# Is the Brain an Organ for Prediction Error Minimization?

**Abstract.** An influential body of research in neuroscience and the philosophy of mind asserts that the brain is an organ for prediction error minimization. I clarify how this hypothesis should be understood, and I consider a prominent attempt to justify it, according to which prediction error minimization in the brain is a manifestation of a more fundamental imperative in all self-organizing systems to minimize (variational) free energy. I argue that this justification fails. The sense in which all self-organizing systems can be said to minimize free energy according to the free energy principle is fundamentally different from the alleged sense in which brains minimize prediction error. Thus, even if the free energy principle is true, it provides no support for a theory of the brain as an organ for prediction error minimization – or any other substantive theory of brain function.

**Keywords:** predictive processing; prediction error minimization; predictive coding; free energy principle

## 1. Introduction

A large number of neuroscientists and philosophers endorse the claim that the brain obeys an overarching imperative to minimize prediction error (Clark 2013; 2016; Friston 2019a; Hohwy 2013; 2016; Seth 2014). Jakob Hohwy, for example, asserts that “the brain is an organ for prediction error minimization” (2016, p.259), that “prediction error minimization is the only principle for the activity of the brain” (2016, p.260), and that “the only processing aim of the… [brain] is simply to minimise prediction error” (Hohwy et al. 2008, p.689). Similarly, Andy Clark claims that brains “are fundamentally prediction-error minimizing devices” (2017a, p.727), that “cognition… is always and everywhere a matter of minimizing prediction errors concerning the evolving flow of sensory information” (2017b, p.115), and that “perception, cognition, and action are manifestations of a single adaptive regime geared to the reduction of organism-salient prediction error” (2016, p.138). I will

henceforth refer to this hypothesis as *Prediction Error Minimization* (PEM), which as a first approximation I will state as follows:

> **Prediction Error Minimization**: The only function of the brain is to minimize prediction error.

PEM is an extraordinary claim. *Prima facie*, the brain performs a multiplicity of distinct social and ecological functions, many of which appear to have nothing to do with prediction or minimizing prediction error. Extraordinary claims require extraordinary evidence and arguments. Have proponents of PEM met this justificatory burden?

Strangely, most existing critiques of PEM do not ask this question. Instead, they pursue two different strategies. The first points to behaviours that appear to be directed at outcomes either orthogonal to or at odds with minimizing prediction error. The "dark room problem," for example, centres on the objection that an imperative to minimize prediction error mistakenly implies that agents will seek out maximally predictable environments such as dark rooms and stay there (see Friston et al. 2012). The standard responses to this objection point to the timescale over which prediction error minimization occurs and the kinds of predictions that drive the behaviour of evolved biological agents (Friston 2013a; Hohwy 2013). This leads to a second persistent criticism of PEM, however: namely, that anything can – with sufficient ingenuity – be described in terms of its contribution to minimizing long-term prediction error, which makes PEM untestable and thus unscientific (Sun and Firestone 2020).

The argument that I advance in what follows is very different from these two critiques. I agree with critics of PEM that it is likely sufficiently flexible to

accommodate any observable behaviour, but I also agree with proponents of PEM that this is not itself a reason for rejecting it (Hohwy 2015). Crucially, however, it is obviously not a reason for endorsing it, either. The important question is thus whether there are independent reasons for endorsing PEM. Proponents of PEM contend that there are. If they are right, its flexibility is irrelevant. If they are not, we have been given no reason to consider it in the first place. It is far better to evaluate these arguments directly, then. At least with respect to one prominent argument for PEM, this is the strategy that I pursue here. Of course, before such issues can be dealt with at all, we first need to know how to understand PEM. Widespread claims to the effect that prediction error minimization constitutes the brain's only *imperative, function*, *processing aim*, or *goal* are not easy to understand. What do such claims mean? How should they be evaluated?

I have two principal aims in this article. The first is to clarify PEM and thus attempt to address the foregoing questions. I take up this task in Section 2. The second is to consider a prominent attempt to justify PEM, according to which prediction error minimization in the brain is a manifestation of a more fundamental imperative in all self-organizing systems to minimize (variational) free energy (Friston 2010; 2019a; Hohwy 2013; 2015; Seth 2014). I outline this argument in Section 3, and I argue that it is unsuccessful in Section 4. Specifically, I argue that the sense in which all self-organizing systems can be said to minimize free energy according to the free energy principle is fundamentally different from the sense in which the brain is alleged to minimize prediction error according to PEM. Thus, even if the FEP is true, it provides no independent support for PEM. Indeed, it provides no independent support for *any* causal theory of brain function. I conclude in Section 5 by considering the implications of this lesson for the epistemic status of both PEM and the FEP.

## 2. Understanding Prediction Error Minimization

Before clarifying how PEM should be understood, it will be useful to provide a brief overview of predictive processing, the broader framework within which it is typically embedded. Predictive processing has been reviewed numerous times in both neuroscience and philosophy (see Clark 2016; Hohwy 2020b; Seth 2014). Thus, here I will restrict myself to just those aspects of the framework that are relevant to my arguments in what follows.

### 2.1. Predictive Processing

What does it mean to minimize prediction error? In the most basic case, at least, "prediction error" names the divergence between the sensory information generated by the body and environment and the brain's attempts to predict that sensory information from a *hierarchical probabilistic generative model* of its bodily and environmental causes. A generative model captures the process by which data are generated. A hierarchical generative model decomposes this generative process into a hierarchy, such that each successive level of the generative model represents the elements and properties responsible for generating the phenomena represented at the level below it. A probabilistic generative model captures statistical relationships that connect the elements of the modelled system and defines probability distributions or densities over its states.

According to predictive processing, the brain's attempts to minimize the error in its predictions of proximal sensory inputs both installs and updates a hierarchical probabilistic generative model of the bodily and environmental causes of those inputs. Crucially, this process of prediction error minimization is thought to be "precision-weighted," such that the degree to which the brain's predictions are

updated in light of prediction errors is constantly adjusted according to their estimated *precision* or *certainty*: the more precise predictions are (i.e. the more confidence associated with them) relative to incoming sensory information, the less they are updated in light of prediction errors, and vice versa. Precision-weighting thus plays an important role in modulating the relative influence of different bodies of information throughout the brain, and it connects prediction error minimization to an approximate form of Bayesian inference. It is widely held by proponents of predictive processing that this process of precision-weighted prediction error minimization is implemented in cortical circuitry through predictive coding, a message-passing scheme in which top-down connections in cortical hierarchies carry predictions about activities at lower levels and bottom-up connections carry information about the errors in those predictions (Friston 2005; Rao and Ballard 1999).

