Bachelor's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Humanities
Department of Philosophy and Religious Studies

Håkon Kjelland-Mørdre

Perceiving the World: Skepticism and the Predictive Mind

Bachelor's thesis in Philosophy
Supervisor: Anders Nes
May 2023

NTNU
Norwegian University of
Science and Technology

Håkon Kjelland-Mørdre

Perceiving the World: Skepticism and the Predictive Mind

Bachelor's thesis in Philosophy
Supervisor: Anders Nes
May 2023

Norwegian University of Science and Technology
Faculty of Humanities
Department of Philosophy and Religious Studies

NTNU
Norwegian University of
Science and Technology

Perceiving the World: Skepticism and the Predictive Mind

I. INTRODUCTION

An increasingly popular theory in neuroscience is that brains are essentially prediction machines (Friston 2010; Clark 2013; Hohwy 2013; Seth 2015; Clark 2016; Hohwy 2020). According to this theory, perception, cognition and action can all be described as the upshots of one underlying computational principle, namely *prediction error minimization*. The simple idea is that by minimizing prediction error, the brain makes sense of the world. To minimize prediction error, the brain constantly tries to explain away sensory input. This is done in a way that approximates Bayesian inference: predictions made by hypotheses stored in hierarchical generative models in the brain are tested against actual sensory input. The ensuing prediction errors are used to update the brain's hypotheses. These updated hypotheses are in turn used to generate new predictions, which are again tested against actual sensory input. And so on.

In what follows I will simply denote this theory 'PEM'. By now PEM has become a prominent theoretical framework in the philosophy of mind and cognitive science as well (Metzinger & Wiese 2017; Hohwy 2020). However, what PEM's philosophical implications are, is far from clear. In this thesis, I discuss the epistemological implications of PEM. More specifically, I discuss whether PEM entails skepticism or not. I will take skepticism to be the view that perception cannot give us justified belief about an external mind-independent world.[1]

Hohwy (2016, 2017), one of the leading proponents of PEM, has recently argued that PEM entails skepticism, saying that 'rejecting skepticism means rejecting Bayesian inference, and hence PEM' (Hohwy 2016, p. 7). PEM has been viewed as attractive, since it promises to offer an elegant and unifying theory of brain function, that would enable explanations of all aspects of mind and action. However, if Hohwy is correct, this makes PEM a less attractive theory to consider. Especially, if one does not see PEM as having the same *unique* explanatory power that Hohwy (who still thinks we should embrace PEM) and others do, why endorse it? Why not look for promising theories of perception and action (perhaps distinct) elsewhere and avoid the suggested descent into skepticism?

In this thesis, I will argue (contra Hohwy) that PEM does not entail skepticism. To argue that PEM does not entail skepticism I will proceed as follows. First, I present Hohwy's reasoning

[1] Under the (reasonable) assumption that knowledge requires justified belief, this means that if we cannot have justified beliefs about an external world, knowledge of an external world is impossible.

for why PEM entails skepticism. Then I try to make plausible a reconstruction of the argument Hohwy seems to be giving – when he reasons to the conclusion that PEM implies skepticism – by comparing Hohwy's reasoning to that of the classical (i.e., Cartesian) skeptic. This comparison motivates a reconstruction of Hohwy's argument, showing it to be structurally similar to, but importantly different from, the traditional skeptical argument – where it becomes clear that Hohwy's argument presents a challenge to all (internalist) theories of perception, not just PEM. Finally, having presented Hohwy's skeptical argument, I argue that PEM does not imply skepticism, by assuming PEM and then rejecting one of the epistemic principles figuring in Hohwy's argument. If my argument is sound, this means that one could very well hold that justified belief about the external world is possible and at the same time embrace PEM. This would make PEM a more attractive theory to consider, also for those who do not take it to have the same unique explanatory power that Hohwy and others do.

The thesis will be structured as follows. In section II, I present the prediction error minimization framework. This will be done in a way that foregrounds Hohwy's version of PEM. Section III concerns Hohwy's notion of brains as self-evidencing and makes clear why Hohwy thinks PEM entails skepticism. In section IV, I provide and try to make plausible one reconstruction of Hohwy's argument meant to establish that PEM entails skepticism. I assume the truth of PEM and then respond to Hohwy's argument in section V. Section VI concludes.

II. BRAINS AS PREDICTION MACHINES

(i) *Perception as (Bayesian) Inference*

In this section, I present PEM in some detail. This will inform the ensuing discussion regarding PEM and skepticism. PEM is motivated in part by the *problem of perception*, here conceived of as the problem of how the brain reaches conclusions about the hidden causes of sensory input. This is not an easy problem, since multiple different causes can have the same sensory effect. In other words, the problem of perception is an inverse problem: if the relationship between external hidden causes C and sensory effects E could be described by the deterministic mapping $f: C \rightarrow E$, the inverse mapping $f^{-1} : E \rightarrow C$ would usually not exist.

The basic idea of PEM is that the brain solves the inverse problem of perception by relying on prior information. That is, since the cause of a given sensory effect is underdetermined by that effect, the brain somehow has to rely on prior information to draw inferences about the hidden causes of sensory effects. Furthermore, PEM holds that it is these inferences from effects to

external causes that determine what we consciously perceive (Hohwy 2013, p. 48; Clark 2013, p. 185; Seth 2020, p. 106). Comparing PEM to Helmholtz, we realize that this response to the problem of perception is not new.[2] In the words of Helmholtz, about the 'physical activities' determining perception,

> [they] are in general not conscious, but rather unconscious. In their outcomes they are like inferences insofar as we from the observed effect on our senses arrive at an idea of the cause of this effect. This is so even though we always in fact only have direct access to the events at the nerves, that is, we sense the effects, never the external objects (Helmholtz 1867, p. 430 as quoted in Hohwy 2013, p. 17).

In sum then, PEM follows Helmholtz in making two general claims about perception: (i) percepts are the result of an unconscious inferential process, and (ii) these percepts present us with properties of external objects, although in fact our brains only ever have access to the effects of external objects.

Thus far, it has been made clear *that* perception is inferential on PEM – that it is determined by the brain inferring the hidden causes of sensory effects. However, it has not been explained *how* perception is inferential on PEM – how exactly the brain infers the hidden causes of sensory input. It is to this *how* I now turn.

A common way to explain 'the how' of PEM, is via the notion of Bayesian inference. Informally, Bayesian inference provides a rational[3] answer to the following question: how should I update my hypothesis given new evidence? Formally, this update involves computing a posterior distribution (aka the *posterior*), which is obtained from the prior distribution (aka the *prior*) multiplied with the likelihood distribution (aka the *likelihood*). Here the prior corresponds to the agent's current hypothesis and codes the information the agent already has, whereas the likelihood encodes the probability of observing the new evidence, or information, given this current hypothesis. By updating the hypothesis according to Bayes' rule the agent tend to reduce uncertainty, since the posterior often has a higher precision (lower variance) than the prior.[4] For example, as I observe your trustworthy behavior over time, my initial hypothesis

[2] This is of course not something PEM proponents deny, often citing Helmholtz as an inspiration.
[3] Rational is here understood probabilistically: it is *rational* to choose the hypothesis that achieves the highest posterior probability, given new evidence.
[4] The rule itself: $P(h|e) = \frac{P(e|h)\cdot P(h)}{P(e)}$. $P(h|e)$ is the posterior – the probability of the hypothesis h given the evidence e. $P(e|h)$ is the likelihood – the probability of the evidence given the hypothesis. $P(h)$

that you are a trustworthy person becomes more and more certain (or, equivalently, more precise).

Treating the brain as a model, which entertains several competing hypotheses about the hidden causes of sensory input, or evidence, we may begin to see how the brain could reduce uncertainty about the hidden causes of sensory input, by engaging in Bayesian inference. By continuously updating its hypotheses about the hidden causes of sensory input in light of new evidence, letting each new posterior serve as a prior for the next round of ever-changing sensory input, the brain would increase the certainty of its hypotheses (about hidden causes) over time. In so doing, the brain would create something like a 'probabilistic inverse mapping' from sensory effects to hidden causes, assigning probabilities to possible hidden causes of measured sensory effects. Here, the most likely hidden cause would come to determine perception. For example, assuming that the brain actually engages in Bayesian inference, the coffee cup in front of me simply is the most likely cause of my current sensory input.

