Predictive Cognition and Its Philosophical Implications: What Doesn't (Yet) Follow from the Enthusiasm about PEM[1]

Submitted by Brendan Fleig-Goldstein
Thesis Committee:
Brian Epstein, Daniel Dennett

Tufts University
Department of Philosophy
Senior Honors Thesis
2015

[1] This title is taken from a comment made by Professor Dennett

0) Introduction

In the past recent years, cognitive scientists and philosophers have put forth a picture of cognition in which the sole activity of the mind is to predict its future states. Under this view, all cognitive processes, including perception, attention, and action, can be analyzed as a part of this predictive process. Many of the theorists arguing for this view—most prominently, Andy Clark, Jakob Hohwy, and Karl Friston—have characterized these mental predictive processes as involving inference to the best explanation (abduction). In the first half of this paper, I will lay out the theory of the predictive mind, and argue that characterizing it in terms of inference to the best explanation is both theoretically unnecessary and not required by the empirical evidence. In the second half of this paper, I will examine what, exactly, the theory of the predictive mind is a theory of, and then discuss various philosophical positions theorists have claimed follow from this theory. I will suggest that most of the proposed philosophical implications do not follow from the core claims of predictive cognition, but instead follow (if they do at all) from various unnecessary ways of characterizing the theory.

First, I will lay out the theory of the predictive cognition. In the prevailing treatments, predictive cognition occurs as a part of hierarchical cognitive processing. I will briefly state what hierarchical processing is, and how predictive theories take place within this picture. Namely, they change the primary direction of processing in the hierarchy from "bottom-up" to "top-down." Next, I will give a brief overview of the empirical evidence for predictive cognition. After, I will introduce theories of Bayesian cognition, and explain how they are distinct from predictive cognitive theories, but can be used to provide a specific account of how

the mind generates and modifies its predictions. Next, I will discuss inference to the best explanation, or abduction, and how it fits into mental prediction. I will argue that, though theorists have couched predictive cognition in terms of an explanatory process, doing so is not necessary given the claims of predictive cognition, as well as the empirical data in support of predictive cognition.

In the second half of the paper, I will discuss how predictive cognition theories have been extended from perceptual processes to include processes such as attention and action. I will then discuss the "dark room" problem, which has been raised against these extended theories. I will review various replies to the dark room problem, and discuss the advantages and disadvantages that come with these replies. After, I will discuss predictive cognition as a complete theory of cognition, as a complete theory of the brain, and as a complete theory of the mind. I will conclude that characterizing predictive cognition as a complete theory of the brain or of the mind requires controversial claims unrelated to predictive cognition. Consequently, when considering philosophical implications of ambitious predictive cognition theories, we should be interested primarily in the philosophical implications of predictive cognition, characterized simply as a complete theory of cognition. Finally, I will go on to suggest that most of the implications theorists have so far claimed follow from predictive cognition in fact follow from predictive cognition as a complete theory of the brain or of the mind, and not as a complete theory of cognition. These proposed implications therefore assume controversial claims unrelated to PEM.

1) What is Prediction Error Minimization?

Let the term "PEM theories" denote a class of theories that treat cognitive processes as prediction error minimization processes. PEM theories claim that: 1) the mind is a device that

takes in input that has causal origins in an external environment 2) the mind generates expectations (or predictions) of future input, and 3) based on discrepancies between expected and actual input, the mind modifies the generation of future expectations in an attempt to minimize these discrepancies (i.e., in an attempt to minimize prediction error).

Contemporary PEM treatments (Clark 2012; Hohwy 2014) characterize prediction processes as occurring hierarchically within the mind. Though predictive minds need not also be hierarchical minds, hierarchical cognition has played importantly in recent PEM theories' ability to explain empirical data, and at this point is more or less as well supported as PEM in general. Hierarchical cognitive theories join nicely with PEM theories, and given the current conversation in the literature, are just as important to understanding contemporary PEM as the concept of prediction itself.

Hierarchical cognitive theories claim that mental processing can be understood as occurring at different "levels." From the hierarchical picture of mind we inherit the metaphor that some cognition is "top-down" and some is "bottom-up." This verticality language is shorthand rarely fleshed out in full, but the basic idea is that we can characterize mental processes in terms of their "distance" from proximal stimuli.[2] Distance here is cashed out in terms of how many temporally intermediary mental processes take place between one mental process and another.[3] Lower level processes are those closer to proximal stimuli, and higher level processes are those farther from proximal stimuli.

[2] The term "proximal stimulus" is both used to refer to the energy from the external world that gets transduced by sensory organs, as well as to refer to the mental activity that is the most immediate result of the process of transduction. Here I use the latter sense, though arguably both meanings could provide a meaningful analysis of mental distance.
[3] The hierarchical mind assumes, if not modularity of mental processes, then a degree of sequentiality of mental processing—though, importantly, the sequential direction need not be specified.

It is important to understand that the hierarchy under discussion is not a hierarchy of consciousness or of mental complexity. Higher level processes are sometimes written about as those processes related to sophisticated types of cognition such as planning, self-reflection, general world knowledge, and abstract reasoning, to name a few. High level processes, however, are not picked out by their sophistication or by being "above the surface" of consciousness. A process could be quite advanced and only present in the most intelligent of creatures, but not be considered high level. A process could also be unconsciously performed and relatively simple and yet be high level. Processes are high level solely in virtue of their sequential or temporal distance in processing from the processing that occurs immediately after transduction of sensory information. We do not need a concept of consciousness, nor do we need a measure of mental complexity, to discuss the relevant cognitive hierarchy. On this characterization, a given cognitive process is not high level *simpliciter*, but rather high level relative to some other cognitive process. We need not draw a line at which "medium level" processes end and high level properties begin—we need not assume any substantive "height divisions" out in the world.

The central insight of PEM is the reversal of the primary direction of processing within this vertical mental framework. Under traditional conceptions of mental processing, it was common to suppose that the external world provides the mind with signals that get processed and passed "vertically upward" for further processing. In contrast, under the new PEM framework, higher levels predict, or make guesses of, what will occur at the lower levels. Each level formulates a guess of what the state of the level below will be, and passes its guess downward. The discrepancy (or error) between the guess and the actual state is then passed back upward. As

error signals are received, states modify their predictions until the error is eliminated or minimized. Instead of signals from the world propagating up the mind, the mind's expectations propagate down, and it is solely the error signals that propagate upward. Instead of perception occurring as a mainly reactive process to the incoming signal, perception is reconstrued as an active process of the mind. Thus we have a picture of the mind as a prediction error minimization device.[4]

Though perception was the first domain of cognition to receive a PEM treatment, this type of analysis has been extended to every type of cognition. Indeed, those who have claimed that the mind is *exhaustively* a PEM device are claiming exactly that all cognition is amenable to a PEM analysis. So far, however, the most sophisticated treatments of cognitive domains other than perception include action and attention, as well as memory, concept formation, and general learning paradigms.

2) Empirical Research and Evidence for PEM

The empirical data used as evidence for PEM frameworks consists largely in computational modeling of cognitive processes, psychophysical data (e.g., quantifications of people's ability to accomplish various cognitive tasks), and to a much lesser degree neuroimaging and neuroscientific data. As Clark (2013) writes, "[d]irect neuroscientific testing of the hierarchical predictive coding model...remains in its infancy. The best current evidence tends to be indirect…" (p. 191). Among the evidence coming from neuroscience and

[4] It is sometimes said that the mind is a predictor of future *input*, and other times that the mind is a predictor of *its own future states*. The two characterizations more or less correspond to the two senses of "proximal stimulus", as described in footnote 2. Predictive minds are trying to guess the mental activity that is the most immediate result of transduction. Phrased as such, the mind is guessing its own future states. But inasmuch as there is a mapping between the energy transduced by the senses and the mental activity that is the immediate result of transduction, we can also describe predictive minds as attempting to predict the energy from the external world. In that sense, predictive minds predict future input.

neuroimaging, one of the most compelling studies comes from Murray et al. (2002). The brain region known as V1, also called the primary visual cortex, is the area of the visual cortex closest via neural connections to proximal visual stimuli (i.e., closest to the eyes). Given the traditional, mainly bottom-up, reactive view of perception, one would expect V1 activity to remain unchanging as a subject looks at an unchanging visual scene. Using fMRI neuroimaging, however, Murray et al. showed that as participants continue to look at an unchanging image, activity in V1 eventually begins to decrease. This decreased activity makes sense within a predictive brain: as error signals from V1 are minimized and higher level predictions are brought in line with lower level states, activity in V1 need not continue.

