R. (2010). "Cognitive integration and the extended mind." In *The extended mind*. R. Menary. Cambridge, Mass., MIT Press: 227–244.
Metzinger, T. (2013). "First-order embodiment, second-order embodiment, third-order embodiment: From spatiotemporal self-location to minimal phenomenal selfhood." *The Routledge Handbook of Embodied Cognition*. L. Shapiro. London, Routledge.
Noë, A. (2004). *Action in Perception*. Cambridge, Mass., MIT Press.
——— (2006). "Experience without the head." In *Perceptual Experience*, T. Gendler and J. Hawthorne, New York, Oxford University Press, pp.411-434Noë, A. and K. O'Regan (2001). "A sensorimotor account of vision and visual consciousness." *Behavioral and Brain Sciences* 24: 939–973.
Orlandi, N. (2013). "Embedded Seeing: Vision in the Natural World." *Noûs* 47(4): 727–747.
Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Fransisco, Morgan Kaufmann Publishers.
Putnam, H. (1981). *Reason, Truth and History*, Cambridge University Press.
Searle, J. R. (1980). "Minds, brains, and programs." *Behavioral and Brain Sciences* 3(03): 417–424.
Seth, A. K., K. Suzuki and H. D. Critchley (2012). "An interoceptive predictive coding model of conscious presence." *Frontiers in Psychology* 2. doi: 10.3389/fpsyg.2011.00395.
Seth, A. K. (2013). "Interoceptive inference, emotion, and the embodied self." *Trends in Cognitive Sciences* 17(11): 565-573
Seth, A. K. (2014). "A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia." *Cognitive Neuroscience*: 1-22.
Shapiro, L. (2011). *Embodied Cognition*. Oxford, Routledge.
——— (2013). "Dynamics and cognition." *Minds and Machines* 23(3): 353–375.
Thompson, E. (2007). *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Harvard, Harvard University Press.
Thompson, E. and D. Cosmelli (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.
Van Doorn, G., J. Hohwy and M. Symmons (2014). "Can you tickle yourself if you swap bodies with someone else?" *Consciousness and Cognition* 23(0): 1–11.
Ward, D. 2011: "Enjoying the spread: Conscious externalism reconsidered." *Mind* 121: 731-752.

[1] I wish to thank seminar participants at the Berlin School of Mind and Brain, Freie Universität Berlin, University of Edinburgh's *Eidyn* Centre , University of Manchester, Johannes Gutenberg University Mainz, Aarhus University's Interacting Minds Centre, and University of Copenhagen's Centre for Subjectivity Research. I am also grateful to Bryan Paton, Thor Grünbaum, John Michael, Adrian Alsmith, Wanja Wiese, Jennifer Windt, Thomas Metzinger, Andy Clark and Karl Friston for comments and suggestions on the ideas presented in this paper.

[2] Formally speaking, the sum of prediction error (which under some simplifying assumptions is the free-energy) can be seen to be equal to surprisal or self-information of the sensory states plus a divergence between the selected hypothesis about the causes and the true posterior probability. This is a Kullback-Leibler divergence, which is always zero or positive and this means that by minimizing the divergence, the prediction error or free-energy must be minimized and the surprisal approximated. This means that a creature who needs to avoid surprising input can be helped in doing so by maintaining a tight divergence (or prediction error bound) on surprise—importantly the creature can do this just by knowing internal models and sensory input, and thus without already, impossibly, knowing surprise. On the basis of such inference, action then moves the agent around to increase the accuracy of models in selective sampling; this is based on a different reorganization of the free-energy principle on which prediction error is expressed as complexity plus accuracy (Friston 2010).

[3] Lipton's classic (2004: Ch. 4) work on inference to the best explanation acknowledges the fit with Bayes. As suggested in the main text, I think the fit is even better than he describes. In addition, simplicity is often taken to be a best-maker, that is, the less complex a hypothesis the better it is. Even though there is much discussion in philosophy of science about what simplicity is, there is a straightforward way to accommodate it within a Bayesian scheme. One can appeal to the Akaike Information Criterion, which penalizes hypotheses according to their respective number of parameters; or one can appeal to the Bayes factor, which uses the (Kullback-Leibler) divergence between priors and the selected hypothesis as a measure of complexity. Minimizing complexity is thus a way of minimizing prediction error or free-energy.

[4] Hempel worked with deductive-nomological patterns of explanation rather than inference to the best explanation, however the self-evidencing phenomenon is very similar for both explanation forms; see also Lipton (2004). Hempel also notices that all explanations are self-evidencing to some degree.

[5] A scientific example of self-evidencing is the observed red shift in the light spectrum of certain stars, which is best explained by the expansion of the universe, where the occurrence of the red shift is a core part of our reason for believing the universe expands (see Hempel 1965).

[6] Formally, at this point it makes sense to talk about models rather than hypotheses, such that hypotheses are generated under a model and shaped through inference to give evidence for the model.

[7] Similarly, attention, in the shape of optimizing precisions, and complexity reduction, in the shape of minimizing the cross entropy between priors and

selected hypotheses, both maximize evidence for the model by making it better able to explain away surprising evidence.

[8] An earlier debate, centered on the dangers of the sensory veil, had a more metaphysical bent (see for example McDowell 1994; Putnam 1981). Clark (2012: 766) rightly points out that it is not clear how to separate out the epistemic and metaphysical aspects of these debates, and he goes on to argue that PEM's sensory veil does not prevent genuine contact with the external world. To me, it seems that PEM is exactly the kind of view targeted by the earlier metaphysical debate.

[9] This can be seen by considering a PEM version (Hohwy 2013: Ch. 8) of Searle's (1980) famous Chinese room argument.

[10] For a framework for building up an internal model of self, based on mounting levels of embodiment in this sense, see Metzinger (2013).

[11] Perhaps this type of position can be embraced, in an argument for indeterminacy of cognitive boundaries (*cf*. Anderson, Richardson *et al*. 2012).

[12] This reiterates in PEM-clothing insights from the good regulator theorem and the self-organization literature (*e.g*., Conant and Ashby 1970).

[13] See Grush (2003) for convincing discussion of a broad range of embodied cognition claims, in the context of an emulation theory of cognition that shares much with PEM.

[14] 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 or external states, are rendered uninformative once the states of the blanket are known. In the same vein, if we treat the mind as a dynamical system, described by a set of differential equations (which in our case describes free-energy minimization), then we can observe that none of the variables in these equations refer to objects beyond this narrow, neurocentric system; for this approach, see Shapiro (2013).

[15] An interesting proposal here is found in Orlandi (2013), which develops the idea that the inferential view is incorrect and that an embedded, evolutionary, Gibsonian notion of perception should be adopted on which the visual system 'relies' on facts in nature but without representing them. If, contrary to the spirit of Orlandi's proposal, the notion of reliance was translated into the PEM framework, it would be the idea that empirical priors are shaped at multiple time scales and under considerations of precision. These represented regularities must be internally represented to take account of non-linear flows of sensory input.

[16] Using PEM, Clark (2009; 2012; see also Ward 2012) comprehensively rebuts another incarnation of enactivism (Noë 2006) that does not rely on self-enabling and self-organisation; see also Seth (2014) who reworks this kind of enactivism in inferentialist terms.

[17] This point can be given an equivalent formulation in terms of attention. As mentioned in the introduction, according to PEM, optimization of the precision of prediction errors is the mechanism that constitutes attention. Applying this to the proprioceptive domain, the idea is that action is triggered when attention is

withdrawn from the actual proprioceptive input (since it is expected to be imprecise). Formulated in these terms, acting is thus allocation of inner attention.