Although Bayesian inference would provide the ideal solution to the problem of perception, it is unlikely that the brain engages in exact Bayesian inference. The reason is that Bayesian inference can be computationally expensive, or even intractable (Hohwy 2017, p. 1; Metzinger & Wiese 2017, p. 7). There is however some reason to believe that the brain can engage in approximate Bayesian inference. It is at this point prediction error minimization enters the scene. In the words of Metzinger and Wiese, 'once Bayesian inference is regarded as a solution to the problem of perception, prediction error minimization can provide a solution to the problem of computing Bayesian updates' (Metzinger & Wiese 2017, p. 7). PEM thus takes the Bayesian claim about what the brain *should be doing* and turns it into a proposal about what it actually *does*.

---

and $P(e)$ are the prior probabilities of the hypothesis and the evidence, respectively. Note that we do not need to care about the probability of the evidence, $P(e)$, when our goal is simply to find the posterior with the highest probability, given the evidence. The reason for this is that $P(e)$ will be the same for all posteriors $P(h_1|e), P(h_2|e), \dots, P(h_n|e)$, and thus cancel out in a comparison between them.

(ii) *Perception as Prediction Error Minimization*

According to PEM, the brain encodes a statistical model[5] of its environment, accumulating evidence for itself by continuously 'explaining away' sensory input. This is done in the following way: hypotheses about hidden causes of sensory input embodied in *hierarchical generative models* are generated by the model (i.e., the brain). These (multiscale) hypotheses are used to continuously predict incoming sensory signal, tracking features at different spatial and temporal scales. Predictions thus made are compared against incoming sensory signals to yield *prediction errors*. These errors are in turn used to update the hypotheses, yielding new (hopefully better) predictions. All this is done so as to minimize prediction error throughout the hierarchies of the generative models. Over time, when a particular (multiscale) hypothesis emerges as best explaining the incoming signal overall, other competing hypotheses are effectively inhibited, or crowded out, and this hypothesis comes to determine perception. More specifically, the overall prediction of this hypothesis comes to determine what we consciously perceive (see e.g. Hohwy 2013, p. 48; Seth 2020, p. 83). If all this is done in a reasonably optimal fashion, the overall prediction made by the winning hypothesis must be (barring skeptical scenarios) 'close to the real world's affairs, at least as signaled in the sensory input: nothing is a surprise, given the hypothesis, which means that everything is fully known' (Hohwy 2013, p. 45).

Let us unpack what all this means, in more detail, focusing on a particular hierarchal generative model, or hypothesis. A hierarchical generative model models how hidden causes *generate* sensory input (e.g. how cats generate *this* input and dogs *that* input) and uses this information to *generate* predictions about what the sensory input to the model (i.e., the brain) will be. The generative model thus tries to create sensory patterns for itself, with an aim to capture the causal/statistical structure of some set of observed inputs by tracking the external causes responsible for that very structure. To successfully recapitulate the causal/statistical structure of the world the generative model needs to be arranged hierarchically, with lower levels of the hierarchy tracking more fine-grained, shorter spatiotemporal regularities in the external world, and higher levels tracking more abstract, slower ones.[6] In other words, to track regularities at multiple spatial and temporal scales, each level of the hierarchy *L* itself must contain a

[5] For the remainder of this thesis I will often just describe the brain 'as a model'.
[6] Generative models can also track internal states, such as sugar levels and body temperature. This is known as *interoceptive inference*. Tracking causes external to the body is known as *exteroceptive inference* (Hohwy 2013, p. 243; Seth 2015, p. 9; Metzinger & Wiese 2017, p. 7). I will return to this below.

hypothesis, or generative model of the level $L-1$ below and try to predict the input from this level (these 'sub-hypotheses' are usually referred to as the parameters of the hierarchical generative model). In general, predictions descend downwards in the hierarchy and are tested against sensory input, giving rise to predictions errors:

$$prediction\ error\ =\ input\ -\ prediction$$

Prediction errors ascend upward in the hierarchy and carry information about the quality of the predictions made by the generative model. These are used to update the model, leading to new predictions:

$$new\ prediction\ =\ old\ prediction\ +\ prediction\ error$$

For example, under the hypothesis that a coffee cup is currently the cause of my sensory input, I predict that that the sensory input should be such and such. The predictions made by the coffee-hypothesis descend downward in the hierarchy of the generative model and are tested against the actual sensory input. The differences between the incoming sensory input and the predictions made by the coffee-hypothesis yield prediction errors, which ascend upward in the hierarchy to inform the generative model. If what I see is in fact a coffee cup, these errors should be small, such that the coffee-hypothesis would continue to determine perception, and the new predictions made by the hypothesis more or less equal to the old ones.

Thus far what has been described is *predictive coding*, where a generative model is used to predict incoming sensory input, and only the resulting prediction errors are used to update the model (Clark 2013, p. 182; Hohwy 2020, p. 210). PEM goes beyond the quite liberal framework of predictive coding (where multiple models, in theory, could be used to predict sensory input) in specifying that the update of the generative model should conform to Bayes' rule. This is achieved by putting a *weight* on prediction errors in the following way:

$$new\ prediction = old\ prediction\ +\ \frac{\pi_L}{\pi_L + \pi_P} \times \text{prediction error}$$

Before I comment on the weight, more details must be added to the PEM story. Thus far I have said that a hierarchical generative model models how hidden causes generate sensory input. Furthermore, I have said that the parameters of a hierarchical generative model themselves are generative models (of the level below). However, I have not specified in detail what these parameters or generative models actually are. They are simply a prior (normal) distribution multiplied with a likelihood (normal) distribution modeling the level below (Metzinger &

Wiese 2017, p. p. 14). Thus, the $old\ prediction$ corresponds to the mean of the likelihood before updating the (sub-) model, i.e., the expected sensory input given the prior hypothesis, whereas the $new\ prediction$ corresponds to the expected sensory input given the weight-updated hypothesis. Furthermore, π_P is the precision of the prior, encoding how certain the (sub-) model is about what is already learnt (about the level below), and π_L is the likelihood precision, encoding how much is learnt from the current input (from the level below). The weight $\frac{\pi_L}{\pi_L+\pi_P}$ thus expresses the *learning rate* of the (sub-) model, which decreases if what is already known is considered certain and increases if much is learned from the current sensory input (i.e., the learning rate decreases with the prior precision and increases with the likelihood precision). The learning rate is crucial because it enables the (sub-) model to trust the learning that has already occurred and ignore imprecise new evidence in a way that approximates Bayesian inference. Importantly, as sub-models approximate Bayesian inference, this means that hierarchical generative models approximate *hierarchical Bayesian inference*.

Perceptual inference is further complicated by the fact that the brain actually has to infer the precisions at any given level of the hierarchical generative model. This means that the brain has to engage in *precision optimization*, making learning rates vary depending on context. The brain is thus not only faced with the challenge of inferring the most likely hidden causes of sensory input but also has to engage in a form of second-order perceptual inference to determine how reliable the relevant sensory input is (Hohwy 2013, p. 65). If the brain expects sensory input to be imprecise the likelihood precision is decreased, and prior learning weighted more in perception, since the brain should not expect to learn much from imprecise input. Conversely, if sensory input is expected to be precise, then the likelihood precision is increased, and prediction errors rendered more informative. For example, walking around in Trondheim without my glasses, I vaguely perceive a red cat far away, relying heavily on prior learning ('red cats are common in Trondheim', say). However, as I put my glasses on and the incoming sensory input becomes more precise, this percept (hypothesis) as of a cat yields non-negligible prediction errors, causing me to form the percept that the red object far away in fact is a fox.

That the brain engages in precision optimization is important, because this means that perceptual inference by minimizing prediction error becomes context-sensitive, such that the brain can deal with an ever-changing environment, where sensory input comes with irreducible noise, and can be both volatile and ambiguous. In short, precision optimization is crucial to

manage the delicate dance between predictions and prediction errors in the actual world, in such a way that the brain succeeds in (perceptual) inference.

To summarize what has been said so far, PEM provides a solution to the problem of perception, where hierarchical Bayesian inference from sensory input to hidden causes is approximated by precision-weighted prediction error minimization throughout the hierarchies of generative models that actively seeks to predict sensory input, where the most successful (multiscale) hypothesis or hierarchal generative model comes to determine perception.[7] For models engaging in such prediction error minimization, generative models are expected to become increasingly accurate with time, in turn making the model (i.e., the brain) more accurate overall.