Consider now an example of computation modeling that has been used as evidence for PEM. PEM theories predict that V2 (the region in the visual cortex second closest via neural connections to proximal visual stimuli) should send predictions to V1, and that V1 should send error signals to V2. As Clark (2013) describes, Rao and Ballard (1999) programed a computational network modeling these areas of the visual cortex with such connections and trained the network on image patches from various scenes. Using learning algorithms designed to minimize prediction error in the long term, the network learned to use different layers of its computational system to perform different types of feature recognition in a similar arrangement as the human visual cortex. That is, the first layers of the network took on roles such as edge detection and evaluating line orientation, while layers further from the input took on roles such as recognizing larger patterns of spatial configuration. It is fairly impressive that a computational model, simply endowed with properties entailed by PEM, generated "on its own"

a visual processing structure isomorphic to our own, as revealed by cognitive neuroscience. The computational model, for these reasons, has been considered evidence for PEM.

Consider next an example of psychophysical data, which both Clark (2013) and Hohwy et al. (2008) have cited as evidence for PEM. Under artificial laboratory settings, scientists are able to elicit in subjects a visual phenomenon known as binocular rivalry. Binocular rivalry occurs, for instance, when researchers present a subject with an image of a house in one eye, and of a face in the other eye. Instead of seeing both images side by side, or a face-house mélange, participants will typically report seeing *either* a house *or* a face. After a time, the image will then "switch" and they will report seeing the other object.

Hohwy et al. explain this phenomenon as follows: It is extremely unlikely to our human minds that an object could be both a face and a house. Accordingly, world knowledge-involving high level processes are tuned such that the mind is unlikely to settle into a perception as of an object that is both a face and a house. When a subject is shown the relevant images, prediction error minimization processes—in an effort to best minimize error over time—consequently settle into a perception as of one or the other (e.g., just the house). As they do so, the signals from the other image (the face) begin to generate more prediction errors at the lower level, and these error signals begin to propagate up the hierarchy. Predictions at increasingly higher levels are gradually modified in an attempt to reduce the new error signals, and eventually the subject reports an image switch from the house to the face. Once this happens, the signals from the *house* at the lower level begin to generate more prediction errors, and these error signals propagate up the hierarchy...and the process repeats itself, causing cycles of upward error propagation that result in the subject reporting oscillating images. Thus PEM provides a

plausible explanation of the phenomenon of binocular rivalry, and binocular rivalry can be seen as evidence for PEM.

Note that the above characterization of PEM has made no mention of Bayesianism. Indeed, the above characterization should illustrate that it is entirely unnecessary to use Bayesian concepts in a treatment of PEM. PEM claims just that certain (or all) cognitive processes can be analyzed as predictive processes.[5] Bayesian PEM makes the more specific claim that these predictive processes are making use of a form of Bayesian inference.[6] That is, mental predictions—which yield guesses of future input—get assigned probabilities, and as more input comes in, the assigned probabilities get updated according to Bayes' theorem.

It is not at all clear, however, why the way in which probabilities get assigned or updated must be specifically Bayesian. It is perfectly within the realm of possibility that the assignment of probabilities more closely resembles, for instance, Solomonoff induction rather than Bayesian inference.

Note, however, that if PEM does involve the assignment and updating of probabilities, it is not at all clear why the way in which that happens must be specifically Bayesian. It is perfectly within the realm of possibility that the assignment of probabilities more closely resembles, for instance, Solomonoff induction rather than Bayesian inference.

And note also, that it is not clear why there need be probability assignments and updates at all in PEM. As has already been mentioned, if a predictive mind is probabilistic, then each level does not arrive at a single, definite prediction of the level below. Instead, each level generates a range of guesses of the level below, and assigns to each guess a different probability.

[5] More specifically, processes involved in making predictions, comparing predictions to actual states, and modifying future generation of predictions.

[6] Or as an approximation of Bayesian inference, or a parameterized version, etc.

Though there may be a dominant prediction—that is, a prediction assigned the highest probability—there remains at each level multiple predictions throughout the process. As more input comes in, and the probability of each guess gets updated, the dominant prediction may change. Notice that the process of assigning a probability to a guess is only useful insofar as that probability gets compared to the probabilities of other guesses. The purpose of having a process of assigning and updating probabilities of hypotheses is to allow for the comparison of competing hypotheses, in an attempt to select the best ones. Bayesian inference is a process of assigning and updating probabilities, and so if PEM is Bayesian, each level of the mind should have competing hypotheses that are compared.

But there need not be multiple predictions at each level. Each level could consist in a process that results in the generation of a single, definite prediction of the lower level. As long as the process of generation is amenable to modification and fine tuning as new input comes in, there is no reason prediction can not occur in such a way. Any means by which levels of the predictive mind get modified in response to error signals so as to dampen them is compatible with PEM theories. There need not be any probabilities within a predictive mind at all, so long as there is error minimization. Moreover, while predictive minds need not be Bayesian, Bayesian minds need not be predictive.[7]

Though theoretically detachable, PEM theories and the Bayesian brain hypothesis do join nicely together. A Bayesian framework is a plausible means by which mental prediction is performed: Bayesian models lend themselves nicely to top-down processing, and a large amount

[7] A mind could employ Bayesian inferential processes to select probable representations of the current state of the external world, and at the same time not be involved in generating guesses of future states or input. Under such a scheme, perception would occur as a reactive process—as in the traditional conception of mental processing—and Bayesian inference would be still be employed.

of the successful computational models of predictive processing have made use of Bayesian statistics. The frequent conflation of the theory of hierarchical predictive processing and Bayesian cognition is, if not justified, understandable. In practice they go together quite frequently. Note, however, that there have been worries as to whether the cognitive processes accomplished by the mind are too complex to be both Bayesian and computationally tractable (Blokpoel et al., 2013). While these worries may not pan out, theorists should do well not to conflate the implications of the different theories, and understand that the success of one does not ride on the success of the other.

How neurons or any other material could be implementing predictive processes (Bayesian or not), remains entirely unknown. We do not understand, for example, how populations of neurons could be making predictions of the activity of other populations of neurons, or what form a prediction takes in a population of neurons. There have been proposals as to how groups of neuronal cells *could in principle* implement Bayesian statistics (e.g., Deneve, 2008), but there is no substantial data to suggest that brains are in fact implementing Bayes in this way. There have also been no single neuron studies in support of PEM. Given such a situation, neuroscience has not been a significant source of data as evidence for PEM or Bayesian theories of cognition.

3) Abductive Minds

I will now lay out the view, which I will refer to as abductive PEM, that explanation plays a central role in PEM. This view holds that the mind predicts by explaining input. This view is prevalent in theoretical treatments of PEM, if not always fully laid out. Explanatory language is present in most of the recent influential works on PEM (Friston 2005, 2010, 2012;

Clark 2013; Hohwy 2014). These authors, however, have not offered extensive or even explicit justification for their employment of explanatory terminology. After laying out the basic abductive PEM position, I will make a few general points about the nature of abduction and explanation. Afterwards, I will turn back to abductive PEM and show that it is an unnecessary way of characterizing PEM. I will then discuss some of the motivations given for invoking mental explanations in PEM, and suggest that they are not particularly convincing.

The basic picture of abductive PEM is as follows. Each level of the mind does not just predict the state of the level below. Each level *explains* why the level below is in its current state, and uses that explanation to generate predictions of future states. As further input comes in, explanations are supported or discouraged depending on whether the predictions they yielded are realized, and future states or input act as evidence for or against explanations.