So far, of course, this story of brain function is highly passive. According to predictive processing, however, action also emerges from the overarching imperative to minimize prediction error, except that rather than updating predictions to bring them into alignment with incoming evidence, action involves intervening on the environment to bring sensory information into alignment with the brain's predictions, a process often described as *active inference* (Friston 2013a). Of course, for this to work an organism's goals must be encoded as predictions and some goals must be encoded as inviolable predictions. It is typically held that the most fundamental of these action-guiding predictions are endowed by evolution (Friston 2010).

This extremely minimal overview of predictive processing will suffice for my purposes in what follows. Needless to say, it neglects many important complexities, nuances, and theoretical extensions of the framework. For example, I have largely glossed over the important temporal dimension of prediction error minimization. As

proponents of predictive processing stress, however, the imperative to minimize prediction error is best understood in terms of the imperative to minimize long-term, average prediction error, which can sometimes be facilitated by short-term increases in prediction error (Hohwy 2013). For expository convenience, I will ignore this complication in what follows.

More generally, the basic framework of uncertainty-weighted prediction error minimization within hierarchical probabilistic generative models has been extended to accommodate different forms of content and structure in generative models, the difference between pragmatic and epistemic motivations, complex computations involving not just actual but expected prediction error, and more (see Friston et al. 2017; Hohwy 2020b; Parr and Friston 2019). Such elaborations demonstrate both that the overarching imperative to minimize prediction error can be decomposed into myriad constitutive functions, and that – when suitably elaborated – prediction error minimization can generate many different capacities captured in a higher-level psychological vocabulary, such as perception, learning, decision-making, planning, and so on. What unites such developments under the same framework, however, is the view that this complexity ultimately exists in the service of minimizing (long-term, average) prediction error (see Clark 2016; Hohwy 2013; 2020b). As Hohwy (2015, p.2) puts it, PEM

> "claims that the brain has *one overarching function*. There is *one thing the brain does*, which translates convincingly to the numerous other functions the brain is engaged in" (my emphasis).

### 2.2. Understanding Prediction Error Minimization

Variants of PEM are expressed in many different ways in the philosophical and scientific literature. We are told, for example, not just that prediction error

minimization constitutes the brain's overarching "function" (Hohwy 2015), but that it is the brain's sole "imperative" (Friston 2018), "processing aim" (Hohwy et al. 2008), and the thing that all brain activity is "geared to" (Clark 2016). What unites such varied articulations of PEM is the assumption that prediction error minimization in some sense constitutes the exclusive *telos* of the brain. This is the claim that I intend to focus on in what follows. Thus, I am not concerned with the claim that prediction error minimization constitutes *one* important function of the brain, or with the claim that some form of generative model-based probabilistic inference plays a central role in neural information processing. I find both such claims highly plausible (see Section 5). Both claims are much weaker than PEM, however.

Nevertheless, it is not clear how to interpret PEM. What does it mean to say that prediction error minimization constitutes the brain's only function or imperative? What is such a claim committed to? The most natural interpretation is that we should understand it in the way that we understand the attribution of functions to biological structures more generally. Regarding PEM, for example, Hohwy (2014, p.1) writes that "there is one main function of the brain, on a par with the heart's pumping of blood."

Of course, there are deep philosophical controversies concerning how to understand functional hypotheses in biology and cognitive science. The two most influential approaches are *etiological* (Wright 1973) and *systemic* (Cummins 1975) theories. Roughly, etiological theories hold that the function of a structure is the effect that explains how it came into existence, either through intentional design or through natural feedback processes such as evolution and reinforcement learning. According to systemic or causal role theories of functions, by contrast, functional hypotheses do

not explain the presence of a trait or structure. Rather, they identify the causal contribution of a component or mechanism to a capacity of a broader system.

I do not want to take a stand on which of these interpretations of functions best applies to PEM. Indeed, I want to allow for the possibility that a wholly different – perhaps *sui generis* – functional interpretation is intended. Instead, I want to focus on what both theories of functions share in common, and what is widely regarded to be a necessary feature of any such theory: functional hypotheses are subject to a *causal constraint*. In the case of etiological-evolutionary theories, for example, functional hypotheses identify the feature or effect of a structure that caused it to be selected (Wright 1973). In the case of systemic theories, functional hypotheses specify the causal contribution of a structure or mechanism to the exercise of particular capacities (Cummins 1975). In the case of PEM, then, it must be the case that prediction error minimization causally explains either why the brain evolved, or its contribution to "perception and action and everything mental in between" (Hohwy 2013, p.1). Of course, the brief overview of predictive processing outlined above shows that PEM *is* widely understood as a causal-explanatory hypothesis of this kind. That is, predictive processing views "prediction-error minimization as the *driving force* behind learning, action-selection, recognition, and inference" (Clark 2013, p.191; my emphasis). Thus, any argument or evidence for PEM must bear on the causal relevance of prediction error minimization. I will henceforth call this the *causal constraint*.

On any interpretation, PEM is extremely radical. As Hohwy (2014, p.1) acknowledges, "Most would agree that it would be controversial or even preposterous to claim that there is one main function of the brain, on a par with the heart's pumping of blood." Of course, it is possible that the apparent multiplicity of

functions and capacities produced by the brain is consistent with – indeed, explained by – an overarching function to minimize prediction error. Nevertheless, radical hypotheses require powerful evidence and arguments. In the next two sections I consider one such argument, according to which PEM receives justificatory support from the free energy principle (FEP) (Friston 2010; 2019a; Hohwy 2013; 2015; Seth 2014). As I will return to in Section 5, this is not the only justification of PEM that has been advanced, and some proponents of PEM are explicitly agnostic about the FEP (e.g. Clark 2013; 2016). Nevertheless, it is one of the most prominent justifications of PEM found in the literature, its scope and ambition are consonant with PEM's radicalism, and – as will become clear – its character as a justification is independently interesting from the perspective of epistemology, psychology, and the philosophy of science.