(iii) *Active Inference and Staying Alive*

I now turn to how prediction error minimization relates to the agent's quest to stay alive. As we shall see, this bears on the skeptical implications of PEM. Although PEM has been introduced as providing a solution to the problem of perception, PEM is not only or perhaps even primarily a theory of perception. That is, PEM is not specifically a theory about how a given model such as the brain can approximate Bayesian *perceptual* inference to determine the hidden causes of sensory input. Rather, PEM focuses on *all* the various ways in which models (such as the brain) can minimize prediction error and observe that such models over time will come to approximate Bayesian inference (in various ways) in the long-term average. Here, the primary role of prediction error minimization is not to infer hidden causes in the world (although this is important too), but to bring about changes in the world that help the agent stay alive. Furthermore, the target for these changes is not primarily the external but rather the internal environment of the agent, i.e., its body. The simple reason for this is that to stay alive a biological organism must maintain its internal integrity: a stable organism, that can control its internal states, can survive in multiple different and challenging environments, whereas an unstable organism may not. This point is emphasized by Anil Seth, one of the leading proponents of PEM:

[7] To be clear, the contents of perception correspond to the top-down predictions of the posterior, $P(h|e)$, of this winning generative model, and are about the hidden causes. These are thus not exactly the same predictions described in the equations above, which are the likelihoods of this generative model, $P(e|h)$, and are about the input, or rather, what the input should be given the hypothesis about the hidden causes (Recall Bayes' rule $P(h|e) = P(e|h) \cdot P(h)$; the posterior at each level of the generative model is given by multiplying the likelihood with the prior for that level).

> [PEM] may apply more naturally to interoception (the sense of the internal physiological condition of the body) than to exteroception (the classic senses, which carry signals that originate in the external environment). This is because for an organism it is more important to avoid encountering unexpected interoceptive states than to avoid encountering unexpected exteroceptive states. A level of blood oxygenation or blood sugar that is unexpected is likely to be bad news for an organism, whereas unexpected exteroceptive sensations (like novel visual inputs) are less likely to be harmful and may in some cases be desirable […] (Seth 2015, p. 9).

Further, the goal of interoceptive inference is not simply to infer the internal conditions of the body but to enable *predictive control* of vital parameters such as blood oxygenation and blood sugar. Consider the following example (due to Seth), illustrating this fact. When the brain detects a decline in blood sugar through interoceptive inference, this results in a craving for sugary things. This 'crave-percept' then leads to prediction errors

> at hierarchically-higher levels, where predictive models integrate multimodal interoceptive and exteroceptive signals. These models instantiate predictions of temporal sequences of matched exteroceptive and interoceptive inputs, which flow down through the hierarchy. The resulting cascade of prediction errors can then be resolved either through autonomic control, in order to metabolize bodily fat stores (active inference), or through allostatic actions involving the external environment (i.e., finding and eating sugary things) (Seth 2015, p. 10).

This example nicely illustrates how interoceptive inference works to keep a biological organism's (such as a human's) vital parameters within viable bounds, which involves *accurately inferring* the current state of these parameters and *actively changing* them when necessary. Interoceptive inference (or prediction error minimization) thus provides an example of how perception and action are coupled, since the hierarchical generative model responsible for the 'crave-percept' can motivate action. More generally, interoceptive inference provides an example of how *perceptual inference* relates to *active inference*, where minimizing prediction error through updating one's hypothesis about the world (or the body) corresponds to perceptual inference, and minimizing prediction error through changing one's sensory input (or internal parameters) corresponds to active inference. That is, in active inference a given

hypothesis is fixed by the model (like the 'crave-percept'), and the resulting prediction errors taken care of by sampling the world (or changing internal estimates) in such a way that the incoming input fits with the predictions made by the hypothesis.[8]

Since active inference is deeply related to the agent's quest to stay alive this means that action only makes sense when conditional on the agent's expected states, i.e., the states compatible with the continued existence of the agent (such that a fish and a human have different expected states, corresponding to environments in water and on land, respectively) (Hohwy 2020, p. 213). These are the states where the internal parameters of the organism are within viable bounds. Active inference thus comes with an element of self-fulfilling prophesying, since the organism tends to occupy the states it prophesies it will occupy, by actively seeking out or sampling those states (like we saw in the example above, where the state of viable blood sugar level was actively sought out).

Further, as the goal of prediction error minimization fundamentally is to stay alive, this means that the organism's ability to minimize prediction error is only evident over a time scale appropriate for the organism in question. In other words, prediction error minimization happens over some appropriate *long-term average*, which means that Bayesian inference is really approximated over time. This allows the agent to explore its environment, take short terms risks when necessary to ensure long-term safety, and so on. Finally, action can be undertaken for *epistemic value*, to reduce uncertainty. For example, I could move my head to confirm that my coffee cup is still at my desk, or I could have walked close enough to the vaguely perceived cat, to see that it was actually a lost fox. This still simply corresponds to minimizing prediction error, since an agent would want to be in certain, not uncertain states.

To summarize, PEM posits that the brain encodes a statistical model of its environment that minimizes prediction error in the long-term average. This is done (primarily) to keep the agent alive. Prediction error can be minimized by perceptual inference (updating the hypothesis) or by active inference (updating the input) and can be aimed at internal states (interceptive inference) or external states (exteroceptive inference). Further, as time goes prediction error is expected to decrease for the brain, as the predictions made by its hypotheses become more and

[8] PEM's notion of active inference is part of a highly original view on action, where actions (e.g. finding something to eat) are fundamentally guided by prediction errors (resulting from e.g. the 'crave-percept'), rather than intentions (e.g. then intention to get something to eat).

more accurate, and effective policies for controlling internal and external states are established.[9] In section V, I will exploit how active inference (on PEM) can bring about changes in the world (to help the agent stay alive) to argue that PEM does not imply skepticism.

III. SELF-EVIDENCING AND EVIL DEMONS

(i) *The Self-Evidencing Brain*

Having presented PEM in some detail, I now turn to its epistemological implications. More specifically, Hohwy's claim that PEM entails skepticism will be examined in this section. As Hohwy tries to make clear prediction error minimization is essentially inference to the best explanation, understood in approximate (long-term average) Bayesian terms, where the brain implicitly infers the best explanation of sensory input simply by minimizing prediction error, where this can be done through both perceptual and active inference (aimed at either internal or external states) – the hypothesis best explaining away sensory input has the highest posterior probability, and determines perception (perceptual inference) and the most effective (prediction error minimizing) policy for sampling the world to conform to a fixed percept (such as the 'crave-percept') determines action (active inference) (Hohwy 2016, p. 5 – 7).

In inference to the best explanation an interesting situation may arise. A given hypothesis, h, best explains away some evidence, e, and e in turn becomes evidence for h. In this scenario, h is said to be a *self-evidencing* explanation or hypothesis since the 'information or assumption that [e occurs] forms an indispensable part of the only available evidential support for [h]' (Hempel 1965, p. 372 – 374 as quoted in Hohwy 2016, p. 5). The pair (h, e) thus gives rise to an explanatory-evidentiary circle (i.e., an EE-circle), where h explains e and e in turn becomes

[9] A point of clarification. Note that PEM could (but need not) be considered under the *free-energy principle* (FEP). Very briefly, FEP leverages the simple truth that biological agents exist, and continue to exist (for some time) because they are able to resist the second law of thermodynamics which states that disorder (entropy) increases over time. Biological agents resist the second law by actively maintaining their boundaries with the environment (compare this to an oil drop added to a glass of water – it continues to exist, but does not actively seek out to maintain its boundaries), which is done (according to FEP) by minimizing an information-theoretic quantity called free energy in the biological organism's exchanges with the environment (for a formal introduction to FEP, see e.g. Friston 2010, for intuitive, informal introductions see e.g. Hohwy 2013, Ch. 2; Seth 2020, Ch. 10). When PEM is seen as an instance of FEP free energy is minimized by minimizing (long-term average) prediction error. Crucially, nothing that is said in Hohwy (2017) actually relies on PEM being an instance of FEP, something pointed out in Clark (2017, p. 1 – 5). Consequently, FEP is ignored here (although everything I say here is compatible with PEM being an instance of FEP). However, to make comparisons between Hohwy (2017) and this thesis easier, the reader may observe that the internal states/model of the FEP agent (= model) conceived of as a Markov blanket roughly correspond to the model (i.e., the brain), the active states correspond to the states where active output is delivered by the model, and sensory states correspond to the states where sensory input is received by the model (i.e., the states at the sensory receptors).

evidence for h. This EE-circle may look suspiciously circular but is in fact a common epistemic pattern.