Abductive inference is inference to the best explanation. Suppose one day you notice that your floorboards are damaged in a particular way, and you come to the conclusion that there are termites in your house. You have just made an inference to the best explanation. The inference is not deductive, as the conclusion is not entailed by logical form. Neither is the inference inductive, as the conclusion does not depend solely on some form of probabilistic reasoning. That there are termites in your house is not just *likely* given the damaged floorboards; it *explains* why your floorboards are damaged. There is disagreement over what exactly qualifies as an explanation. But it is relatively uncontroversial that an explanation is a relation between some state of affairs and another state of affairs. That is, when one observes certain events or conditions of the world, one explains them by appeal to other events or

conditions of the world. You observe that you feel nauseous, and you explain this fact by appealing to *the further fact* that you have not eaten well today.

Note that explanation is not a reflexive relation; events cannot explain themselves. That the floorboards are damaged does not explain why the floorboards are damaged. To explain must be to make a claim about the world that goes beyond the the claims about the event in need of explanation. Check to make sure you are on board with this characterization of explanation, as it will be important in what follows.

As has already been mentioned, abductive PEM is the view that the mind predicts future input by explaining the input that has occurred so far. By generating explanations, the mind is making inferences to the best explanations of its own current states. It is important to understand that in abductive PEM, the explanations are in addition to, and not the same thing as, the predictions. The explanations generate, or give rise to, the predictions. So there is an additional, third element involved in abductive PEM, and not just a richer characterization of what is normally involved. In non-abductive PEM, there are predictions and input.[8] In abductive PEM, there are events in need of explanation (the input), explanatory hypotheses intended to explain these events, and also the predictions of future events that the explanatory hypotheses give rise to.

I will now argue that abductive PEM is an unnecessary characterization of PEM. As a first observation, the above sections of this paper demonstrate that it is possible to lay out the claims of PEM theories of cognition without using any explanatory language whatsoever. And as the above paragraph argues, abductive PEM requires a further, additional specification of the

[8] More fully, predictions of the state of the level below and the actual state of the level below.

PEM process. Namely, abductive PEM requires that predictions are generated from explanations. But predictions do not need to be generated from explanations, and can instead be generated from other processes. For example, statistical processes can be used to generate predictions, and these need not involve explanations.

One might claim that any time you generate a prediction, even by statistical means, you are also (inadvertently, as the case may be) offering an explanation. For example, one might argue that if you are trying to predict subsequent numbers in a sequence, and you come upon an algorithm that reliably predicts those numbers, then that algorithm *explains* the number sequence. For example, if you are presented with the sequence (1, 1, 2, 3, 5), then it is explained by the rule '$A_0 = 1$, $A_1 = 1$, and $A_X = A_{X-2} + A_{X-1}$'. In the same way, when the brain predicts future states, one might argue, *whatever* process is used to generate those predictions, that process explains the input. The underlying idea here is that all successful predictions can be considered explanations (and perhaps all unsuccessful predictions can be considered failed explanations). If this claim is true, then any and all types of predictive cognition would automatically involve explanation. Consequently, PEM would entail abductive PEM, and any use of explanatory language in characterizing PEM would be justified.

Predictions are not explanations, however, and this fact follows from the above points about the general nature of explanation. In the following paragraphs, I will illustrate clearly the difference between predicting and explaining, and how the former can be accomplished without the latter.

Consider again the process of guessing the next number in a sequence. There is a difference between predicting the next number in a sequence and explaining why the next

number in a sequence will be what it is. Imagine that I am in a room and writing down numbers on pieces of paper and sliding them out to you from under the door. Suppose that inside the room, "$A_0 = 1$ and $A_n = A_{n-1} + 2$" is written on a chalkboard and I am writing down the numbers I send to you by following this rule. You receive the numbers (1, 3, 5, 7). To predict the next number, you decide to use the rule '$A_0 = 1$ and $A_1 = 3$ and $A_n = A_{n-2} + 4$': a rule that, had you had it from the start, would have successfully predicted all the numbers so far. Using your rule, you predict I will slip '9' under the door, and I do so. Your prediction is correct. In fact, using your method, your predictions will be correct for as long as this process continues. You have truly come upon a successful method of prediction. But you have not explained anything yet. Just because you can say "9" before a 9 comes out, and even though you have a systematic approach to producing your predictions, it does not mean you possess an explanation.

To give an explanation of the numbers you are getting, you must make a claim about causal relationships in the world giving rise to the numbers. This point follows from what was said earlier about about the nature of explanation: explanations must describe certain states of affairs in the world to explain other states of affairs. In this case, the relevant causes happen to be inside the room. So to offer an explanation, you cannot just *use* a mathematical rule, you must *claim* something along the lines of "the rule '$A_0 = 1$ and $A_1 = 3$ and $A_n = A_{n-2} + 4$' is written on a chalkboard inside the room and Brendan is following this rule to produce the numbers." Now you have offered an explanation—though it happens to be incorrect. A different rule is written on the chalkboard, and therefore you have incorrectly described the condition the world is in, and therefore your explanation is wrong.

But the point has been made: the process of prediction can be done without engaging in a process of explanation. You can use a method of prediction without that method making claims about conditions of the world, and doing the latter is a necessary part of offering an explanation. To predict by using a process like a mathematical algorithm or another type of rule by itself does not explain anything. In the same way, mental prediction processes need not be explaining anything, as they need not be making claims about the conditions of the world in order to arrive at predictions of future input. They need only be employing processes that result in better and better predictions of input. PEM consequently need not be abductive PEM.

These points are not intended to argue that abductive PEM is false. They are intended only to show that there is no reason PEM must involve explanation. Motivations for the PEM picture, therefore, are not automatically going to be motivations for abductive PEM. Further motivations must be given by any theorist who wishes to characterize PEM in such a way.

In the last part of this paper, I will briefly survey possible motivations for abductive PEM. I invite the reader to consider further motivations. Ultimately, however, I will suggest that these motivations fall short of what is necessary to fully motivate abductive PEM.

The first appealing quality to abductive PEM is that it fits nicely with the phenomenology of our thought processes and how we often consciously generate predictions. That is, it often seems like we, from the first person, generate predictions by using explanations as a starting point. In ordinary life (although in scientific theorizing as well) explanations lend themselves nicely to the generation of predictions. You see the damaged floorboards, and explain this fact by claiming that there are termites in the house. From this explanation, you come quite quickly to the prediction that if you were to open up the floors or walls, or really look through the house,

you would be able to find the termites. Without that explanation, the prediction seems unlikely to have followed. Or say you believe you are nauseous because you have not eaten. Out of this explanation, you come to the prediction that if you eat, you will feel better. At the personal level, explanations beget predictions, and so it is not entirely unreasonable to suppose that at the subpersonal level, the mind generates predictions from explanations.

Second, it is easy to see how abductive PEM works with Bayesian PEM. Under an abductive Bayesian PEM framework, each level of the mind produces multiple explanations for the state of the level below. These explanations are assigned prior probabilities. The explanations produce various predictions, and each prediction is assigned a different probability under each explanation. As input comes in, the probability of the explanations are updated according to Bayes' theorem. While this point may not seem exhilarating, if abductive PEM were not compatible with Bayesian PEM, we might have less reason to believe in abductive PEM. As the case is, however, abductive PEM and Bayesian PEM work fine together.

Third, it makes sense that a model of external causes of stimuli would be able to give rise to accurate predictions of future stimuli. That is, it is not difficult to see how having an idea of what is going on out in the world will give you an idea of what is to come from there.

These, in my mind, are the greatest virtues to abductive PEM. It is not clear that abductive PEM is a simpler or more elegant theory of cognition than non-abductive PEM. And it is not clear what empirical research abductive PEM is supposed to better explain. Theorists who have characterized PEM as abductive have not spent the time citing studies that they believe specifically motivate the abductive version of PEM.

PEM is a theory intended to describe the primary activity of the mind. And so to describe PEM as a process of abduction is a fairly substantial claim about the nature of cognition. I suggest it is not clear that the above motivations alone are satisfactory for fully motivating abducting PEM.