## 3. The Free Energy Principle and Prediction Error Minimization

The FEP developed primarily by Karl Friston (2009; 2010; Friston et al. 2006) is highly controversial, with debates raging in philosophy and the cognitive sciences concerning how to understand its justification, content, epistemic status, and implications (see Colombo and Wright 2018; Hohwy 2020b). Nevertheless, two ideas in this burgeoning literature appear to be widely shared among proponents of the FEP. The first is that the FEP is supported by a transcendental argument aimed at establishing *from first principles* the claim that all self-organizing systems obey an imperative to minimize (variational) free energy (see Friston 2009; 2010 2019). The second is that prediction error minimization in the brain can be viewed as a manifestation of this fundamental imperative of self-organization (Buckley et al. 2017; Friston 2019a; Hohwy 2020a; Seth 2014). In this section I review both ideas, postponing any evaluation of them until Section 4.

### 3.1. The Transcendental Argument

The attempt to establish the FEP from first principles is widely understood as a transcendental argument (Colombo and Wright 2018; Friston and Stephan 2007; Hohwy 2020b). Thus, Friston (2019a, p.175) writes that the argument "starts by asking fundamental questions about the necessary properties thing must possess, if they exist." In this section I outline just those aspects of the argument that are relevant for understanding and evaluating the position that I defend in Section 4.

Applied to biological systems,[1] the transcendental argument involves three core ideas that I have highlighted in Buckley et al's (2017, p.56) succinct summary:

> "**[1]** …[A]ll (viable) biological organisms resist a tendency to disorder as shown by their homoeostatic properties… **[2]** [They] must therefore minimise the occurrence of events which are atypical ('surprising') in their habitable environment… **[3]** Because the distribution of 'surprising' events is in general unknown and unknowable, organisms must instead minimise a tractable proxy, which according to the FEP turns out to be 'free energy'."

First, then, the transcendental argument begins with the tautology that a necessary condition for survival is that biological systems maintain themselves in those states consistent with their survival. Thus, for any biological system one can define a state space characterising the range of possible states it could be in, with each dimension of this space corresponding to the range of possible values that variables representing the system could take. In order to preserve its structure and organisation, a biological system must limit itself to a highly constrained subset of

[1] In more recent formulations, the FEP subsumes *all* things that persist as distinctive systems over time (Friston 2019a; 2019b; see Section 4.2.2 below).

such states, which constitute the system's attracting set (Friston 2013b). The transcendental argument thus purports to address

> "how a biological system, exposed to random and unpredictable fluctuations in its external milieu, can restrict itself to occupying a limited number of states, and therefore survive in some recognisable form" (Friston 2012, p.2100).

This first stage of the transcendental argument is often framed in terms of homeostasis and the second law of thermodynamics (Friston 2009; 2010; 2012). The second of law of thermodynamics states that closed systems tend towards a state of thermodynamic equilibrium or maximum entropy (i.e. disorder). "Homeostasis" refers to the process by which organisms exchange matter and energy with their environments to maintain their structure and organisation in the face of this tendency towards disorder. In doing so, they maintain a steady state *far from* thermodynamic equilibrium – or, equivalently, a *nonequilibrium steady state* (Friston 2013b). Because the impact of the environment is mediated through a biological system's sensory transducers, understood broadly to include all features of a system's boundary through which external states influence its internal states, the "system's 'states' can… be understood in terms of its sensations, which mediate the influence of the external world upon the system" (Hohwy 2020b, pp.3-4; see also Friston 2010; 2019a).

According to the first stage of the transcendental argument, then, biological systems must maintain themselves within those sensory states consistent with their survival. Thus, if we observe such systems when they are alive, we will find that there is a high probability that they will be in such survival-consistent states and a low

probability that they will be in states inconsistent with their survival (Friston 2010; Hohwy 2015). For any self-organizing system, one can therefore define a probability distribution over all of its possible states that captures this fact, referred to as a *nonequilibrium steady state distribution* or *density*.[2] Given this probability distribution or model, survival can now be described in terms of the avoidance of improbable states *relative to this probability distribution*. The negative logarithm of the probability a state, S, given a model or probability distribution M, P(S|M), is known in information theory as *surprisal* or *self-information*, the long-term average of which is Shannon entropy (Friston 2010). Thus, surprisal is large if the probability of the observed data given the model is low. This implies that "existence entails minimizing surprise," such that "any self-organizing system that is at nonequilibrium steady-state with its environment must minimize surprise, given a model" (Hohwy 2020b, p.4). Equivalently, minimizing surprising sensory states can be thought of as "*maximizing* the sensory evidence for the agent's existence, if we regard the agent as a model of its world" (Friston 2010, p.128; my emphasis).

Summarising the first two stages of the transcendental argument, Friston (2010, p.128) writes, "So far, all we have said is that biological agents must avoid surprises to ensure that their states remain within physiological bounds. But how do they do this?" The third stage of the transcendental argument answers this question. It involves two central claims: first, that evaluating surprisal directly is impossible; second, that systems can approximate the minimization of surprisal by minimizing

[2] I will use "distribution" to subsume both probability distributions over discrete states and density functions over continuous states throughout.

*variational free energy*, a quantity that places an upper bound on surprisal. As Friston (2010, p.128) writes,

> "A system cannot know whether its sensations are surprising and could not avoid them even if it did know. This is where free energy comes in: free energy is an upper bound on surprise, which means that if agents minimize free energy, they implicitly minimize surprise."

"Free energy" here (and henceforth) refers to *variational free energy,* an information-theoretic quantity that roughly captures the improbability of an observational conditional on a model of its causes (Friston 2010). For understanding my argument in what follows, all one strictly needs to know is that variational free energy minimization provides a computationally tractable means of minimizing surprisal. Thus, readers exclusively concerned with my argument can skip the rest of this sub-section. To get an intuition for this part of Friston's argument, however, it is useful to approach this topic from the perspective of Bayesian inference (see Buckley et al. 2017; Gershman 2019).

(**Bayes' Theorem)**

$$p(H|E) = \frac{p(H)p(E|H)}{p(E)}$$

If {E} is the set of sensory states encountered by a system and {H} is the set of possible hypotheses about its environmental causes, Bayes' theorem specifies the optimal procedure for updating the probabilities assigned to such hypotheses in light

of novel sensory states. To calculate this posterior, however, one must evaluate the denominator p(E), known as the *marginal likelihood* or *model evidence*, which is equal to a summation (for discrete states) or integration (for continuous states) over the product of the priors p(H) and likelihoods p(E|H) for all possible hypotheses. When dealing with large discrete hypothesis spaces, this is often practically infeasible. When dealing with continuous states, it can be analytically intractable. Thus, when Friston (2009, p.294) claims that biological systems cannot evaluate surprisal directly because "this would entail knowing all the hidden states of the world causing sensory input," this is because of the connection that he draws between surprisal and the (negative) model evidence in Bayesian inference, which requires knowledge of the probability of E conditional on all possible hypotheses (see Friston 2010; 2019a). This equivalence enables him to draw on the mathematics of variational inference, which provides methods for replacing exact Bayesian inference with optimization techniques for approximating its results without having to compute the model evidence directly (Bishop 2007). As Gershman (2019, p.1) puts it, "The basic idea of the FEP is to convert Bayesian inference into an optimization problem."