To illustrate, Hohwy (2016, p. 5) provides the following example (originally from Lipton 2004, p. 24). Imagine one day looking out the window, observing footprints in the snow. You consider several hypotheses about what might explain the occurrence of this surprising evidence (i.e., the footprints). The hypothesis that a burglar is afoot turns out to be hypothesis best explaining the footprints, and so you naturally come to believe in this hypothesis. Asked to give evidence for your hypothesis that a burglar is afoot, you would now clearly be justified in citing the occurrence of the footprints. This way of reasoning is, as already mentioned, both common and unproblematic. However, as Hohwy makes clear, there is one situation in which this EE-circle turns vicious.

If someone raises doubts about the occurrence of e, then h cannot be used to deal with this doubt. The fact that e actually occurred must be established independently of the information about the occurrence of e, and in this sense h is not independent. So, for example if someone should suggest that the footprints outside your window were really produced by some hoaxers, tricking you into believing that a burglar is afoot, then you cannot simply dismiss these people by appealing to your favored burglar hypothesis, even though the footprints outside justify this hypothesis. Of course, if independent evidence against the hoaxers-hypothesis is available to you (e.g., you happen to know that there are no hoaxers in town), the hoaxer-hypothesis can be effectively dealt with, and the accompanying doubt it creates about the occurrence of the footprints put to rest. However, if no such independent evidence is available, nothing can be done to address this doubt.

Hohwy (2016, p. 6) takes this to illustrate that an EE-circle gives rise to an *evidentiary boundary* between (h, e) on the one side, and the hidden causes of e on the other. This is a boundary because causes beyond it must be inferred, and it is evidentiary because it is defined by the occurrence of the evidence. Hohwy then observes that self-evidencing and EE-circles apply to Bayesian inference in the shape of PEM as well. Consider what a prediction error minimization model (such as the brain) is doing. In perceptual inference, it actively seeks to generate hypotheses, h, about the hidden causes of sensory input, e, that best explain away incoming sensory input, simply by minimizing prediction error. In so doing, the brain approximates Bayesian inference, and maximizes evidence for itself. Furthermore, in active inference the brain too maximizes evidence for itself, by causing action in the world so as to make incoming

sensory input conform to the predictions made by its hypotheses. In other words, the (continuous) occurrence of sensory input becomes indispensable evidence for the existence of the model, both in perceptual and active inference. Thus, the brain is self-evidencing: the model simply would not exist unless the sensory input it actively sought to explain away occurred.

Importantly, Hohwy (2016, p. 6 – 7) makes clear that both forms of inference correspond to trying to infer the hidden causes of sensory input, and thus happen 'within' the evidentiary boundary induced by the self-evidencing brain 'which begins where sensory input is delivered through exteroceptive, proprioceptive and interoceptive receptors and […] ends where proprioceptive predictions are delivered, mainly in the spinal cord' (Hohwy 2016, p. 18).[10]

That the brain maximizes evidence for itself, by minimizing prediction error, is a fundamental feature of PEM. Thus, the prediction error minimizing brain must necessarily be self-evidencing.[11] However, Howhy goes further and turns the observation that the prediction error minimizing brain is self-evidencing into a proposal about why PEM entails skepticism. He points out that

> 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 (Hohwy 2016, p. 7).

From these observations Howhy concludes:

> 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 normal states of affairs we normally believe in. The Bayesian framework thus entails skepticism.

[10] As Hohwy point out in footnote 14 'the sensory input and the 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' (Hohwy 2016, p. 25; my emphasis). I mention this here to, once more, relate the concepts applied in this thesis, and in Hohwy (2016) to those applied in Hohwy (2017). That is, to relate 'PEM, evidentiary boundaries and the brain' to 'FEP, Markov blankets and the internal states/model'.

[11] The claim that PEM entails self-evidencing is uncontroversial (see e.g. Clark 2017, p. 5; Seth 2020, p. 306). However, there is disagreement about how PEM is self-evidencing, and the mutability of the evidentiary boundary (see e.g. Clark 2017; Fabry 2017).

> Consequently, rejecting skepticism entails rejecting Bayesian inference, and hence PEM (Hohwy 2016, p. 7).

Hohwy adds further important details about his skeptical position in Hohwy (2017). First, Hohwy makes explicit what the evil demon/scientist scenario figuring in the reasoning above actually is. In what I will call the *evidentiary evil demon scenario* we are fed with sensory input by an evil demon, which causes the beliefs we form to be radically false. They are false because the causes no longer are as we suppose them to be: 'the familiar people, houses, trains, trees, fruits, etc. that we perceive in everyday life' (Hohwy 2017, p. 4). Rather, sensory input is caused by an evil demon (or evil scientist) with total control of [our] sensory states' (Hohwy 2017, p. 4). In other words, it seems to be the case that 'the very same states that were assumed to represent people, houses, trains, trees, fruits, etc. are now not representing those things. They are all misrepresentations because really the hidden causes now belong to the cunning machinery and states of the evil demon' (Hohwy 2017, p. 4). Then, crucially, Hohwy makes clear that since the model (i.e., the brain) will not be able to distinguish between this non-veridical (demon) scenario and the veridical (world) scenario 'there is never any *justification* for any perceptual belief that p because the evidence for those beliefs cannot exclude the possibility that not-p' (Hohwy, 2017, p. 4; my emphasis).

Concluding this section, we see that Hohwy thinks PEM entails skepticism because an evidentiary boundary or veil exists between the brain and the world, where the brain is unable to distinguish between sensory input being caused by an evil demon or by the world as we normally think of it, causing our perceptual beliefs to be unjustified. Furthermore, it seems clear that Hohwy operates with the same definition of skepticism as laid down in this thesis. That is, Hohwy claims that PEM entails skepticism because (assuming PEM) we are not justified in our perceptual beliefs (i.e., perception cannot give us justified belief about an external mind-independent world).

IV. PEM AND JUSTIFICATION

In this section, I offer one possible reconstruction of the argument Hohwy appears to be giving, when he reasons to the conclusion that PEM implies skepticism. I proceed by first presenting a related skeptical argument, namely the (Cartesian) classical skeptical argument. This argument will, I hypothesize, help us to reconstruct Hohwy's own skeptical argument. In fact, Hohwy himself refers to the evidentiary evil demon scenario as raising 'a familiar skeptical spectre […], versions of which are familiar from much philosophy going back to Descartes'

Meditations' (Hohwy 2017, p. 4), suggesting a very close relationship between the arguments, as well as the epistemic principles figuring in them. Having presented the classical skeptical argument, I compare the reasoning motivating this argument to Hohwy's reasoning that PEM entails skepticism, as presented above. I will try to make clear that whereas the traditional skeptic presents an argument relying on epistemic principles associated with a classical skeptical scenario, and sensory veil between mind and world, suggesting that perception is metaphysically indirect, Hohwy, it seems, takes his argument to rely on epistemic principles associated with the evidentiary evil demon scenario, and evidentiary veil between brain and world, suggesting that perception is psychologically indirect. Furthermore, it seems that both Hohwy and the traditional skeptic takes the respective psychological and metaphysical indirectness to imply that arguments purporting to establish the reliability of perception are prone to vicious circularity, where this fact (allegedly) shows that we do not have good reason to think our perceptual appearances (or percepts) to be veridical.

Comparing Hohwy's reasoning to that of the traditional skeptic motivate a reconstruction of Hohwy's argument, where it becomes clear that this argument is structurally similar to, but importantly (I shall argue) different from, the classical skeptical argument. Furthermore, I will make clear that Hohwy's argument presents a challenge to all (internalist) theories of perception, not just PEM.

(i) *The Classical Skeptical Argument*

Let med now introduce the *classical skeptical argument*. The classical skeptical argument is motivated by Cartesian or *classical evil demon scenarios*. In these scenarios, things appear to us just as things normally do, but the beliefs we form are radically false. For example, a Cartesian evil demon might make us perceive as hot what we normally would perceive as cold, as black what we normally would perceive as white. Or, a brain in a vat might be stimulated in a way that causes it to have sensory appearances as of apples, trees and cats, when no such things really exist, and so on.

