Philos Stud (2022) 179:1693–1714
https://doi.org/10.1007/s11098-021-01722-0

Is the brain an organ for free energy minimisation?

Daniel Williams[1,2]

Accepted: 13 August 2021 / Published online: 7 September 2021

Abstract Two striking claims are advanced on behalf of the free energy principle (FEP) in cognitive science and philosophy: (i) that it identifies a condition of the possibility of existence for self-organising systems; and (ii) that it has important implications for our understanding of how the brain works, defining a set of process theories—roughly, theories of the structure and functions of neural mechanisms—consistent with the free energy minimising imperative that it derives as a necessary feature of all self-organising systems. I argue that the conjunction of claims (i) and (ii) rests on a fallacy of equivocation. The FEP can be interpreted in two ways: as a claim about how it is possible to redescribe the existence of self-organising systems (the *Descriptive FEP*), and as a claim about how such systems maintain their existence (the *Explanatory FEP*). Although the Descriptive FEP plausibly does identify a condition of the possibility of existence for self-organising systems, it has no important implications for our understanding of how the brain works. Although the Explanatory FEP would have such implications if it were true, it does not identify a condition of the possibility of existence for self-organising systems. I consider various ways of responding to this conclusion, and I explore its implications for the role and importance of the FEP in cognitive science and philosophy.

Keywords Free energy principle · Predictive processing · Predictive coding · Active inference · Process theories · Mechanism

✉ Daniel Williams
dw473@cam.ac.uk

[1] Early Career Research Fellow, Corpus Christ College, University of Cambridge, Cambridge, UK

[2] Associate Fellow, Leverhulme Centre for the Future of Intelligence, University of Cambridge, Cambridge, UK

"…the free energy principle (and the predictive processing this entails) must necessarily be in play for any person or system we care to study" (Friston, 2019a, p.184).

1 Introduction

Among recent ideas in cognitive science and philosophy, none are more simultaneously ambitious, interesting, and enigmatic than the free energy principle (FEP) developed primarily by Karl Friston (2010, 2013), according to which all self-organising systems—or, in more recent formulations, all physical systems that persist over time (Friston, 2019a)—obey an imperative to minimise variational free energy, an information-theoretic quantity that roughly scores the improbability of an observation conditional on a model of its causes (see Buckley et al., 2017; Colombo & Wright, 2018; Friston, 2010; Hohwy, 2020a).

Much of the current interest and controversy surrounding the FEP arises from two striking claims: (i) that it identifies a condition of the possibility of existence for self-organising systems, such that self-organising "systems that do not minimise free energy cannot exist" (Friston, 2013, p.2; see also Friston, 2010, 2013, 2019a, 2019b; Hohwy, 2013; 2014; 2020b); and (ii) that it has important implications for our understanding of how the brain works, providing "a unified brain theory" (Friston, 2010) and "grand unifying principle for cognitive science and biology" (Hohwy, 2020b, p.1). In conjunction with each other, such claims thus present the FEP as "an attempt to explain the structure and function of the brain, *starting from the very fact that we exist"* (Friston, 2009, p.293, my emphasis). Specifically, proponents of the FEP allege that it defines a set of "process theories"—roughly, theories of the structure and functions of neural mechanisms—consistent with the free energy minimising imperative that it derives as a necessary feature of all self-organising systems (Allen & Friston, 2018; Friston, 2019a; Hohwy, 2018, 2020a, 2020b).

My primary aim in this article is to argue that the conjunction of claims (i) and (ii) rests on a fallacy of equivocation. The FEP can be interpreted in two ways: as a claim about how it is possible to redescribe the existence of self-organising systems (the *Descriptive FEP*), and as a claim about how such systems maintain their existence (the *Explanatory FEP*). Although the Descriptive FEP plausibly does identify a condition of the possibility of existence for self-organising systems, it has no important implications for our understanding of how the brain works. Although the Explanatory FEP would have such implications if it were true, it does not identify a condition of the possibility of existence for self-organising systems. Thus, the only interpretation of the FEP on which it plausibly establishes a necessary imperative for all self-organising systems provides no reason for thinking that free energy minimisation is implemented in the brain.

My second aim is to explore the implications of this conclusion. Nothing in this article is intended to challenge the value of the FEP and the evolving formal apparatus surrounding it when it comes to generating or inspiring process theories in cognitive science. Nevertheless, once one abandons the idea that such work has a first principles *justification*, I argue that it becomes difficult to see what could motivate some of the most ambitious claims advanced on its behalf.