Nevertheless, PEM thinkers have been recently using explanatory language, and explicitly claiming processes of abduction, in their PEM treatments. Philosopher Jakob Hohwy, author of The Predictive Mind (2013), has most explicitly put forth the abductive PEM position. He states in no uncertain terms (2014) that "PEM is essentially inference to the best explanation, cast in (empirical, variational) Bayesian terms" (p. 5). Philosopher Andy Clark, while not writing explicitly of inference to the best explanation, has used unambiguously explanatory language in describing mental prediction. He (2013) writes,

> Perception thus involves "explaining away" the driving (incoming) sensory signal by matching it with a cascade of predictions pitched at a variety of spatial and temporal scales. These predictions reflect what the system already knows about the world (including the body)...On this model, perception demands the success of some mutually supportive stack of states of a generative model...at minimizing prediction error by hypothesizing an interacting set of distal causes that predict, accommodate, and (thus) "explain away" the driving sensory signal. (p. 187)

And neuroscientist Karl Friston's influential characterizations of PEM (2005; 2010; 2012) have also relied heavily on the notion of explanation. Friston (2010) writes,

> In this view [referring to the Bayesian Brain Hypothesis], the brain is an inference machine that actively predicts and explains its sensations...central to this hypothesis is a probabilistic model that can generate predictions, against which sensory samples are tested to update beliefs about their causes. This generative model is decomposed into a likelihood (the probability of sensory data, given their causes) and a prior (the a priori probability of those causes). Perception then becomes the process of inverting the likelihood model (mapping from causes to sensations) to access the posterior probability of the causes, given sensory data (mapping from sensations to causes). (p. 3)

In their papers, these authors do not support these abductive claims. They instead support the theory of the mind as a prediction device. But it was the intention of this paper to show that PEM does not entail any abductive process. After surveying motivations for abductive PEM, I propose that abductive PEM is insufficiently motivated.

I have attempted to show that PEM, Bayesian cognition, and abductive cognition—though all blended together in the recent literature—are each different, separable claims about the nature of cognition. When Hohwy (2014) writes that "PEM is essentially inference to the best explanation, cast in (empirical, variational) Bayesian terms," he is simply wrong. PEM is not essentially any of those things. Abductive PEM is not entailed by general PEM theories of cognition, and is not noticeably motivated by empirical research. It is unclear what advantage there is to adopting abductive PEM. It is my hope that this paper encourages theorists who support abductive PEM to offer further motivations for, and clarify the advantages of abductive PEM.

4. The Whole PEM Iguana

In the next half of this paper, I intend to examine three different ambitious claims about the principles of PEM: 1) all cognitive processes can be understood using the principles of PEM 2) the entire biological function of the brain, qua organ, can be understood using the principles of PEM, and 3) all psychological facts can be explained by the principles of PEM. I will argue that these are three distinct, separable claims. Further, I will argue that it is claim 1) that is both the most plausible and philosophically relevant position. Next, I will argue that most of the philosophical implications theorists have claimed follow from PEM theories do not follow from claim 1), but follow instead from claims 2) or 3).

Before turning to these three different claims, however, I will spend this section illustrating how PEM analyses can be extended to areas of cognition beyond perception. The paper so far has focused on PEM as it applies to perception. This focus was made mainly for simplicity's sake, but the overall above points—including the main point that PEM does not require abduction—were made without reference to PEM as a specifically perceptual process, and the above arguments can be run for PEM as it applies to any specific domain of cognition, *mutatis mutandis*. I now turn to PEM as it applies to various other domains of cognition. I will then discuss what is known as the "dark room problem" and various replies to this problem. These points will serve to illustrate the strong PEM picture, as well as show that PEM has the potential to elegantly tie together several domains of cognition.

First, let us consider PEM theories of attention. The idea here is that the mind has limited resources for predicting input. Given the limited processing power of the mind, and the large quantity of incoming mental activity that is produced by various sensory organs transducing environmental energy, it is plausible that the mind can make predictions—and then correct for error—for only some of this mental activity. That may seem bizarre, but the ultimate goal is to minimize the total prediction error in the long run, and so the mind really only has to engage with the input that is most valuable for that end.

And which input is the most valuable for that end? Under PEM theories of attention, it is this question that attentional processes seek to answer (Feldman and Friston 2010, Hohwy 2012). Attention is reconstrued as the process of selecting where to devote predictive resources, so as to minimize prediction error in the long run. In order to accomplish this task, it is posited that the mind, in addition to predicting future input, makes predictions about how unpredictable different

input will be. Input that is expected to be highly unpredictable (and remain highly unpredictable) will be deemed not particularly useful to devote predictive resources to, and the same goes for input that is expected to be highly predictable. The most valuable input to direct predictive processes toward is input somewhere in between very predictable and very unpredictable: predictable enough to be manageable, but unpredictable enough to allow the mind to learn something new when it minimizes the error.

Second, let us consider PEM theories of action. PEM is a process of getting a better and better match between predictions and input. There are logically two distinct ways of achieving such a better match: the mind can change its predictions, or the mind can change its input. The former is perception, and the latter is action. During exteroception and interoception, the mind minimizes prediction error by changing its predictions to match the input. But during proprioception, for at least some proprioceptive input, the mind minimizes prediction error by actuating muscle movements to change proprioceptive states and match the input to the predictions. One raises one's arm by predicting that it is raised, and subsequently moving one's arm to make such a prediction come true. In this sense, PEM characterizes perception and action as more deeply linked and symmetrical than many other theories of action.

Why would the mind make incorrect predictions of proprioceptive input just so that it may correct them? Hohwy and others argue that the mind, by being wrong in the short term, can be more accurate in the long term. That is, by being able to reposition one's sensory organs and gain a new perspective on the world, the mind can reposition itself to a better vantage point in which it may more successfully make predictions. And the way to reposition one's sensory organs is to be wrong about the immediate proprioceptive input. A small price to pay to, say, be

able to move one's head and direct the eyes toward all sorts of input from new areas of the world. Such "sampling" of the world will greatly help to tune the predictive machinery of the mind to be more successful in the long run.

How will an organism endowed with PEM perceptual, attentional, and actional processes behave? Given the comments above about the value of input that is somewhere between predictable and unpredictable, it seems the following may be the case. The principles of PEM-style perception and action entail that the best behavior for a PEM organism to exhibit is a balance between exploring novel parts of the environment, and remaining in predictable, stable parts of the environment. Such a presciption for animal behavior would not be a ridiculous one.

Andy Clark (2013), however, reminds us of a worry brought up by David Mumford, a mathematician who has looked at visual and cortical systems from a computational perspective, back in 1992. Mumford worries that a predictive cognitive architecture, as we have been discussing, entails that a PEM organism should trend towards an ideal "ultimate stable state" in which there is perfect prediction and no prediction error. Such a state would be accomplished if an organism were to confine itself to a particularly stable part of its environment. For example, if an organism were to bury itself in a hole and block out all light or sound. If this is the ideal state for a PEM organism, and we humans are PEM organisms, it is a wonder that we do not devote all our energy into the construction of sensory-deprivation tanks, rooms to enjoy silent darkness, and other technologies to deliver perfectly constant sensory experiences.

Clark (2013)'s response to this "dark room problem" is to suppose that our predictive machinery puts high probabilities on an organism exploring its environment, moving around, changing states, and so forth. Since organisms as a matter of fact live in changing environments

that require them to move and change states, an organism's mind either learns, or is programmed, to expect that the organism will move and change states. Thus the mind predicts that the organism will move and change states, and consequently the organism does move and change states.

Hohwy (2015)'s response to the dark room problem is similar, though he brings in the concept of homeostasis and puts it to good use. His response is essentially as follows. Each organism has a range of possible biological states that it can occupy. An organism can, for example, be very hungry, very full, or be somewhere in between; or their blood can be very acidic, alkaline, or somewhere in between; or their body temperature can be high or low or somewhere in between, etc. These "somewhere in between" states are the organism's homeostatic states. PEM processes are tuned to predict (or put high probabilities on, in a probabilistic mind) an organism being in homeostasis, and when perceptual input reports that an organism has strayed from homeostasis, the mind, in an attempt to minimize the prediction error, will act to bring the organism back into its homeostatic state. Such actions can unfold over a long period of time and have multiple sub-actions. Thus, PEM theories allow for an organism capable of rich planning and having sub-goals as commonly conceived (e.g., opening the car door to get in, to drive to the store, to get the bread, to make the sandwich…). The dark room is avoided, as moving an organism into a dark room will not bring the organism back into homeostasis, and will therefore not minimize prediction error in the long run.