The technical details here are not relevant to my argument (see Buckley et al. 2017 and Gershman 2019 for an overview). The underlying intuition is relatively straightforward, however. Assume that a biological system encodes a *generative model* (or G-density) in its internal states capturing the joint probability of sensory states and environmental causes p(E, H), factored into the prior p(H) and likelihood p(E|H) distributions in Bayesian inference. Rather than trying to compute the exact Bayesian posterior directly, p(H|E), variational inference involves optimizing a *recognition model* (or R-density), q(H), in such a way as to minimize its divergence

from p(H|E). The smaller this divergence, the better the recognition model approximates the true posterior. Crucially, variational free energy is a quantity that enables a system to evaluate this divergence *without knowing the true posterior* because it is dependent on three things which a system can access: (1) data (i.e. sensory states), (2) the aforementioned generative model p(E, H), and (3) the approximate recognition model q(H) over the parameters of this generative model that it is free to optimize (Friston 2010).[3] As Buckley et al. (2017, p.57) point out, variational free energy therefore "has two functional consequences":

> "First it provides an upper bound on sensory surprisal. This allows organisms to estimate the dispersion of their constituent states and is central to the interpretation of FEP as an account of life processes. However, VFE [i.e. variational free energy] also plays a central role in a Bayesian approximation method."

The upshot is that the impossible task of minimizing surprisal directly can be replaced with the tractable task of minimizing variational free energy, which is mathematically constructed to place an upper bound on surprisal. This establishes the FEP itself: "Any self-organizing system that is at nonequilibrium steady-state with its environment must minimize its free energy" (Hohwy 2020, p.1).

---

[3] The distance between q(H) and p(H|E) is given by the Kullback-Leibler (KL) divergence:

$$D_{KL}(q(H) \,||\, p(H|E)) \;=\; \int dH \;\; q(H) \ln \frac{q(H)}{p(H|E)}$$

The denominator in the right-hand side of equation requires knowledge of p(H|E), however. Nevertheless, it is well-established in statistics that one can rewrite this equation as,

$$D_{KL}(q(H) \,||\, p(H|E)) \;= F \;+\; \ln p(E)$$

Here, F is known *as variational free energy* or (more commonly) the negative of the *evidence lower bound* (see Bishop 2007):

$$F = dH \; q(H) \ln \frac{q(H)}{p(E,H)}$$

As this equation shows, F can be computed solely from the recognition and generative models (see Buckley et al. 2017 for a review).

### 3.2. Connecting PEM to the FEP

The FEP does not say anything specifically about the brain. Nevertheless, it is widely held by proponents of the FEP and PEM that the imperative to minimize prediction error in the brain emerges "as a consequence of a more fundamental imperative towards the avoidance of "surprising" events" established by the FEP (Seth 2014, p.5). Thus, Friston (2009, p.293) claims that "the free-energy principle is an attempt to explain the structure and function of the brain, *starting from the very fact that we exist*" (my emphasis), and Hohwy (2020b, p.1) claims that the FEP can be regarded as "a grand unifying principle for cognitive science and biology."

Friston's remark about the structure and function of the brain suggests that the FEP has two interrelated implications for neuroscience. The first concerns the brain's function. As Friston (2009, p.300) puts it, the FEP provides a "mathematical specification of 'what' the brain is doing." Specifically, it implies that "*everything we do* serves to minimise surprising exchanges with the environment" (Friston and Stephan 2007, p.417; my emphasis), such that "*all* neuronal processing (and action selection) can be explained by maximizing Bayesian model evidence – or minimizing variational free energy…" (Friston et al. 2017, p.1; my emphasis).

The second implication concerns the brain's structure or causal organization. We have seen that minimizing variational free energy implicates a distinctive computational architecture involving probabilistic generative and recognition models and free energy minimization. Thus, the FEP implies that the brain and its constituent parts and operations implement this abstract computational scheme and the architecture that it involves. Specifically, it is widely held that the FEP specifies a space of so-called *process theories* that describe "concrete algorithmic

implementations of the overall computational scheme set out by FEP's use of variational Bayes, often given various assumptions" (Hohwy 2018, p.164; see Clark 2017b; Friston 2019a).

The reference to assumptions here is crucial. Although free energy minimization involves a distinctive computational architecture, this overarching architecture is consistent with multiple "different generative models, different algorithmic approximations, and different neural implementations" (Gershman 2019, p.4). PEM can be understood as a way of applying the FEP to the brain by making specific assumptions about these features. Specifically, PEM assumes that the generative model is hierarchically structured, that probability distributions are Gaussian and thus encoded by their sufficient statistics (i.e. their mean and precision), and that the approximate posterior factorizes across hidden states both within and between levels of the model, such that non-adjacent levels of the hierarchy are conditionally independent (Friston 2005; Hohwy 2020b). Further, it is associated with an evolving implementational theory in which this process of hierarchical precision-weighted prediction error minimization involves (among other things) canonical cortical microcircuits, cortical hierarchies, and the role of various neuromodulators in realising precision-weighting (Bastos et al. 2012; Friston 2005).

Stepping back, then, the FEP does not logically imply any *specific* process theory. Rather, it demarcates a space of process theories consistent with the overall computational scheme and objective function that it describes (see Allen and Friston 2016; Clark 2017b; Friston 2019a; Hohwy 2018). This nuanced relationship between PEM and the FEP is succinctly expressed by Hohwy (2020a; p.5):

> "FEP fundamentally analyses existence in terms of the probability of finding the system in certain states and the corresponding surprise of finding it in others. This

> leads to the imperative to minimise surprise, for which free energy minimisation is required. Free energy connects to PP [predictive processing] as the long-term average of prediction error (given some assumptions…)"

In this way, "PP [predictive processing] can be considered a special case of the free energy principle" (Seth 2014, p.5), with the transcendental argument *in conjunction with certain assumptions*

> "used to display the prediction error minimization strategy as itself a manifestation of a more fundamental mandate to minimize… free energy in a system's exchanges with the environment" (Clark 2016, p.305).