I structure the article as follows. Section 2 outlines the attempt to derive the FEP from first principles. Section 3 describes the FEP's alleged relationship to process theories and broader theoretical frameworks in cognitive science. Section 4 then argues that there is no interpretation of the FEP on which it establishes both a necessary imperative for all self-organising systems and the computational scheme by which the brain works, and Sect. 5 considers two objections. I conclude in Sect. 6 by summarising the article's conclusions and exploring their implications for questions about the scientific and philosophical importance of the FEP.

2 The transcendental argument for the free energy principle

It is widely held by proponents of the FEP that the imperative to minimise free energy can be derived from transcendental reflection on conditions of the possibility of existence for self-organising systems (Friston & Stephan, 2007; Friston, 2010, 2013, 2019a; Hohwy, 2020a). Thus, Friston (2019a, p.175) writes that what I will call the *transcendental argument* for the FEP "starts by asking fundamental questions about the necessary properties things must possess, if they exist" (see Colombo & Wright, 2018, p.3; Friston & Stephan, 2007; Hohwy, 2020a, p.9). Applied to self-organising *living* systems,[1] the transcendental argument involves three core stages succinctly captured in Buckley et al.'s (2017, p.56) summary:

> **[1]** …[A]ll (viable) biological organisms [must] 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'.

In this section I expand on this chain of reasoning, postponing broader philosophical questions about its epistemic status until Sects. 4 and 5. My overview is brief, non-technical, and geared only towards a basic understanding of those features of the FEP relevant to evaluating the arguments that I advance in later sections.

2.1 Nonequilibrium steady states and markov blankets

The first stage of the transcendental argument notes that living systems must maintain themselves within those states consistent with their survival if they are to survive. Drawing on dynamical systems theory, proponents of the FEP claim that one can formalise this truism in terms of a state space whose dimensions range over the possible states of a living system. As per the second law of thermodynamics, closed systems inevitably tend towards a state of thermodynamic equilibrium. By

[1] In more recent formulations the FEP is intended to apply not just to living or biological systems but to all bounded systems that preserve their structure and organisation over some time period (Friston 2019a; 2019b).

contrast, living systems are open systems that exchange matter and energy with their environments to maintain themselves in a *nonequilibrium steady state*, often described in biological terms as homeostasis (Friston, 2010, 2013).

It is implicit in this description that living systems conserve a boundary that distinguishes them from their environments. Proponents of the FEP allege that this boundary can be formalised as a Markov blanket, a set of states that constitute "a surface or boundary that defines the thing that exists (e.g., a cell membrane)" (Friston, 2019a, p.176). First introduced to capture certain conditional independencies that obtain in statistical networks, the FEP draws on Markov blankets to partition a set of states into states internal to a system {I}, states external to it {E}, and states of the system's boundary, which can in turn be decomposed into sensory states {S} directly influenced by external states but not internal states and active states {A} directly influenced by internal states but not external states (Friston, 2013). Those states necessary for maintaining survival within a given environment can then allegedly be understood in terms of a subset of sensory states, "which mediate the influence of the external world upon the system" (Hohwy, 2020b, pp.3–4; see also Friston, 2010, 2019a, 2019b).

2.2 Survival as surprisal avoidance

The second stage of the transcendental argument notes that if we observe living systems when they are alive, we will therefore find that there is a high probability that they will be in survival-consistent sensory states and a low probability that they will be in states inconsistent with their survival (Friston, 2010; Hohwy 2016). Thus, for any living system, one can define a probability distribution over its possible states that captures this fact, grounded in the probability of finding the system in different states when sampled at random (Friston, 2013).[2] Relative to this probability distribution, survival can then be understood in terms of the avoidance of improbable states. 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 (Friston, 2010). Thus, surprisal is large if the probability of the observed data given the model is low, which 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, minimising surprising sensory states can be thought of as "maximising the sensory evidence for the agent's existence, if we regard the agent as a model of its world" (Friston, 2010, p.128), a process often referred to as "self-evidencing" (Hohwy, 2016).

2.3 Avoiding surprises by minimising free energy

The third stage of the transcendental argument involves two central claims: that evaluating surprisal directly is impossible, and that systems can nevertheless

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

approximate the minimisation of surprisal by minimising 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 minimise free energy, they implicitly minimise surprise."

"Free energy" here (and henceforth) refers to variational free energy, an information-theoretic quantity that roughly scores the improbability of an observational conditional on a model of its causes (Friston, 2010).