The virtue to Clark and Hohwy's similar solutions to the dark room problem is that there is no need to supplement a PEM theory of cognition by positing a box in the mind with things called "desires"—the desire for food, the desire to survive—that communicates with the PEM

box. We do not need to create a separate system for desires, and "hook" this system up to the PEM system. Instead, we need only suppose that the PEM system is capable of keeping track of an organism's biological states, and predicts that the organism will be in its homeostatic states. Note that the PEM system does not need to have any understanding or sophisticated awareness of what homeostasis is, it just needs to predict that the organism is in the states *that happen to be* the organism's homeostatic states. Once these conditions have been met, we should get desires, goals, and something like planning, "for free" out of the PEM framework.

The downside to both Clark and Hohwy's solutions to the dark room problem is that both seem to involve the positing of "hyperpriors."[9] Hyperpriors are either very high level predictions, or tunings throughout the hierarchy, that place restraints on the sorts of predictions the system can make. Hyperpriors are difficult, or impossible to modify during the PEM process, and it is this feature that makes hyperpriors capable of providing a solution to the dark room problem. In the dark room case, the relevant hyperpriors are the predictions that the organism will stay in its homeostatic states (in Hohwy's version), or continue to explore the environment and be in motion (in Clark's version). If these predictions were subject to revision as much as any other prediction, then if an organism ever had easy access to a dark room, it is not clear why the mind would not just ditch those predictions for the chance to enjoy the perfectly predictable states the dark room would afford. That is, these predictions (the hyperpriors) have to be more or less unrevisable, or else they are not solutions to the dark room problem.

[9] 'Hyperpriors' connotates probability, and I have been trying to remain neutral to whether PEM minds are probabilistic. In the discussion that follows, the term 'hyper-predictions' can be used interchangeably with 'hyperpriors'. I use the term 'hyperpriors' though, as it is what others use and it reads better.

Any theory that posits hyperpriors of these sorts then evokes the question "where do the hyperpriors come from?" This question is more difficult than answering "where do the first priors (or predictions) come from?" The latter question can be answered by supposing that the first guesses are random, or an equal probability is assigned to all guesses. As more input comes in, the process of generating future predictions is modified, and the organism is off and running. But what would make guesses unrevisable?

Hyperpriors are employed in PEM elsewhere beyond the dark room problem. In a sense, they are what cause visual illusions. Whenever we perceive a visual illusion, it is a case where the mind makes a prediction (for example, that the light is coming from above in a visual scene), that prediction is incorrect, but because the mind is so unwilling to revise that prediction, we end up with a visual illusion. The fact that the mind is unwilling to revise these sorts of predictions is what makes them hyperpriors in the sense I have been discussing.

There is a difference, however, between the sorts of hyperpriors involved in perceptual illusions and those that are needed to solve the dark room problem. Namely, hyperpriors of the first sort reflect a truth about the environment an organism finds itself in (light usually does come from above in visual scenes in natural environments). Hyperpriors of the second sort, on the other hand, reflect a truth about the organism-environment only insofar as the organism already possesses these hyperpriors: the prediction that an organism will remain in homeostasis is, according to Hohwy, what causes the organism to remain in (or close to) homeostasis. Hyperpriors of the second sort make themselves true, while the hyperpriors of the first sort are made true by features of the environment.[10]

[10] It may be tempting to suppose that the hyperpriors that reflect a truth about the environment are predictions about perception, while the "self-fulfilling prophecy" hyperpriors are predictions about actions, and that that difference

Call the first sort of hyperpriors—hard to revise predictions that exist because of features in the environment in which an organism (type) evolved or an organism (token) developed—descriptive hyperpriors. And call the second sort—ones that make themselves true—prescriptive hyperpriors. Descriptive hyperpriors are descriptive because they describe how the organism-environment *is*. They go wrong whenever the environment does not meet that description. Prescriptive hyperpriors are prescriptive because they describe how the organism-environment *should* be for the organism to be evolutionarily fit. They go wrong whenever the organism-environment changes in such a way, such that satisfying that description no longer makes the organism evolutionarily fit.

Possessing both sorts of hyperpriors makes an organism evolutionarily fit. It is (evolutionarily) good to have heuristics for perceiving your environment,[11] and it is (evolutionarily) good to be in homeostasis. But the evolutionary advantage of having certain descriptive hyperpriors depends only on features of the organism-environment that organisms are capable of perceptually discerning (organisms have eyes that can see light coming from different directions). The evolutionary advantage of having certain prescriptive hyperpriors, on the other hand, depends on facts about evolutionary fitness of an organism relative to its environment, and it is not clear that organisms are capable of perceptually discerning *these* facts.

So, then, how does an organism acquire prescriptive hyperpriors, if it is unable to perceptually discern the facts on which prescriptive hyperpriors depend? The point is that since an organism can perceptually discern the facts on which descriptive hyperpriors depend, it seems

---

explains the relevant difference above. But both are about perceptual input. The latter may be predictions about proprioceptive input, but proprioception is still perception.

[11] Although it might be even better to not have heuristics, and instead have more fool-proof techniques, if an organism has the computational power to spare. They are heuristics afterall.

descriptive hyperpriors could be acquired through environmental input. Since the facts on which hyperpriors depend cannot be perceptually discerned, it is hard to understand how they could be acquired through environmental input. Such a line of thinking would suggest that prescriptive hyperpriors would have to be "pre-wired" into the PEM machinery, presumably by the process of evolution.[12]

PEM, as a theory of cognition, has the capacity to be neutral to questions of innate versus acquired knowledge. PEM theories have a potential to adopt an ultra-empiricism about the mind: genetic and epigenetic materials provide the instruction for neural development to result in just the PEM apparatus, but all predictive weightings come from environmental input. PEM theories also have the potential to incorporate theories of innate knowledge, however, and posit that genetic information provides instruction for neural development resulting in both the PEM apparatus, as well as some preliminary or permanent predictive weightings. It is a virtue of PEM that it leaves open the possibility for both these options.

How much we should suppose that predictive weightings have origin in genetic material versus environmental input is a matter of how much evidence we have elsewhere for thinking a particular cognitive process is (in part or in whole) innate. For example, there is a reasonably fair amount of evidence from linguistics for thinking that humans have some innate linguistic capacities—even if that capacity just involves attending to words. If PEM is to be extended to language cognition, therefore, it seems necessary to suppose that some predictive weightings for linguistic input come from genetic information. These may take the form of predictive

[12] Hohwy (2014), not specifically on this topic, and rather matter of factly, remarks that "priors are shaped through experience, development and evolution" (p. 4).

weightings that result in attentional processes to devote predictive resources to the words that people around us utter.

What we do not want to do, is suppose, just for the sake of protecting our theory of PEM against the dark room problem, that certain predictive weightings (hyperpriors) are pre-wired, without significant evidence from elsewhere. The more we do so, the more we add untestable (at the moment) assumptions to our theory. And the more we become partisan to a debate about empiricism, and the question of how much of the mind was once in the senses, without good cause.

But the body does remain in (or strive toward) homeostasis. We know this much. And the organism must accomplish this task somehow. Perhaps we have reason to believe that, whether that somehow is PEM or not, in general, genes cannot provide an organism with just the capacity to acquire (given the right environmental input) a drive for homeostasis. Instead, genetic materials must provide an organism with instructions for a full-blown drive for homeostasis.[13] If that is true, then the fact that PEM requires hyperpriors for homeostasis, which find their instruction in genetic material, just to protect itself against the dark room problem, is not a strike against it. For every theory of homeostasis would get to help itself to certain innate structures. It is then a virtue of PEM that it can provide a single framework for understanding, not just perception and action, but also this drive for homeostasis.

Notice that the dark room problem only arises in the first place when considering an organism that performs both perception and action only via PEM processes. The dark room problem does not arise for a creature that perceives with PEM, but has other strategies for action.

[13] Though perhaps the organism can learn which states are its homeostatic states.