### 3.3. Summarising the High Road

To summarise, proponents of the FEP advance a transcendental argument aimed at establishing from first principles that all self-organizing systems must minimize variational free energy. PEM can then be understood as a way of applying this principle to neuroscience through certain assumptions about how the brain minimizes variational free energy and how this minimization is implemented in neural circuitry. Friston (2019a) calls this the "high road" justification of predictive processing. "The high road," he writes,

> "stands in for a top-down approach that starts by asking fundamental questions about the necessary properties things must possess, if they exist. Using mathematical (variational) principles, one can then show that existence is an embodied exchange of a creature with its environment – *that necessarily entails predictive processing as one aspect of a self-evidencing mechanics*" (Friston 2019a, p.175; my emphasis).

In this way the high road allegedly "takes us on a top-down journey from near existential nihilism to the riches of predictive processing" (Friston 2019a, p.175).

## 4. Evaluating the link between PEM and the FEP

In this section I argue that the high road justification described in the previous section fails as a justification of PEM. Specifically, I argue that it provides no *independent* support for PEM – or indeed *any* causal-explanatory theory of brain function. To be as explicit as possible, I *do not* argue that PEM is false or that the FEP is false. Rather, I argue that the sense in which all self-organizing systems can be said to minimize free energy according to the FEP is fundamentally different from the sense in which brains minimize prediction error according to PEM. Thus, even if one accepts the FEP, it offers no independent support for PEM. Further, appealing to auxiliary assumptions about how free energy minimization is implemented in the brain does not address this problem.

### 4.1. The FEP and the Causal Constraint

My argument is straightforward:

> P1. Any justification of PEM must bear on the causal relevance of prediction error minimization.
>
> P2. The FEP has no causal implications.
>
> C1. Therefore, the FEP does not provide any justification of PEM.

I first clarify and lay out the case for P1 and P2 in this sub-section, before considering several responses to this argument in Section 4.2.

First, any argument for PEM must satisfy what I called the "causal constraint". That is, it must provide reason to believe that prediction error minimization plays a *causal* role in brain functioning. Indeed, given the extreme scope of PEM and the transcendental argument outlined above, PEM appears to assign prediction error minimization an extremely important causal role: it is viewed as the "driving force" (Clark 2013, p.191) behind "perception and action and everything mental in between"

(Hohwy 2013, p.1), and it is ultimately responsible for the capacity of an organism to maintain a nonequilibrium steady state within its "window of viability" (Clark 2016, p.269).

Second, the FEP does not advance a causal hypothesis. Specifically, it provides *no* information about *how* self-organization is causally generated and sustained in the systems that it applies to. Instead, it provides a formal re-description of the dynamics of self-organizing systems, demonstrating that all such systems can be described *as if* they involve the minimization of variational free energy. This might be true, and it might be pregnant with profound theoretical and philosophical implications. It cannot provide any support for a causal hypothesis about how self-organization – or anything else – is achieved, however. Thus, it cannot provide any support for PEM, which constitutes such a hypothesis.

It is difficult to see how Premise 2 could be denied. First, the transcendental argument outlined above – and thus the FEP established by it – are wholly a priori. Consider the three stages of the transcendental argument outlined in Section 3: the first draws on dynamical systems theory and statistics to formalise a necessary condition for survival; the second draws on information theory to formalise what is required for the satisfaction of this condition; and the third draws on the mathematics of variational calculus to clarify how this condition can be satisfied in a way that is computationally tractable. As Hohwy (2020b, p.8) puts it, the transcendental argument thus "moves a priori – via conceptual analysis and mathematics – from existence to notions of rationality (Bayesian inference) and epistemology (self-evidencing)."

There are principled epistemological reasons for thinking that one could not derive information about how things work – specifically, the causal structure underlying the

capacities of a contingent biological system such as the human brain – from a priori reflection of this kind. Debates continue to rage in contemporary philosophy concerning the scope of a priori knowledge, with positions ranging from empiricists who deny its existence or seek to reduce it to analytic knowledge to modern rationalists who defend synthetic a priori knowledge of the sort allegedly found in mathematics and certain parts of metaphysics. Even among staunch rationalist metaphysicians, however, I am aware of no contemporary epistemological framework that would countenance a priori knowledge of contingent truths about biological mechanisms. The reason is obvious: There is no way that we could evaluate such hypotheses without observing how *our* world – out of the vast space of possible worlds that we *could* inhabit but do not – is causally structured.

Second, consider the scope of the FEP. It applies not just to all biological systems that *do* exist but to all biological systems that *could* exist. Indeed, in more recent formulations it subsumes all possible systems that conserve a boundary that distinguishes them from their environment and that preserve their structure and organisation over time, which includes inanimate objects such as rocks and drops of oil (Friston 2019a). Thus, the FEP can place no constraints on the space of possible causal mechanisms over and above those generated by existence itself. Further, when we examine the mechanisms by which existing self-organising systems persist over time, not only do we encounter enormous variation; we also encounter many whose causal structure clearly does not involve the implementation of variational inference and the representations and algorithms that this minimally implicates.

Consider simple regulatory mechanisms to which the FEP is supposed to apply, for example, such as thermostats and the Watt governor. Our knowledge of how such mechanisms work is sufficiently detailed that we can build them. In the case of the

Watt governor, for example, a simple homeostatic mechanism involving interactions among a handful of parts and operations (e.g. the angle of the spindle arms, the rotation of the flywheel, the engine output, etc.) enables it to regulate the output of steam from a steam engine (see Van Gelder 1995). [4] Nowhere in this simple mechanism is there anything resembling the implementation of variational Bayesian inference (Baltieri et al. 2020). Similarly, we easily can – and often do – build artificial intelligence systems that persist over time without implementing algorithms involving variational Bayesian inference of any kind. To say that such systems nevertheless "implicitly" encode probabilistic models (e.g. Friston 2013; 2019a) or behave "as if" they minimize variational free energy (e.g. Friston 2019a) is simply to concede this point: variational inference is not part of the causal mechanism by which they work.