The evil demon scenario suggests a *sensory veil* between us and the external world, where perception must be *metaphysically indirect* – where our access to the external world is mediated by perceptual appearances. The sensory veil and the related metaphysical indirectness of perception give rise to a problem. If our only access to the external world is through perceptual appearances, we want some assurance that the appearances we are relying on are not of the misleading kind – that they are not, e.g., caused by an evil demon. Here trouble arises, since

there is no way we could have any evidence for the reliability of perception (i.e., perceptual appearances) without relying on other perceptions. To put it differently, it appears that arguments aimed at establishing the reliability of perception are predisposed to circularity, thus undermining their ability to provide a sound justification. The classical evil demon scenario thus gives rise to the following skeptical argument, which I will refer to as the classical skeptical argument (the CSA, argument due to Lyons 2016, p. 5):

(1) Nothing is ever directly present to the mind in perception except perceptual appearances (**Metaphysical Indirectness Principle**).

(2) Without a good reason for thinking perceptual appearances are veridical, we are not justified in our perceptual beliefs (**Metaevidential Principle, CSA**).

(3) We have no good reason for thinking perceptual appearances are veridical (**Reasons Claim, CSA**).

(4) Therefore, we are not justified in our perceptual beliefs.

Comments on the argument are in order. First, notice that the skeptical conclusion is entailed by (2) and (3) alone. This means that (1), which is motivated by the classical skeptical scenarios mentioned above and the associated sensory veil, is unnecessary to derive the skeptical conclusion, as are those scenarios. However, (1) is taken to render perception inferential, such that if (1) is true, then plausible (2) is. That is, if our access to the world is mediated by potentially misleading, non-veridical appearances (fed to us for example by the evil demon), then we should trust only those appearances we have good reason to think are veridical. (3) in turn derives its plausibility from the fact that the only way to verify the veridicality of appearances would itself depend on perception, in the question-begging manner described above. Furthermore, the Metaphysical Indirectness principle is, as the name suggests, a *metaphysical* principle – it says something about the nature of perceptual experience, the world, and the relation between them. The Metaevidential Principle and the Reasons Claim are *epistemic* principles – the former specifies a normative requirement for justified (perceptual) beliefs, whereas the latter denies that this requirement is met. Finally, notice that the CSA presents a prima facie challenge to *all* (metaphysical) theories of perception.

(ii) *Hohwy's Skeptical Argument*

Having presented the CSA and commented upon it, I now compare the reasoning motivating this classical argument to Hohwy's reasoning that PEM entails skepticism, as presented in

section III. I will try to show that Hohwy, like the traditional skeptic, takes his argument to rely on epistemic principles associated with an evil demon scenario, and veil, suggesting that perception is indirect, where this perceptual indirectness imply (according to Hohwy) that we do not have good reason to think our percepts to be veridical. Further, I show that Hohwy motivates these epistemic principles independently of PEM. This should encourage (and hopefully make plausible) the reconstruction of Hohwy's skeptical argument that follows.

As we saw in section III, Hohwy first established that self-evidencing gives rise to an evidentiary boundary (or evidentiary veil) between (h, e) on the one side, and the hidden causes of e on the other. Hohwy made clear that this is a boundary because causes beyond it have to be inferred, and that it is evidentiary because it is defined by the occurrence of the evidence. Hohwy then observed that self-evidencing and the corresponding evidentiary boundary apply to PEM. This, in turn, lead Hohwy to say that PEM entails skepticism because the self-evidencing brain, inferring hidden causes beyond the evidentiary boundary, will not be able to distinguish between the evidence (or sensory input) being caused by an evil demon, or by external objects as we normally think of them, (somehow) implying that 'there is never any justification for any perceptual belief that p because the evidence for those beliefs cannot exclude the possibility that not-p' (Hohwy, 2017, p. 4).

To show that Hohwy, in his reasoning to the conclusion that PEM implies skepticism, relies on epistemic principles that he motivates independently of PEM, it should suffice to make clear that Hohwy takes the evidentiary evil demon scenario to pose a challenge to all internalist theories of perception, where Hohwy takes PEM to be one such theory. According to internalist theories of mind, the mental supervenes on the neural such that e.g. cognitive, perceptual and psychological states (both conscious and unconscious) are fully determined by brain states, or computations done by the brain.[12] Especially considering internalist theories of perception, Hohwy makes clear that the 'general kind of point [concerning why PEM entails skepticism] is familiar […]: 'Given computation [e.g. prediction error minimization] 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' (Hohwy

12 It should here be made clear that internalism about the mind is not an all or nothing position, such that one could for example be an internalist about consciousness, but an externalist about cognition (i.e., cognitive states do not supervene on brain states). That being said, I will in thesis follow Hohwy in taking PEM to be an internalist theory of mind on which all mental states supervene on the neural, where these various states arise from different types of prediction error minimization done by the brain, where it trivially follows that PEM in particular is an internalist theory of perception.

2016, p. 7). Internalist theories of perception thus give rise to a *perceptual veil*[13] between brain and world. The entailed skepticism is then argued for by pointing out that the brain will not be able to distinguish between this input being caused by an evidentiary evil demon, or the world as we normally think of it, making veridical and non-veridical percepts (which arise from computations done by the ignorant brain) introspectively indistinguishable, causing our perceptual beliefs to be unjustified.

It thus becomes clear that Hohwy, like the traditional skeptic, motivates his epistemic principles by an evil demon scenario, and an associated veil, where these principles figure in an argument presenting a challenge to all internalist theories of perception, not just PEM. However, it should here be made clear that Hohwy does not, like the traditional skeptic, use the (evidentiary) evil demon scenario to make plausible an (evidentiary/perceptual) veil between the brain and the world, but rather observes that given a perceptual veil, the brain will not be able to distinguish between sensory input being caused by an evil demon, or by the world as we normally think of it. However, this difference in argumentation seems irrelevant to the respective arguments they are presenting. To see this, observe that Hohwy could very well have argued that on internalist theories of perception, the brain only ever has access to sensory input, such that the brain can't tell whether this input is caused by an evil demon or not, suggesting a perceptual veil between the brain and the world, beyond which hidden causes have to be inferred. This suggests that the epistemic principles Hohwy motivates by the evidentiary evil demon scenario and the associated evidentiary veil should be (quite) similar the Metaevidential Principle and the Reasons Claim figuring in the CSA.

Further reflection on Hohwy's reasoning from section III shows that the epistemic principles figuring in his argument indeed are similar to those figuring in the CSA. To see this, observe that the perceptual veil between brain and world suggests another reason (apart from the problem of perception) why perception must be *psychologically indirect* (Helmholtzian) – where percepts are the result of an unconscious inferential process, and where these percepts present us with properties of external objects, although in fact we only ever have access to the effects of external objects. Observe further that the perceptual veil and the related psychological indirectness give rise to a problem, similar to the problem arising for metaphysically indirect theories of perception. If the external causes (that our percepts are about) cannot be known

13 The perceptual veil arising when PEM is considered specifically as a theory of perception can in some sense be understood as a sub-veil of the more general, abstract evidentiary veil/boundary arising when active inference is taken into account as well.

directly by the brain, but have to be inferred, we want some assurance that these causes really are as we perceive them – that sensory input is not caused by an evil demon. Here trouble (once more) arises, because there is no way we could have any evidence for the reliability of perception (i.e., percepts) without obtaining further independent evidence about the occurrence of sensory input. Of course, averaging over the entire sensory input, there is no possibility of independent evidence, which (paraphrasing Hohwy 2016, p. 7) would require us to 'crawl outside our own brains' – something we cannot do. In other words, it seems that arguments purporting to establish the reliability of perception, once more, are prone to vicious circularity, albeit this time of the self-evidencing kind described in section III.[14] The evidentiary evil demon scenario and the associated perceptual (evidentiary) veil thus motivates the following reconstruction of Hohwy's skeptical which I will refer to as the evidentiary skeptical argument (the ESA)[15]:

(A) Nothing is ever directly present to the brain in perception except sensory input (**Psychological Indirectness Principle**).

(B) Percepts are the result of an unconscious inferential process happening in the brain. (**Psychological Indirectness Claim**).