And it also does not arise for a creature that perceives and acts with PEM, but also has non-PEM frameworks attached (such a framework for desires or other drives). Either are live options for those who take the dark room problem seriously. It is important to understand the worry the dark room poses: it is a worry only for those who seek a complete and strictly PEM theory for both perception and action.

Given the above points, even if biological creatures do not perceive, act, attend, and maintain homeostasis via the principles of PEM, such a unified framework for cognition provides Artificial Intelligence researchers with an interesting research avenue. While PEM has been put to the test as part of numerous computational models of perception, to my knowledge, no one in AI has attempted "the whole [PEM] iguana" (Dennett, 1978); no one has attempted to create an embodied agent, capable of homeostasis, perception, action, and which performs all cognitive functions with a PEM cognitive apparatus. Such a project would be a beautiful proof of concept for strong PEM.

Consider Ian Kelly's "SlugBot"—a robot designed to rove around, detect slugs, pick them up and put them in its hopper, and use bacteria to digest the slugs to produce energy to run on. Dennett notes this robotic agent as a decent attempt at the whole iguana. It would be interesting to attempt a version of the SlugBot that has a PEM mind, places predictions on its hopper being filled with slugs, and figures out how to navigate around the world so as to find and capture slugs, so as to minimize its prediction error and make its predictions about its slug filled hopper a reality. Such a project would also be interesting to see how little pre-wired "hyperpriors" such a robotic agent would require to get going. Would the PEM SlugBot only need the homeostasis hyperprior pre-wired (i.e., the prediction about its hopper being filled)? If

there were other "parent" SlugBots around who showed the "baby" SlugBot that slugs are the right things to put in the hopper, would that be enough? Such an AI project, if the agent could successfully navigate the world, would be a demonstrative reply to the dark room problem.

Before ending this section, one take away message should be noted: the question of "where the priors come from" in PEM is a difficult and messy one (unless they are random). It is perhaps one of the most unattractive features of both Bayesian and PEM theories.

5. Different Ambitious PEM Theories: Theories of Cognition, the Brain, and the Mind

The previous section looked at PEM theories as applied to action and attention, in addition to perception. But humans do much more than perceive, act, attend, and maintain homeostasis. We perform many other cognitive processes not yet discussed. Can PEM theories be extended to every domain of cognition? Can every cognitive process be understood as a process of prediction error minimization? When we do mental math, or produce a long, complex, and grammatical sentence of English, do we do so through prediction and error minimization?

Hohwy (2014, 2015) and Friston (2010, 2012) acknowledge that the answer to such a question is ultimately an empirical one, but suggest we have reasons to take seriously the possibility that the principles of PEM can be extended to all areas of cognition. Call such a position "strong PEM." Strong PEM is different from two other ambitious PEM positions, which I will call "brain PEM" and "mind PEM." In the next few pages, I will clarify these three positions and show how they are different from each other. I will also argue that brain and mind PEM depend on controversial, unknown, or simply irrelevant claims that are unrelated to PEM.

Therefore, it makes the most sense to consider what follows from strong PEM when thinking about the philosophical implications of PEM theories.

First, let us turn to brain PEM. Jakob Hohwy (2014) writes, "PEM says that prediction error minimization is the only principle for the activity of the brain" (p. 2). In his next paper, Hohwy (2015) slightly modifies his statement in a subtle but important way. Hohwy (2015) claims that not all brain activity, but all activity of the brain qua organ, is to be understood as a PEM process. In other words, it is not the case that we can understand everything that occurs in the brain as part of a PEM process, but it is the case that the brain's only biological function for the organism is to perform PEM.

This change in proposal is a nice theoretical move on Hohwy's part. The claim that all occurrences in the brain—e.g., neural behavior, all subcellular processes, etc.—can be understood using the principles of PEM is almost certainly false. Although perhaps some subcellular activity might play a role in cognition, some subcellular (and probably supercellular) activity is not going to play a role in cognition; some activity of neurons will be best explained by theories of cognition, but other activity of neurons will be best explained by general principles of microbiology. It is one of the goals of cognitive science to discover where such a line is in the brain, and understand which neural (and glial, and other) processes take part, and do not take part, in implementing cognitive functions.

That all brain activity cannot be understood using the principles of PEM does not, however, exclude the possibility that the brain is an organ with a single function or purpose. Not all activity in the heart (for example, cellular metabolism in heart tissue) can be explained as a blood pumping function, but the heart, qua organ, may still only have one function: to pump

blood. Brain PEM is the claim that the brain, qua organ, has only one function: to minimize prediction error.

I will now sketch three ways in which brain PEM could be false, while strong PEM (the claim about all cognitive processing) could be true. These points will serve to draw out the difference between these two positions. In each hypothetical case, brain PEM is false for a not particularly interesting or relevant (to cognition or PEM) reason. Such uninterestingness is intentional: it is meant to illustrate the fact that brain PEM builds in too many details about the implementation of cognitive functions, and is therefore vulnerable to many considerations that are irrelevant to thinking about the philosophical implications of specifically PEM theories. While many theoretical positions may follow from brain PEM, these claims may or may not follow as a result of the principles of PEM. Instead, these claims may follow from empirical facts about what substances implement cognitive functions, totally unrelated to PEM theories. Considering implications of brain PEM therefore ceases to be particularly philosophically instructive.

First, suppose tomorrow we were given substantial reason to think the brain is actually an organ for cooling the blood, just as Aristotle claimed. Cognitive processes are instead actually implemented inside the intestines (it only seems like the brain is involved because of complicated connections the brain has to the intestines). Such an insight into the human body, while very surprising, would not disprove PEM as a theory of perception, attention, action, etc. It would, however, show that brain PEM is false.[14]

[14] Note that one could not save brain PEM, in this circumstance, by claiming that we were just wrong about the location of the brain. The brain is an anatomically defined structure. Namely, it is the clump of neural and glial cells in our skulls. So we cannot say, in this circumstance, that it turns out the brain is in the intestines.

Second, suppose that though the brain implements the vast majority of PEM functions, it turns out that the human appendix plays a small but indispensable role in realizing part of the PEM process. I.e., the brain and appendix are jointly necessary to accomplish the function of PEM. Such a discovery would show the claim (that the brain is an organ that does PEM) to be, if not wrong, then incomplete or lacking in important ways. Brain PEM would have to be modified or extended to something like brain-appendix PEM. But once again, PEM as a theory of cognitive functions would need no modifications. This new piece of information, after all, is just a detail about the implementation of PEM.

Third, consider the fact that the brain engages in many processes of maintaining and regulating certain physiological states of our body. Suppose that of these processes, some consisted entirely in repetitive and constant neural activity that actuated muscles in another part of the body, without any feedback whatsoever. That is, suppose there are certain muscles in our body that need to be moving in the same repetitive fashion at all times, and there are neurons devoted to this cause that are more or less unconnected to any other neurons. It is not entirely clear in what sense such neural behavior could be considered cognitive. In such situations, there is no "steering" or responding on the brain's part, no need for information, propositions, or beliefs. There would just be neurons actuating muscles in a repetitive, but biologically important fashion. While the fact that neurons undergo certain non-cognitive metabolic processes is irrelevant to a description of the brain's organ function (as was discussed above), it seems that the non-cognitive processes described here do belong in a description of the brain's organ functions. And since, by stipulation, there is no feedback, there is no perceptual input, and hence no input to predict, and hence no prediction error minimization.

To be clear, this third worry is a worry about whether the brain is an organ for both cognition, and the (non-cognitive and non-PEM) maintenance of certain physiological states. In such a case, PEM as a unified framework for understanding all human cognition would again be unhindered, but brain PEM would be falsified. All cognition could be understood through PEM principles, but it would turn out that neurons serve biological purposes for the organism other than allowing it to process information about, and navigate through, its environment,[15] and therefore other than PEM.

The above paragraphs show three different ways in which Hohwy's claim—which I have been calling "brain PEM"—about the brain qua organ, could be false. First, if none of the brain turns out to be involved in PEM. Second, if more than the brain is involved in PEM. And third, if the brain, qua organ, has other non-cognitive, non-PEM functions. These may seem like trivial worries, but that is exactly the point. In each case, Hohwy's position goes wrong because of details about the implementation of PEM (as in the first two cases), or what other functions neurons implement (as in the third).