In fact, when one examines the FEP, it is clear that it concerns merely how it is possible to *describe* the dynamics of systems that satisfy certain formal conditions, not how the dynamics of such systems are causally generated and sustained. This interpretation makes sense of the FEP's a priori character: Even though one cannot discover the causal structure of contingent biological structures a priori, one can derive facts about how it possible to describe the dynamics of systems that satisfy certain formal conditions a priori. Insofar as the brain satisfies such formal conditions, the FEP thus tells us that its dynamics can be described in a particular way. Similarly, this interpretation also makes the scope of the FEP intelligible: To the extent that the FEP merely specifies how it is possible to describe the dynamics of systems that satisfy certain formal conditions, the massive variation in the mechanisms by which systems generate and sustain such dynamics – and the fact

[4] Note that the set of differential equations that describe the dependencies between the Watt governor's variables in a way that abstracts away from concrete implementation also contains no probabilities (see Van Gelder 1995).

that many of these mechanisms evidently do not work by minimizing free energy – are irrelevant.

Most importantly, this interpretation receives substantial support from the literature. Thus, Friston (2012, p.2101) writes that the FEP "connects probabilistic descriptions of the states occupied by biological systems to probabilistic modelling or inference as described by Bayesian probability and information theory." Probabilistic descriptions of biological systems are an artifact of a contingent decision about how to *describe* such systems, however, and not – or at least not necessarily – a feature of biological systems themselves. Thus, any connection between such probabilistic descriptions and variational approximations to Bayesian inference does not have any logical implications for our understanding of the systems themselves. To assume that it does is to confuse properties of a possible representation of a system with properties of the system being represented, a tendency sometimes called *Pygmalion syndrome* after the mythological sculptor who fell in love with a statue (Sharvy 1985).

*Which* features of a representation map onto its target is always an open empirical question that cannot be decided a priori. Thus, even if one can *describe* self-organization in terms of a probability distribution defined over an abstract state space, there is no justification for assuming that the properties, constraints, and implications *relevant to this description* will map onto the causal structure of the self-organizing systems being described (Chater and Oaksford 2000; Colombo and Wright 2018, p.12). Specifically, the fact that one can describe the dynamics of a system in terms of free energy minimization does not imply that there is a meaningful mapping between this description and the concrete parts and operations that constitute the causal mechanism by which such dynamics are generated (Kaplan and Craver 2011). Whether this is so is always an *a posteriori* matter. Thus, claims

to the effect that the "adaptive exchange" involved in self-organization "can be *formalised* in terms of free-energy minimisation" (Friston and Stephan 2007, p.451; my emphasis) and that the FEP implies that "you will *appear* to sample your world *as if* you were trying to maximize the evidence for your own existence" (Friston 2019a, p.179; my emphasis) are irrelevant to our understanding of the causal structure of the world.

To summarise, then, any justification of PEM must satisfy a causal constraint. The FEP does not satisfy this constraint. Indeed, it has *no* implications for our understanding of how the systems that it applies to are causally structured. This fact is manifest in its epistemic status, its scope, and in its explicit commitment to the "as if" nature of its application. Thus, the FEP provides no justificatory support of any kind for PEM – or any other causal-explanatory theory of brain function.

### 4.2. Responses

In articulating the foregoing argument, I have encountered four recurring responses by proponents of PEM and the FEP. In this section I outline each response and I argue that it is unsuccessful.

#### 4.2.1. Rejecting the Causal Constraint

One might respond to my argument by abandoning the idea that PEM constitutes a causal hypothesis. That is, perhaps claims to the effect that prediction error minimization constitutes the brain's overarching function or imperative should – *contra* my claim in Section 2 above – not be understood as hypotheses about how the brain works. If true, the fact that the FEP has no causal implications would not stop it from providing justificatory support for PEM. Friston (2013a, p.212-3) seems to endorse this idea in the following passage:

> "The imperative to minimise surprise rests on the need to resist a natural tendency to disorder – to minimise sensory entropy. The Bayesian brain and predictive coding are then seen as *a consequence of*, or requirement for, this fundamental imperative – *not as a causal explanation for how our brains work*" (my emphasis).

This response is deeply unattractive, however. First, is unclear why we should care about PEM or predictive processing more generally if they do not constitute a purported causal explanation for how our brains work. Such an interpretation is certainly at odds with the extraordinary amount of excitement surrounding PEM and the growing attention paid to it within cognitive science and philosophy. Second, and more importantly, PEM *does* in fact constitute a causal hypothesis about brain function as it is typically presented in the literature (Clark 2013, p.235). Thus, my argument is that PEM *as it is typically presented* receives no justificatory support from the FEP. Of course, one can stipulate a distinct interpretation of PEM – call it PEM2 – according to which brains can merely be described *as if* they seek to minimize prediction error (see, e.g., Baltieri et al. 2020). In conjunction with certain auxiliary assumptions (e.g. about the parameterization and factorization of probability distributions), the FEP does imply PEM2. However, PEM2 is theoretically uninteresting, inconsistent with standard presentations of PEM in the literature, and irrelevant to neuroscience.

#### 4.2.2. An Alternative Interpretation of the FEP?

I have interpreted the "as if" phrase central to presentations of the FEP in the same way that this phrase is understood in the biological and social sciences more generally: namely, to indicate that although a system can be described as maximizing (or minimizing) some objective, the mechanism underlying the system's behaviour need not work *by* maximizing (or minimizing) the objective. For example,

the *individual-as-maximizing agent* principle in evolutionary biology holds that organism can be described *as if* they seek to maximize expected fitness, where "as if" indicates that calculations and representations of fitness need not play any causal role in the mechanisms underlying their behaviour (Del Giudice 2018; p.50). Similarly, rational choice models in the social sciences typically describe agents only *as if* they seek to maximize expected utility, where, again, "as if" is used to indicate that such models are silent on the causal mechanisms by which actions are generated (Chater and Oaksford 2000).

I have argued that the fact that systems can merely be described *as if* they minimize free energy should be interpreted in the same way: namely, as a purely descriptive analysis of a system's behaviour that is silent on the mechanism by which that behaviour is generated. In a recent article, however, Ramstead et al. (2020, p.17) appear to push back against this interpretation and its apparent implication that the FEP should be interpreted purely instrumentally (see Van Es 2020):

> "[A] system equipped with such a partition that exists at nonequilibrium steady state will act in a way that *looks as if* it has an intentional relation with some features of its environment. *We now know what this "as if" character amounts to*: it refers to the duality of information geometries and thereby the duality of possible descriptions (in terms of a flow towards nonequilibrium steady state and in terms of belief updating under a generative model)" (my emphasis).