(C) Without a good reason for thinking percepts are veridical, we are not justified in our perceptual beliefs (**Metaevidential Principle, ESA**).

(D) We have no good reason for thinking our percepts are veridical (**Reasons Claim, ESA**).

(E) Thus, we are not justified in our perceptual beliefs.

Comments on the argument are in order. First, notice that an additional premise (premise (B)) is included, to make clear how Hohwy's claim about our brains, relates to his claims about us. Other than that, the argument has the same structure as the classical skeptical argument (the CSA). The skeptical conclusion is, once more, entailed by (C) and (D) alone. This means that (A) would be unnecessary to derive the skeptical conclusion, as would the associated evidentiary evil demon scenario, where it not for the fact that (A) is taken to render perception psychologically indirect (Helmholtzian), such that if (A) is true, then plausibly (B) is. (C) in turn derives its plausibility from (B). That is, if we cannot know causes directly, but have to

[14] This suggests that Hohwy conceives of all internalist theories of perception as being self-evidencing in some sense.

[15] It is perhaps worth mentioning that the Psychological Indirectness Principle is really just the problem of perception, as defined in section II.

infer them, then we want some assurance that these causes (which our percepts are about) are really as we think they are. (D) is then made plausible by the fact that the only way to provide this assurance would be to obtain further independent evidence about the occurrence of our sensory input, which would be impossible (i.e., no one can crawl outside their own brain). The Psychological Indirectness Principle is (as the name suggests) a *psychological* (computational, informational) principle[16]: it tells us that causes cannot be directly accessed by the brain. In other words, it tells us that sensory information somehow has to be turned into perceptual information, or percepts (as perceptions are about causes). The Metaevidential Principle and the Reasons claim are *epistemic* principles – the former lays down a normative requirement for justified perceptual beliefs, and the latter denies that this requirement is met. Finally, observe that the ESA presents a prima facie challenge to all (internalist) theories of perception.

(iii) *Are the CSA and the ESA really distinct?*

Given that the classical skeptical argument (the CSA) and Hohwy's skeptical argument (the ESA) both rely on epistemic principles associated with an evil demon scenario and veil, suggesting that perception is metaphysically/psychologically indirect, one might come to wonder whether these arguments really are different. Before I provide a response to Hohwy's argument (the ESA), let me briefly argue that this is so. More specifically, let me give some reason why the classical and the evidentiary evil demon scenario really are different, and that metaphysical and psychological indirectness are not the same, or rather, do not describe the same perceptual relationship with the world.

To make clear that the evidentiary evil demon scenario is different from the classical evil demon scenario, I should here stress that Hohwy does not take the evil demon to possess 'Laplacian powers', i.e., the ability to fully predict the evolution of the internal states of the modeling brain. As Hohwy (2017, p. 11) makes clear the brain (as a model) is a self-organizing, noisy system, with some autonomy and individual learning history, and an ability to vary parameters in approximate Bayesian inference. This means that the demon could provide sensory input to the brain, know that the brain engages in prediction error minimization, but nonetheless (due to noise) be unable to fully control or predict what percepts would be, given this input. That is, the demon could have some best guess about what percepts would result from the input it fed to the brain, but never *know* what those percepts would be. Consequently, the fact that the brain

[16] In fact this is a metaphysical principle as well – it says something the nature of perceptual experience, the world and the relation between them. However, it is not the same metaphysical principle as the principle figuring in the CSA, which I will shortly make clear.

is a noisy system implies that the perceptual veil between brain and world (or demon) is symmetric, where the brain's states are hidden from the demon (not only the other way around). This point will be crucial in what follows. Concluding, one could not say that the evidentiary scenario (as Hohwy conceives of it) really is the same as the classical skeptical scenario, since the evil demon could control percepts or perceptual appearances by controlling the brain's sensory input.

Regarding the difference between metaphysical and psychological indirectness, it should suffice to show that these notions of indirectness do not entail each other (such that they must be different). For example, we might observe that PEM could be considered psychologically indirect (Helmholtzian), but metaphysically direct.[17] PEM could e.g. be metaphysically direct in the sense that perception of external objects is not mediated by the perception of something else – mental states that somehow represent these external objects (this is perceptual directness, see e.g. Lyons 2016, p. 10). This fact is nicely summarized by Hohwy, who makes clear that

> Perception is of course not indirect in the sense that there is an inner representation somehow emblazoned on a mental screen observed by some homunculus. That is not a very attractive conception of how representation occurs. It is indirect in the sense that what you experience now is given in your top-down prediction of your ongoing sensory input, rather than in the bottom-up signal from the states of affairs themselves (Hohwy 2013, p. 48).

Here the top-down predictions are, as made clear in section II, the predictions made by the brain's best guess of its sensory input, i.e., the predictions made by the hypothesis with the highest posterior probability. This example should suffice to show that psychological and metaphysical indirectness do not entail each other, and thus describe different perceptual relationships with the world.

Given these brief comments and clarifications, I think it is clear that the CSA and the ESA should be considered different arguments. Furthermore, it seems clear that the CSA applies more naturally to metaphysically indirect theories of perception, whereas the ESA applies more naturally to internalist (psychological, neural computational) theories of perception.

[17] For a careful treatment of the difference between metaphysical and psychological indirectness, and how this relates to PEM, see Drayson (2018).

V. PERCEIVING THE WORLD: INITIAL MOVES TOWARDS A PEM-INSPIRED EPISTEMOLOGY

In section IV I presented one possible reconstruction of Hohwy's skeptical argument (the ESA), commented on it, and distinguished it from the classical skeptical argument. In this section I assume PEM, and provide a response to the ESA, rejecting premise (C) (i.e., the Metaevidential Claim), stating that we need good reason to think our percepts to be veridical in order to have justified belief about the external world. To be clear, I will not actually try to create a PEM-inspired theory of justified belief as such. However, my argument could aid such a theory. The argument is inspired by the argument made in Chalmers (2005)[18] and observations made in Hohwy (2017, p. 8 – 13) – regarding the epistemic value of action.

(i) *Entraining the Demon*

To argue that the Metaevidential Principle ought to be rejected, first, let us follow Hohwy (2017) and carefully reconsider the evidentiary evil demon scenario in the context of PEM. As Hohwy points out, this will give us valuable insight into what the relationship between the predictive mind and the world must be like.

In the evil demon scenario, an evidentiary veil exists between the model (i.e., the brain) and that which is modeled (in this case the evil demon), where this veil is symmetric. On internalist theories of perception this veil is, as has been made clear, merely a perceptual veil. However, in the case of PEM, this veil incorporates active inference as well. Action is inference in the same fundamental way that perception is inference, namely approximating Bayes' rule through prediction error minimization. Hence, active inference cannot break down the evidentiary veil induced by the self-evidencing brain. In fact, the opposite happens to be true: the existence of the evidentiary veil is evidenced by action. But this is not to say that action cannot do 'epistemic work'.

Consider what happens during active inference. For example, my brain may favor the hypothesis that the incoming proprioceptive input is of the kind that occurs when my legs are stretched. Since my legs are not currently stretched (we assume), the prediction made by the

[18] My argument will, however, differ from Chalmers' in important respects. For example, the nature of our minds will be the same the non-demon scenario and the demon scenario(s). This follows from the assumption that PEM is an internalist theory of mind – brain states/processes constitute mental states, combined with the fact that the evil demon can only deliver sensory input, which is the same in all scenarios. In Chalmers' argument the nature of our minds is thought to depend on the scenario (though not, Chalmers argues, in a way that matters for justified belief).

currently favored hypothesis leads to a prediction error. The proprioceptive prediction, which is an active output at the evidentiary boundary, triggers reflex arcs to move the limbs around until my legs actually are stretched, the prediction is realized, and the expected proprioceptive sensory input registered by the brain (at the evidentiary boundary). This example illustrates how the brain can casually affect hidden causes beyond the evidentiary boundary with its hypothesizing, and how these hidden causes in turn affect the brain. More generally, this example captures how active inference induces a circular causality between brain and world, or in our case brain and demon (see figure below).