The PEM theories I have covered in the previous section are theories of perception, attention, action, and so forth. PEM theories should be understood as, first and foremost, theories of cognitive processes. They are theories of neural behavior only insofar as neurons implement cognitive processes. But when considering the philosophical implications of PEM theories, we should for the most part ignore where and how PEM is implemented, as these are further empirical questions, and we do not want to mix further empirical questions into our reasoning about the upshots of PEM as a unified theory of human cognition. I take the

[15] An organism's environment includes the organism itself, of course. But neuronal activity that simply tells muscles to move in a constant and repetitive manner, and has no feedback processes, will not count as cognition.

interesting and relevant philosophical questions regarding PEM to be about what follows from PEM as a unified theory of cognition, and not about what follows from PEM as a unified theory of cognition plus some further empirical, and contingent, facts about where and how PEM is implemented. An ambitious, interesting, and philosophically relevant theory of PEM should not hinge on the details of implementation, or whether the brain is involved in cognition at all. It should instead hinge on whether PEM can account for processes like language or mental mathematics. And so characterizing PEM as a theory of the brain, as opposed to a theory of cognition, is inaccurate.

I now turn to mind PEM, which is the claim that all psychological facts can be explained by PEM, and the way in which this position is different from strong PEM. Hohwy (2014) writes in the first couple of pages, "[This 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" (p. 2). Does the position that all cognitive processes are PEM processes entail this bold claim, as Hohwy appears to think? No, it does not; strong PEM and mind PEM are importantly different and separable claims.

On this subject, it helps to look at what sophisticated (but also bold) Bayesian cognition theorists are saying about the applicability of Bayesian principles to the human mind. Many of those who are making quite ambitious claims about the role of Bayesian inference in the mind do not attempt to explain certain psychological facts via reference to Bayesianism (Tenenbaum et al., 2008):

> For example, that a certain behavior takes people an average of 450 milliseconds to produce, measured from the onset of a visual stimulus, or that this reaction time increases when the stimulus is moved to a different part of the visual field or decreases when the same information content is presented auditorily, are not facts that a rational computational theory is likely to predict. Moreover, not all computational level models of

> cognition may have a place for Bayesian analysis. Only problems of inductive inference, or problems that contain an inductive component, are naturally expressed in Bayesian terms. Deductive reasoning, planning, or problem solving, for instance, are not traditionally thought of in this way. However, Bayesian principles are increasingly coming to be seen as relevant to many cognitive capacities, even those not traditionally seen in statistical terms (Anderson, 1990; Oaksford & Chater, 2001), due to the need for people to make inherently underconstrained inferences from impoverished data in an uncertain world. (p. 3)

Some psychological findings may not be fully explicable through some abstract principle about the mind's general structure, and need to be explained with a reference to the brute or contingent facts about the wiring of the brain (as may be the case in the example given above about a certain behavioral response taking 450 milliseconds). As Tenenbaum et al. note, Bayesian cognition applies best to those domains of cognition where a signal needs to be extracted (inferred) from noisy or ambiguous input (the clearest example being the domain of perception). Not all cognitive activity requires such inference. As they also note, however, Bayesian cognition theories are increasingly being applied to more unexpected areas of cognition. The bold claim that Bayesian cognition theorists are making, therefore, is that processes like deductive reasoning and planning are, in fact, Bayesian processes. These theorists do not seem to argue that Bayesianism should be applied to processes (of the mind, brain, or anything) that are not in any sense cognitive at all—that is, to processes not dealing with information, propositions, beliefs, or something of that nature. Nor do they argue that the principles of Bayesianism can explain every psychological fact. It is bold and interesting enough to claim that all (or a lot of unexpected) cognitive processes can be understood as operating according to Bayesian principles.

The same exact points apply to PEM theories. It is implausible that the principles of PEM can explain every psychological fact, including why a certain reaction speed is 450

milliseconds, as opposed to 500 milliseconds. It is still ambitious and interesting, and also more plausible, to instead claim that all cognitive processes are PEM processes. It is bold because in depth and evidenced PEM theories have not yet been offered for every cognitive domain. Most notably, there are not yet rich PEM accounts of deductive reasoning or language processing. But it is perfectly plausible that we will eventually have such accounts, as well as empirical evidence for them. It is less plausible that we will be able to eventually explain every single psychophysical data point with the principles of PEM.

Here is another reason PEM theories are not poised to "account for everything the mind is and does": PEM—as is—is not a theory of consciousness, and it is not clear how it can be turned into a theory of consciousness. At the very least it seems, a PEM theory of consciousness will need to be supplemented with some non-PEM concepts. It will probably be necessary to add something like a "global workspace hypothesis" (Baars, 2005) or "fame-in-the-brain" (Dennett, 2001) theory of consciousness on top of our PEM framework to get us an explanation of consciousness. It is not clear how a concept like the global workspace could be reduced to PEM concepts, nor why it should be.

Although unlikely, it is possible that the PEM framework alone is sufficient to explain consciousness. It may turn out that within the PEM framework there are principles that can explain both cognition and consciousness, just as within Maxwell's equations there are principles to explain both electricity and magnetism. Imagine that it turned out, for example, that prediction error simply is consciousness; conscious states consist of only and all the error signals that get propagated up the mental hierarchy. In such a scenario, PEM would provide a comprehensive framework for both cognition and consciousness. But that seems unlikely, and

so we should expect to have to supplement PEM with other non-PEM principles (like fame-in-the-brain) to get a theory of consciousness.[16]

I have just mentioned two types of psychological facts that the principles of PEM are not likely to explain: psychological facts that have more to do with the brute wirings of the brain; and psychological facts concerning conscious states. I will now mention a third type of psychological fact that seems to escape the grasp of PEM principles. These facts have to do with the individuation of mental states.

Consider two different brains, brain A and brain B, that are structural duplicates of each other down to the atomic level. They have the same synaptic connections, predictive tunings, and so forth. Imagine, however, that brain A is in a brain-in-a-vat situation, and that brain B is in a body-in-an-environment situation. Imagine that the machine feeding stimulation to brain A has replicated the same patterns of stimulation that brain B has enjoyed for its entire existence (which is why both have the same predictive tunings, and will retain the same predictive tunings over time).[17] Now imagine that the body of brain B has its eyes directed toward Barack Obama, and the machine is feeding brain A the same sorts of stimulations, so that both brains are having perceptions "as of" Barack Obama. Are the persons of brains A and B both enjoying the same mental state? You will get a different answer to this question depending on your theory of mental content.

---

[16] PEM principles are certainly capable of shedding light on our phenomenology, and heterophenomenological data. See Dennett (2013), who argues that PEM can provide an understanding of how we "project" many properties, such as the cuteness of babies, or the sweetness of super sugary foods, onto the world and the objects we perceive. See also Hohwy (2015), who argues that "The way inference is put together in the brain recapitulates the causes we represent in perception" (p.23). But explaining certain, many, or all (hetero)phenomenological facts does not equate to a theory of consciousness.

[17] I am using the same thought-experimental set up Uriah Kriegel (2013) uses to explain the difference between what he calls "subjective" and "objective" mental representations, but for a different purpose.

Internalists about mental content would claim that brains A and B share the same mental state. Both share a similar perception "as of" Barack Obama, and that is all that matters. Whether a brain is on Earth, Twin Earth, or is a brain-in-a-vat light years away, as long as the functional states (in this case, as we are assuming strong PEM, the PEM states) are the same, then the mental states are the same.

Externalists about mental content, on the other hand, would claim that brains A and B are in different mental states. "Tracking" theories of mental content, which include bio-semantic, teleosemantic, causal, and co-variational theories, will specifically claim that the content of brain B's mental state is Barack Obama, the content of Twin brain B's mental state is Twin Barack Obama, and the content of brain A's mental state is the state of the machine feeding it stimulation. The mental contents are different, and so the mental states are different.

The point is that brain A and brain B share the same PEM states, yet depending on different theories of mental content, they will be in different mental states. Cognitive processing can be entirely PEM (and therefore strong PEM can be true), yet mental states may supervene on facts about PEM states *and* facts about the environment outside of the PEM system. Put a bit differently, PEM states may fix the narrow content of any and all mental states, but if there is wide content, non-PEM states will in part determine this wide content. If such is the case, all psychological facts, and particularly the individuation of mental states, cannot be fully explained by PEM theories, and consequently mind PEM is false.