The reference to "information geometries" here touches on more recent formulations of the FEP, in which Friston (2019b) has sought to demonstrate that any self-organizing system at nonequilibrium steady state equipped with a boundary (technically, a Markov blanket) that statistically separates it from its environment is amenable to two mathematically conjugate descriptions couched in the vocabulary of

information geometry. Roughly, information geometry provides a formalism for describing the distance between probability distributions in an abstract space, where each point in the space represents a possible probability distribution. According to Friston (2019b), all random dynamical systems that satisfy certain formal conditions (i.e. that possess a Markov blanket and a nonequilibrium steady state distribution) can be described in terms of both an *intrinsic* (or state-based) and *extrinsic* (or belief-based) information geometry, where the former describes the probabilistic evolution of its internal states and the latter describes the distance among probability distributions defined over external states, which are assumed to be parameterized by its internal states (see Friston et al. 2020). It is this extrinsic information geometry that can then be couched in terms variational free energy minimization (Friston 2019b).

Setting the technical details of this work aside, can it help to address the argument of this article? It is difficult to see how. A "duality of possible descriptions" still concerns possible descriptions, and this formulation of the FEP is equally vulnerable to the points made about its epistemic status and scope above. This point is effectively acknowledged by Friston et al. (2020, p.17), who note that

> "the existence of an extrinsic information geometry only means that one can map internal states to conditional probability distributions (over external states, given blanket states). *It does not mean that the resulting descriptions refer to entities that actually exist* (just as we can ascribe to a lectern the propositional belief that the best way to persist is to do nothing…)" (my emphasis).

More generally, formalism and a priori mathematical reasoning do not *in themselves* carry causal implications (see S4.2.4 below).

#### 4.2.3. The Role of Auxiliary Assumptions

The most compelling response to my argument is that it neglects the crucial role of auxiliary assumptions in connecting the FEP to PEM. Specifically, recall that PEM is not a logical implication of the FEP, but rather the FEP *in conjunction with* assumptions about how free energy is minimized (e.g. how probability distributions are parameterized and factorized) and how this process is implemented in neural circuitry. Thus, one might argue that the foregoing argument is irrelevant: Nobody believes that the FEP implies the truth of PEM anyway.

This response rests on a subtle confusion. To see this, it is crucial to distinguish between what I will call the *free energy hypothesis* about brain function from the FEP. The free energy hypothesis alleges that the brain is an organ for free energy minimization *in the sense that* free energy minimization provides a schematic mechanism sketch for how the brain works (see, e.g., Clark 2013; Friston 2005). Given this distinction, the following conditional is true: If the free energy hypothesis about brain function is true, then this hypothesis in conjunction with auxiliary assumptions implies the truth of PEM. My thesis can now be stated differently: The *FEP* – that is, the thesis that all self-organizing systems can be described *as if* they minimize free energy – provides *no support for* the free energy hypothesis. Thus, the fact that one can derive PEM from the free energy *hypothesis* in combination with certain auxiliary assumptions provides no support for PEM in itself, because we have been given no reason to endorse this hypothesis in the first place. Interestingly, this distinction is acknowledged by Friston et al. (2006, p.71) in an early article on the FEP, commenting on the evolution of work from models of free energy minimization as a purported causal explanation of perception to the all-encompassing framework of the FEP itself:

> "Previous treatments of free energy in inference (e.g., predictive coding) have been framed *as explanations or mechanistic descriptions*. In this work, we try to go a step further by suggesting that free energy minimisation is mandatory in biological systems and therefore *has a more fundamental status*" (my emphasis).

Whether or not the FEP has a more "fundamental" status than the free energy hypothesis about brain function, it certainly has a *different* epistemic status. For this reason, the FEP provides no justificatory support for the free energy hypothesis – and thus for PEM. It is worth noting that this situation is not unique. For example, the fact that an agent can be described as if it seeks to maximize expected utility provides no support for the claim that expected utility maximization forms part of the causal mechanism underlying its behaviour (Chater and Oaksford 2000).

This helps to address a related response that I often encounter in presenting the foregoing argument: namely, that my analysis of the FEP *must* be wrong because the FEP "has its roots" in explicitly causal-mechanistic work on the structure of cortical mechanisms (Friston 2005) and the development of neural networks in artificial intelligence (Dayan et al. 1995). Although the FEP draws on the same formal apparatus underlying much of this work, however, it has a fundamentally different epistemic status. Thus, the fact that there is a historical continuity and formal connection between free energy hypotheses about specific causal mechanisms and the FEP does not imply that the latter has causal implications, any more than the fact that contemporary rational choice theory has its historical roots in nineteenth and eighteenth century work that treated utility maximization as a psychological process entails that contemporary applications of rational choice models in economics should be interpreted in this way.

#### 4.2.4. The Formalism

Finally, by far the most common response that I encounter in presenting my argument is that it fails to properly engage with the complex and evolving formal apparatus surrounding the FEP. For example, my treatment has ignored the subtle statistical properties of Markov blankets, the mathematical framework of information geometry, and more generally the complex equations, equivalences, and derivations that saturate the work of Friston and colleagues. According to this final response, my failure to deal with this technical material invalidates the argument that I have advanced.

There are two ways of understanding this response. The first is roughly methodological. It states that *any* argument concerning the FEP and its epistemic status and implications must deal exhaustively with the formal apparatus surrounding the FEP, *whatever the value of that argument*. This methodological stricture on any criticism of the FEP is difficult to understand, however. My focus in this article has been on the epistemic link between the FEP and neuroscience. In advancing my argument, I have *granted* the validity and soundness of the mathematical work that underlies Friston's derivation of the FEP on the grounds that such details are irrelevant to the truth of my conclusion. It is unclear what methodological rationale could prohibit arguments of this kind if it rests on reasons independent of the soundness of such arguments.

The second and more plausible interpretation of this response is substantive. It states that some specific aspect of the formal apparatus surrounding the FEP that I have neglected undermines my argument. If true, one would have to show that a feature of this formal apparatus is capable of demonstrating that the FEP does in fact carry causal implications. For reasons outlined above, I am sceptical that *any* kind of a priori mathematical reasoning can carry substantive causal implications of this

kind. It would be foolish to rule out this possibility with any certainty, however. Thus, I welcome proponents of the FEP to identify technical features of the FEP and its derivation that carry these implications.

## 5. Conclusion

The argument of this article is simple: Given that PEM has causal implications, and the FEP does not, the FEP provides no justificatory support for PEM – or any other causal theory of brain function. Rather than elaborating on the positive implications of this argument, I will conclude by being as explicit as possible about what it does *not* imply.