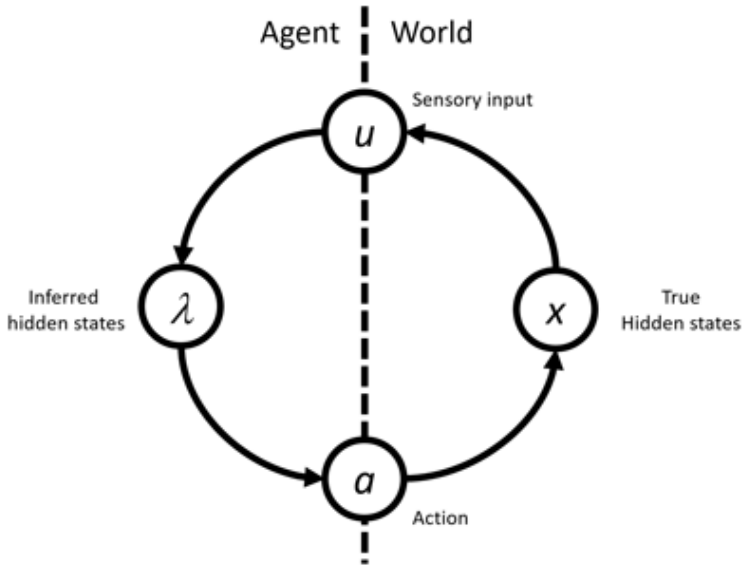


Figure 1: An illustration of the circular causality induced by active inference. The dashed line marks the evidentiary boundary induced by the self-evidencing brain (i.e., the mind-world divide). u are the sensory states (i.e., the states where sensory input is received by the model), λ are the internal states (i.e., the states of the brain, considered as a model), a are the active states (i.e., the states where active output is delivered by the model and x the true hidden causes in the world (or true hidden states of the world). The figure is taken from Hohwy (2017, p. 10).

What does this circular causality imply for our scenario? As Hohwy (2017, p. 10) makes clear not only is the demon causally entraining (to use Hohwy's words) the brain's (model's) internal states, but the brain entraining the demon's. That is, the demon causes the brain to continually update its (perceptual) hypotheses about the hidden causes of its sensory input. However, given active inference, the same might be said about the brain: through active inference, the brain causally entrains the demon so that it too has to change its states according to its sensory input, which is now the active output of the brain (which would then make the demon a model of the brain, though not necessarily a model engaged in prediction error minimization). If the demon does not follow suit – is not entrained in this manner – this would yield increasing prediction error for the brain. In the short run, this would merely make the brain a more perceptual model,

spending more time optimizing its perceptual hypotheses before acting. However, in the long run, this demon-strategy would cause the brain to perish, since increasing prediction error for the brain amounts to decreasing evidence for its existence. In other words, for as long as the brain exists, the demon would have to be entrained.

This means that in their quest to model each other (where the brain models the demon by engaging in prediction error minimization) the brain and the demon begin to 'oscillate' together – locked in a system of coupled oscillation, where the demon's input to the brain supervises the brain's perceptual hypotheses (as prediction errors), and these in turn guide active inference, yielding active output from the brain, which figure as sensory input for the demon. And so on. We see then how action levels the playing field between the brain and the demon, and causally ties them together. As Hohwy points out

> It may be that the demon started things of by giving the [brain] its basic expectations (as evolution arguably does in the non-demon case) but this is not tantamount to know a priori what active inference the [brain] will engage in since the [brain] is a self-organizing, noisy system with some autonomy, individual learning history, and ability to vary parameters in approximate Bayesian inference (Hohwy 2017, p. 11).

As such, the demon can do nothing but follow suit when the brain engages in active inference, changing the sensory input through action to fit its predictions.

Carefully considered, Hohwy's evil demon scenario reveals the demon to be more like a world and less like the classical manipulating demon completely in charge of what we perceive, found in the classical scenario. That the evil demon starts to look more and more like a world, the more information we extract from the scenario, should come as no surprise. This has to do with why the brain cannot distinguish between the evil demon scenario and the non-demon scenario in the first place. The reason the brain cannot distinguish between the two scenarios is that 'the demon world and the non-demon world must have overlapping causal/statistical structure, given the [brain's] model' (Hohwy 2017, p. 13). Otherwise, the brain (which models that structure) would not be the same in these scenarios, nor would our conscious percepts. That is, as the brain carries information about the statistical relations and causal interactions amongst the modeled hidden causes in the world, the relations and interactions would have to be the same in both scenarios.

(ii) *Perceiving the World*

That the brain carries information about the causal/statistical structure of the world, and that this structure is the same in both non-demon world and demon world, seems to suggest that our percepts should be considered veridical also in demon world. Yet, Hohwy maintains that how action entrains the environment merely 'make inroads on the evil demon scenario' (Hohwy 2017, p. 13), and thus that 'evil-demon skepticism remains' (Hohwy 2017, p. 11). It is not clear to me how Hohwy, despite his original observations, can hold that percepts are veridical in non-demon world, but not in demon world.

Three differences between non-demon world and demon world are apparent, and could cause our percepts to be veridical in non-demon world, and not in demon world. These differences are metaphysical and concern (i) the underlying, intrinsic nature giving rise to the (overlapping) causal/statistical structure in these worlds, (ii) causal regularities in these worlds that the brain does not track, and (iii) how these worlds were created.[19] I will argue that none of the aforementioned differences have any bearing on the veridicality of percepts. If I am correct this would show that our percepts are veridical both in non-demon world and in demon world, implying that a distinction between veridical and non-veridical percepts is untenable, thus undermining the Metaevidential Principle (C).

To be absolutely clear, I will in what follows assume (like Hohwy) that percepts in non-demon world are veridical. Afterall, a model like the brain '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' (Hohwy 2016, p. 8). That is, since percepts are Bayesian best guesses on the hidden causes of sensory input, where these are directly supervised by truth, I think they must be considered real, veridical and so on. Crucially, as the causal structure of non-demon and demon world are the same, where the brain receives information about this structure through sensory input, the brain is supervised by exactly the same truth in demon world as in non-demon world. This means that the differences between the worlds as listed above must be such that they (somehow) cause percepts to be non-veridical in demon world, despite the brain being supervised by the same truth in this world.

[19] That these are the only differences between non-demon world and demon world, is also something Hohwy acknowledges (see Hohwy 2017, p. 9 – 13).

Let us first consider whether a difference in intrinsic nature in non-demon and demon world could bear on the veridicality of percepts, such that this intrinsic nature makes percepts veridical in non-demon world but not in demon world. As emphasized above (and in section II) our percepts are tracking the causal/statistical structure of the external world, not the intrinsic nature giving rise to it. It is with respect to this structure that the brain seeks to minimize prediction error, and it is with respect to this structure that the brain is supervised. That is to say, percepts track things-in-themselves, by tracking the causal structure they give rise to. However, the things-in-themselves are fundamentally unknown to us – we can only know them implicitly, by knowing the causal structure they give rise to. This is of course something Hohwy acknowledges, saying that 'what makes prediction error minimization succeed is only information about the causal-statistical properties […], rather than about the intrinsic aspect of the objects themselves' (Hohwy 2017, p. 12).

It is thus unclear to me what Hohwy has in mind when he thinks intrinsic nature contributes to my percept as of a cat, say, being veridical in non-demon world, but non-veridical in demon world. In non-demon world, some external object *unknown* to my brain, X_W, gives rise to the causal structure, Y, that I come to represent as a cat. In demon world, some object *unknown* to my brain, X_D, gives rise to exactly the same structure, Y, which I once more come to represent as a cat. Furthermore, my percepts in both worlds are my brain's current best guess about the hidden causes of sensory input, are supervised by the same truth, and equally well track Y (and thus implicitly X_W/X_D). I believe these reflections show that intrinsic nature does not bear on veridicality of percepts, for how can the unknown ever serve as a standard for what is real?

This means that there must be some other difference between non-demon world and demon world that causes percepts to be veridical in non-demon world and non-veridical in demon world. Although it has been stressed that the causal structure of non-demon and demon world overlap, there could actually be some deep-seated regularities in these worlds that differ, namely those the brain does not track. However, as the brain does not track these regularities, they do not contribute to perception and, I think, should not bear on the veridicality of percepts. Compare: we do not say that a deaf person's visual percept as of a house is non-veridical because he is deaf. More generally, the veridicality of percepts should be assessed in terms of how well they track the regularities they are about, not in terms of regularities they are not about.