6. Final Thoughts: Hohwy on PEM entailing Internalism

Hohwy (2014) has argued that PEM is incompatible with externalist theories of mental content, and therefore entails a form of internalism. In fact, he believes that any theory that

claims psychological facts supervene on anything outside of states of the brain is incompatible with PEM. He believes this for the following reasons. The goal of this section is to illustrate that many of the implications that Hohwy argues follow from PEM do not follow simply from strong PEM.

First, he argues that neurons are the only sort of objects involved in implementing PEM processes (in organic creatures at least). Hohwy is not a neuro-chauvinist; he does not disallow that silicon chips could have a PEM mind. He is instead objecting to the notion that notebooks, smartphones, and other similar objects that have been proposed to play a role in our cognition (see Chalmers and Clark, 1998) play any part in PEM processes.

The claim that notebooks and smartphones help implement certain cognitive processes, and therefore can be considered a part of our minds, is known as the theory of the extended mind. The idea is that, if remembering a fact, for example, by writing it down in a notebook and looking at it later is cognitively similar to "storing" a fact in your brain and "retrieving" it later, then notebooks should be considered a part of your cognitive machinery just as much as your brain.

According to Hohwy, however, normal (i.e., non-notebook involving) memory processes are PEM processes, and notebooks do not implement PEM (how can a piece of paper compute predictions and minimize error?). Therefore using a notebook to "remember" is cognitively dissimilar to using your brain to remember, and notebooks cannot be considered a part of your cognitive machinery. Therefore, brains are likely to be the only realizers of the mind.

Next, Hohwy argues that, assuming that all cognitive processes are PEM processes, the principles of PEM supply us with a general criterion for determining the boundary between the mind and the outside world. He characterizes this criterion in terms of abductive PEM.

To review, the idea behind abductive PEM is that the mind is constantly trying to explain its own states, and judges the success of its explanations based on the accuracy of the predictions that the explanations produce. Abductive PEM therefore has three major elements involved: events in need of explanation (the states of the mind, which can also be characterized as input), hypotheses intended to explain the events (the explanations), and predictions about future states or input. Since events cannot explain themselves, the mind must hypothesize about states of affairs in the external world beyond the mind to explain its current states. Thus the mind attempts to infer the latent, or hidden causes, responsible for giving rise to the input, and in this fashion, construct a model of the world.

Hohwy argues that this architecture implies a clean division between the mind and the world, as well as a method of discerning where that division falls. That which is part of the modelling process—the input, the predictions, the explanations (which all together comprise the model)—is part of the mind. And that which is modelled—that which the mind attempts to infer through abductive PEM—is part of the world external to the mind. Put slightly differently, that which is part of the explanandum is the mind, and that which is part of the explanans is part of the world external to the mind. The mind is what is in need of explanation (and also what does the explaining), and the world is the explanation.

Hohwy is claiming that the brain is the sole realizer of our PEM processes, and that the mind is solely an (abductive) PEM framework, which is engaged in a process of representing the

world. The upshot is that, for Hohwy, neurons, and only neurons, are engaged in a process of representing the world. Therefore, he concludes, theories such as embodied cognition, which claim that there is no need for the brain to model or represent its environment (because the environment furnishes its own representations) are incompatible with abductive PEM. And similarly, any theory that claims that states of the environment outside an organism's mind contribute in any way to what mental state an organism is in is incompatible with (abductive) PEM. PEM states fix mental states alone, and therefore externalism about mental content is incompatible with PEM.

I agree with Hohwy that, if it is the case that all cognitive processes are PEM processes, and the brain is the only object involved in implementing PEM, then smartphones and notebooks are not a part of our mind in the sense that the theory of the extended mind claims. Such a conclusion does seem to follow. It is, however, unclear that neurons are the only objects involved in implementing human PEM processes. We do not know enough about how PEM is implemented to yet know what is or is not capable of participating in this process. Still, I admit that PEM has the potential to show that a lot of the claims the extended mind theorists have been making are incorrect. This point of Hohwy's is a significant one.

I disagree, however, that we can use the principles of PEM to discern the boundary of the mind and the external world. Hohwy's criterion for determining the boundary depends on PEM being an abductive process, and as I have argued toward the beginning of this paper, PEM need not be abductive. It is not at all clear how Hohwy's criterion could be re-interpreted into non-abductive PEM terms. Further, even if abductive (and strong) PEM is the case, it is not clear why PEM states should fix mental states. Even if all cognitive processes are PEM

processes, and even if such PEM processes consist of constructing explanations or models of the world, there still can be disagreement about what mental state a person is in (depending on if the PEM system is a brain-in-a-vat, on Twin Earth, etc.). Abductive PEM may very well fix the narrow content of a mental state, but these points do not show that there is not a wide content to mental states, and that this wide content is fixed by conditions of the environment outside of the mind and PEM framework.

Hohwy seems to take the excitement over the possibility of strong PEM, and leap to the position of mind PEM (the position that the entire mind can be understood with the principles of PEM). He seems to then uses this latter position to exclude the possibility of externalism. But it was argued in the previous section that the very possibility of externalism is one of the reasons that disallows us from automatically leaping to mind PEM in the first place.

In general, Hohwy appears to be bringing in many claims beyond the claims of strong PEM. And it was the purpose of this paper to try to suggest that we should be concerned mostly with what follows from strong PEM. The other claims—regarding abduction, what exactly implements PEM functions, and to what degree mental states supervene on states outside of PEM states—are all interesting questions. Much of the answers to these questions, however, depend on further empirical or philosophical questions that are unrelated to PEM. If involving these sorts of claims in a discussion of PEM, then, theorists should work hard to be clear about what follows strictly from the PEM principles, and what follows from other matters.

References

Anderson, M. L., & Chemero, T. (2013). The problem with brain GUTs: conflation of different senses of "prediction" threatens metaphysical disaster.Behavioral and Brain Sciences, 36(03), 204-205.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Progress in brain research, 150, 45-53.

Blokpoel, M., Kwisthout, J., & van Rooij, I. (2012). When can predictive brains be truly Bayesian?. *Frontiers in psychology*, *3*.

Clark, A., & Chalmers, D. (1998). The extended mind. analysis, 7-19.

Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. Mind, fzs106.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(03)

Deneve, S. (2008). Bayesian spiking neurons I: inference. *Neural computation*, *20*(1).

Dennett, D. C. (1978). Why not the whole iguana?. Behavioral and Brain Sciences, 1(01), 103-104.

Dennett, D. (2001). Are we explaining consciousness yet?. Cognition, 79(1), 221-237.

Dennett, D. C. (2013). Expecting ourselves to expect: The Bayesian brain as a projector. Behavioral and Brain Sciences, 36(03), 209-210.

Friston, K. (2005). A theory of cortical responses. Philosophical Transactions of the Royal Society of London B: Biological Sciences 360(1456)

Friston, K. (2009). The free-energy principle: A rough guide to the brain? Trends in Cognitive Sciences 13(7).

Friston, K. (2010). The free–energy principle: a unified brain theory? Nature Reviews. Neuroscience 11(2): 127–138.

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*,*62*(2).

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition.

Hohwy, J., Roepstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. Cognition 108(3).

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

Hohwy, J. (2014). The self-evidencing brain. *Noûs*.

Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt, eds. Open MIND. Frankfurt a. M., GER: MIND Group.

Kelly, I., Holland, O., & Melhuish, C. (2000, January). Slugbot: A robotic predator in the natural world. In Proceedings of the Fifth International Symposium on Artificial Life and Robotics for Human Welfare and Artificial Liferobotics (pp. 470-475).

Kriegel, U. (2013). Two notions of mental representation. Current Controversies in Philosophy of Mind. Routledge. 161-179.

Lipton, P. (2004). *Inference to the best explanation*. Psychology Press.

Mumford, D. (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biological Cybernetics 66(3):241–51.

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P. & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. Proceedings of the National Academy of Sciences USA 99(23).

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