First, as I have already stressed, the argument does not imply that the FEP is false or theoretically unimportant. The fact – if it is a fact – that all systems that conserve a boundary and persist over time can be described as engaging in variational Bayesian inference is fascinating and likely rich in mathematical and philosophical implications. Further, there is no doubt that the evolving formal apparatus surrounding the FEP has been extraordinarily fecund in developing models of how the brain and specific neurocognitive mechanisms work. My claim is simply that the FEP itself provides no justificatory *support* for such models. Here it might be useful to draw on the classic distinction between the psychological *context of discovery* and the epistemic *context of justification* (Reichenbach 1938). Whereas the theoretical fecundity of the FEP and the formal apparatus surrounding it attests to its fruitfulness in the construction and development of causal models of brain functioning, my claim is that the FEP itself cannot play any role in *justifying* such models.

Second, I have not argued or implied that PEM is false. As already noted, some proponents of PEM are explicitly agnostic regarding the FEP (Clark 2013; 2016), and

proponents of PEM have offered independent arguments and evidence in its defence. Here, however, it is crucial to distinguish the claim that *one* important function of the brain is prediction error minimization from the much more radical claim that the *only* function of the brain is prediction error minimization. The FEP is important precisely because its scope and ambition are consistent with this much more radical hypothesis. Once we abandon this justification of PEM, it becomes unclear what grounds there could be for assigning even a moderate degree of confidence to such a radical claim about brain function given our current state of understanding in neuroscience and psychology. Even focusing just on the domain of perception, for example, the evidence for predictive processing remains controversial (see Walsh et al. 2020).

Of course, adjudicating such complex empirical issues falls far beyond the scope of the article. I hope that this article demonstrates that this is the domain in which the truth of PEM must be adjudicated, however. *Contra* Friston (2019a, p.175), there is no alternative "high road" justification that "takes us on a top-down journey from near existential nihilism to the riches of predictive processing," at least if predictive processing is viewed as a causal explanation of how our brains work.

## REFERENCES

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695-711.

Baltieri, M., Buckley, C. L., & Bruineberg, J. (2020,). Predictions in the eye of the beholder: an active inference account of Watt governors. In *Artificial Life Conference Proceedings* (pp. 121-129).

Bishop, C. M. (2007). *Pattern recognition and machine learning*. Cordrecht: Springer.

Buckley, C. L., Kim, C. S., Mcgregor, S. and Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology,* 81, 55-79.

Chater, N., & Oaksford, M. (2000). The rational analysis of mind and behavior. *Synthese*, *122*(1-2), 93-131.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03), 181–204.

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

Clark, A. (2017a). Busting Out: Predictive Brains, Embodied Minds, and the Puzzle of the Evidentiary Veil. *Noûs,* 51, 727-753.

Clark, A. (2017b). Predictions, precision, and agentive attention. *Consciousness and cognition*, *56*, 115-119.

Colombo, M. and Wright, C. (2018). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese*, 1-26

Cummins, Robert, 1975, “Functional Analysis”, *The Journal of Philosophy*, 72(20): 741–765

Dayan, P., Hinton, G.E., Neal, R.M., (1995). The Helmholtz machine. *Neural Comput*. 7, 889–904.

Del Giudice, M. (2018). *Evolutionary psychopathology: A unified approach*. Oxford: Oxford University Press.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), 815–836.

Friston, K. (2009). The free-energy principle: a rough guide to the brain?. *Trends in cognitive sciences*, *13*(7), 293-301.

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

Friston, K. (2012). A free energy principle for biological systems. *Entropy*, *14*(11), 2100-2121.

Friston, K. (2013a). Active inference and free energy. *Behavioral and brain sciences*, *36*(3), 212.

Friston, K. (2013b). Life as we know it. *Journal of The Royal Society Interface,* 10.

Friston, K. (2018). Does predictive coding have a future?. *Nature neuroscience*, *21*(8), 1019-1021.

Friston, K. (2019b). *A free energy principle for a particular physics*. arXiv 2019, arXiv:1906.10184.

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

Friston, K. (2019a). Beyond the Desert Landscape. In Colombo, M., Irvine, E., Stapleton, M. (Ed.) *Andy Clark and His Critics*. Oxford: Oxford University Press, 174-190.

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active Inference: A Process Theory. *Neural Computation*, *29*(1), 1-49.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, *100*(1-3), 70-87.

Friston, K., Thornton, C., & Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in psychology*, *3*, 130.

Friston, K.J, Wiese, W., Hobson, J.A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy*, 22, 516.

Gershman, S. J. (2019). What does the free energy principle tell us about the brain?. *arXiv preprint arXiv:1901.07945*.

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

Hohwy, J. (2015). The Neural Organ Explains the Mind. In Metzinger, T. K. and Windt, J. M. (eds) *Open MIND.* Frankfurt am Main: MIND Group, 1-122

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, 50(2), 259–285.

Hohwy, J. (2018). Prediction error minimization in the brain. In Sprevak, M. and Colombo, M. (eds) *Handbook to the Computational Mind.* Oxford: Routledge.

Hohwy, J. (2020a). New directions in predictive processing. *Mind & Language*, *35*(2), 209-223.

Hohwy, J. (2020b). Self-supervision, normativity and the free energy principle. *Synthese*, 1-25.

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687-701.

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of science*, *78*(4), 601-627.

Parr, T., & Friston, K. J. (2019). Generalised free energy and active inference. *Biological cybernetics*, *113*(5-6), 495–513

Ramstead, M. J., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, *22*(8), 889.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a Functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2, 79–87

Reichenbach, H., (1938). *Experience and Prediction. An Analysis of the Foundations and the Structure of Knowledge*. Chicago: The University of Chicago Press.

Seth, A. K. (2014). *The cybernetic Bayesian brain*. Open MIND. Frankfurt am Main: MIND Group.

Sharvy, R. (1985). Searle on programs and intentionality. *Canadian journal of philosophy*, *15*(sup1), 39-54.

Sun, Z., & Firestone, C. (2020). The dark room problem. *Trends in Cognitive Sciences*.

van Es, T. (2020). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*,

Van Gelder, T. (1995). What Might Cognition Be, If Not Computation?. *Journal Of Philosophy*, *92*(7), 345-381.

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242.

Wright, Larry, 1973, “Functions”, *The Philosophical Review*, 82(2): 139–168.