If the veridicality of percepts is unaffected by the intrinsic nature giving rise to the causal structure determining these percepts, and by deep-seated regularities that the brain does not track, it seems that a difference regarding how non-demon world or demon world was created must be what causes percepts to be veridical in non-demon world and not in demon world. I do not, however, think that how our world was created should affect the veridicality of percepts. That is, I think the veridicality of percepts should depend on the structure they are about, not on how that structure was created in the first place.

I think this is in tune with common sense. We do not consider how non-demon world – with its causal structure – was created to bear on the veridicality of percepts. That is, we do not normally take the veridicality of our percepts as of familiar people, houses, trains etc. to depend on whether the universe was created by a God outside the universe (i.e., created by a structure outside our structure), or by the Big Bang (i.e., created by the structure itself) etc. Likewise, we should not care if God turned out to be some 'evil' scientist (this would correspond to the BIV-scenario), or if the Big Bang turned out to be a demon instantiating the causal dance described above.

In sum, it seems that none of the three differences between non-demon world and demon world bears on the veridicality of percepts. Thus, if percepts are veridical in non-demon world, they are veridical in demon world too. In other words, whether in non-demon world or demon world, it seems that we (veridically) *perceive the world* – our world. This collapses the distinction between veridical and non-veridical percepts drawn by Hohwy and serves to (if I am correct) undermine the Metaevidential Principle stating that we need good reason to think our percepts to be veridical in order to have justified perceptual belief about the external world. Thus, it seems that we could very well have justified (perceptual) belief about the external world, while at the same time maintaining the truth of PEM.

VI. CONCLUSION

I now briefly conclude. In this thesis, I have argued that PEM does not entail skepticism. To support this claim, I began by presenting a possible reconstruction of Hohwy's argument to the conclusion that PEM entails skepticism. Subsequently, I assumed PEM and proposed an argument meant to undermine the Metaevidential Principle figuring in the ESA, by exploiting how active inference on PEM can bring about changes in the external world (to help the agent stay alive). If my argument is sound, it would suggest that justified belief about the external world is possible even while accepting PEM. That PEM does not imply skepticism would make

PEM a more attractive theory, also for those who do not take it to have the same explanatory power that Hohwy and others do.

REFERENCE LIST

Anderson, M. & Chemero, A. (2013). The problem with brain GUT's: Conflation of different senses of 'prediction' threatens metaphysical disaster. *Behavioral and Brain Sciences, 36* (3), 204–205. https://doi.org/10.1017/S0140525X1200221X

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

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

Clark, A. (2016). *Surfing Uncertainty*. Oxford University Press.

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

Drayson, Z. (2018). Direct Perception and the Predictive Mind. *Philosophical Studies, 175* (12), 3145–3164.

Fabry, R. (2017). Transcending the evidentiary boundary: Prediction error minimization, embodied interaction, and explanatory pluralism. *Philosophical Psychology, 30* (4) 395–414. https://dx.doi.org/10.1080/09515089.2016.1272674

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

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

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

Helmholtz, H. (1867). *Handbuch der Physiologishen Optik*. Leopold Voss.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press.

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Hohwy, J. (2016). The Self-Evidencing Brain. *Noûs, 50* (2), 259–285. https://doi.org/10.1111/nous.12062

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

Hohwy, J. (2020). New directions in predictive processing. *Mind and Language, 35* (2), 209–223. https://doi.org/10.1111/mila.12281

Lipton, P. (2004). *Inference to the Best Explanation*. Routledge.

Lyons, J. (2016). Epistemological Problems of Perception. In E. N. Zalta & U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy.*

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

Nagel, J. (2014). *Knowledge. A Very Short Introduction*. Oxford University Press.

O'Brien, D. (2023). The Epistemology of Perception. In J. Fieser & B. Dowden (Eds.), *Internet Encyclopedia of Philosophy.*

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers.

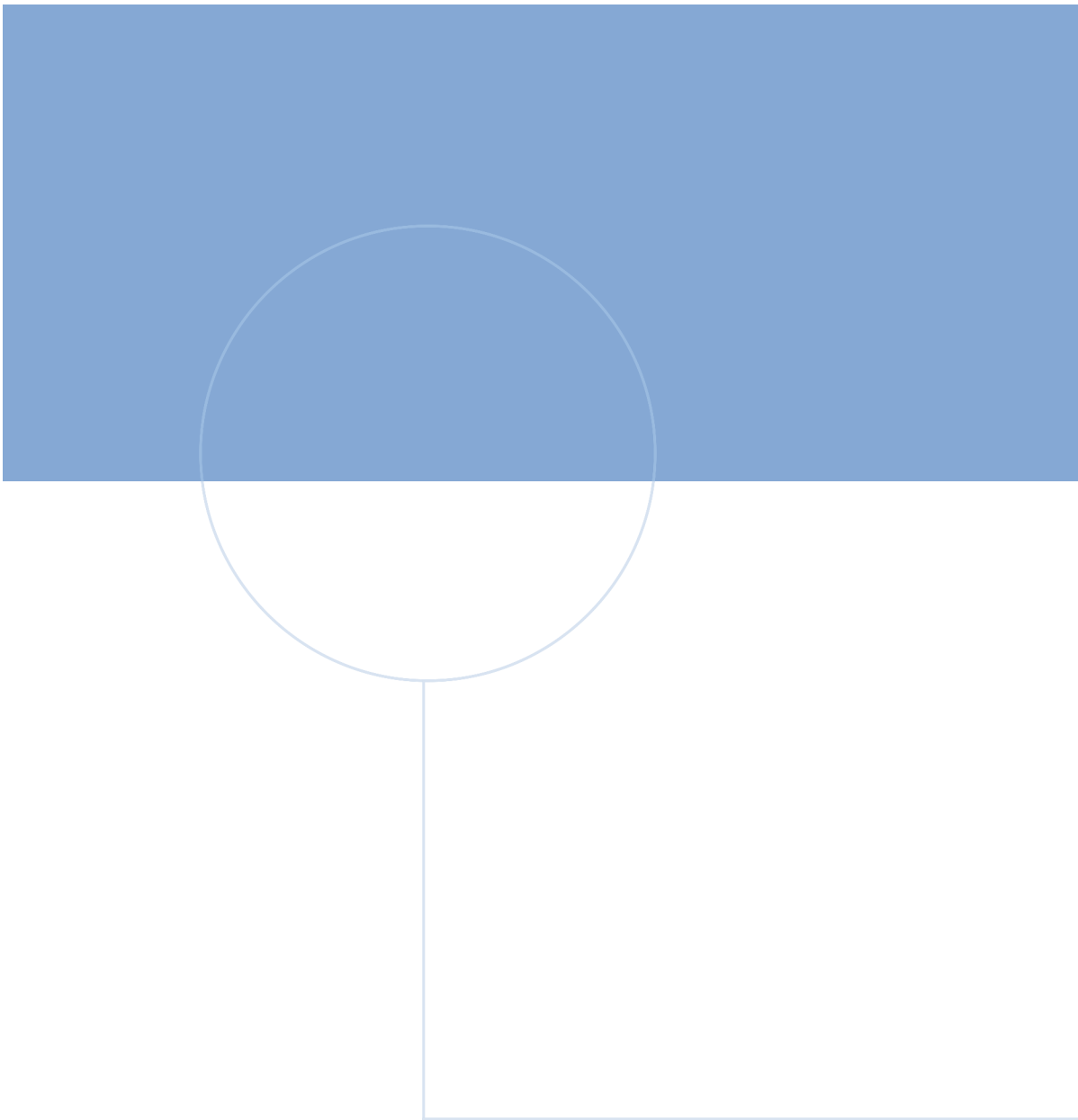
Seth, A. (2015). The Cybernetic Bayesian Brain: From Interoceptive Inference to Sensorimotor Contingencies. In T. Metzinger & J. M. Windt (Eds.), *Open Mind*. Frankfurt am Main: Mind Group. https://dx.doi.org/10.15502/9783958570108

Seth, A. (2021). *Being You*. Faber.

Shapiro, L. (2011). *Embodied Cognition*. Routledge.

Shapiro, L. & Spaulding, S. (2021). Embodied Cognition. In E. N. Zalta & U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy*.

Thompson, E. & Cosmelli, D. (2011). Brain in a Vat or Body in a World? Brainbound versus Enactive Views of Experience. *Philosophical Topics*, *39* (1), 163–180.



NTNU

Norwegian University of
Science and Technology