To get an intuition for how this works, it is useful to approach it from the perspective of Bayesian inference. Specifically, its driving assumption is "that the log probability of being in a particular (sensory) state is the marginal likelihood or Bayesian model evidence of that state" (Friston, 2019a, p.177). The marginal likelihood is the denominator in Bayes' theorem, which can be interpreted as a method for inferring the probability of external states from the sensory states that living systems have access to:

(Bayes' Theorem)

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

However, calculating this denominator p(S) requires calculating the sum of the product of the priors p(E) and likelihoods p(S|E) for all possible environmental states, $\sum_E p(S|E)p(E)$ (for continuous probability distributions, the summation is replaced with the integral, $\int p(S|E)p(E)dE$). This calculation is often intractable. Thus, Friston's (2009, p.294) claim that living systems cannot evaluate surprisal directly because "this would entail knowing all the hidden states of the world causing sensory input" rests on the connection that he draws between sensory surprisal and (the negative of) this marginal likelihood, p(S). Crucially, this connection then enables him to draw on methods of variational inference established in physics and machine learning for replacing exact Bayesian inference with approximation methods that rest on optimisation (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. The underlying intuition is relatively straightforward, however. Assume that a living system somehow encodes a generative model (or G-density) capturing the joint probability of sensory states {S} and hidden environmental causes {E}, p(S, E), factored into the prior p(E) and likelihood p(S|E) distributions in Bayesian inference. Rather than trying to compute the exact Bayesian posterior directly, p(E|S), variational inference involves optimising a recognition model (or R-density), q(E), in such a way as to minimise its divergence from p(E|S). Crucially, variational free energy is a quantity that enables a system to evaluate this divergence without access to the true posterior because it depends on three things that the system allegedly can access: (1) data (i.e. sensory states), (2) the aforementioned generative model p(S, E), and (3) the approximate recognition model q(E) over the parameters of this generative model that it is free to optimise (Buckley et al., 2017; Friston, 2013; Hohwy, 2020a).[3] Further, because variational free energy is mathematically constructed to place an upper bound on surprisal, minimising variational free energy provides a computationally tractable means of approximating the minimisation of surprisal. Thus, the claim that living systems must minimise surprisal can be replaced with the FEP itself, according to which "*any 'thing' that attains nonequilibrium steady-state can be construed as performing an elemental sort of Bayesian inference*," namely, *variational free energy minimisation* (Friston et al., 2020, p.2; see also Friston, 2010, 2013, 2019a, 2019b).

3 From the FEP to predictive processing

The FEP does not say anything specifically about the brain. Nevertheless, it is widely held that it has important implications for neuroscience. Specifically, Friston's (2009, p.293) remark 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," highlights two such implications. The first concerns the brain's *function*, *objective*, *processing aim*, or *imperative* (Friston, 2010, 2013, 2019a; Hohwy, 2013).[4] 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 & Stephan, 2007, p.417), such that "all neuronal processing (and action selection) can be explained by… minimising variational free energy" (Friston et al., 2017, p.1).

The second implication concerns the brain's structure. Minimising variational free energy implicates a distinctive computational scheme involving probabilistic models and variational Bayesian inference. Applied to the brain, the FEP is thus taken to imply that neural mechanisms implement this scheme (Friston, 2009, 2010). Specifically, it is taken to define a set of 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 Allen & Friston, 2018; Clark, 2017; Friston, 2019a). Process theories are

[3] Technically, the distance between q(E) and p(E|S) is given by the Kullback–Leibler (KL) divergence:
$DKL(q(\mathrm{E})\,||p(\mathrm{E}|\mathrm{S}))\ = \int dH\ q(E)\ \ln q(\mathrm{E})p(\mathrm{E}|\mathrm{S})$.

The denominator in the right-hand side of the equation requires knowledge of p(E|S), however. Nevertheless, one can rewrite this equation as.

$DKL(q(\mathrm{E})\,||p(\mathrm{E}|\mathrm{S}))\ = F +\ \ln \mathrm{p(S)}$.

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

$F = dH \cdot q(\mathrm{E})\ln \cdot\ \mathrm{q(E)}p(\mathrm{S},\mathrm{E})$

Here, F is known as variational free energy or (more commonly) the negative of the evidence lower bound (Bishop 2007), which can be computed solely from the recognition and generative models (see Buckley et al., 2017 for a review).

[4] Different formulations of this claim are used in the neuroscientific and philosophical literature.

thus mechanistic theories, albeit—like mechanistic theories generally—theories that can be pitched at multiple levels of detail and abstraction (Luce 1995).[5] That is, they purport to describe how the process of free energy minimisation is implemented in the brain's mechanisms: their constituent parts, properties, and activities, and the way in which these parts, properties, and activities are causally, spatially, temporally, and hierarchically organised (see Kaplan & Craver, 2011). As Friston (2019a, p.179) puts it, to explain "neurobiology…one has to move from variational principles to particular process theories that conform to those principles."

The role of assumptions in connecting the FEP to process theories is important (see Hohwy, 2018, 2020a). Free energy minimisation is consistent with multiple "different generative models, different algorithmic approximations, and different neural implementations" (Gershman, 2019, p.4). Individual process theories, including influential theories such as predictive coding (Bastos et al., 2012) and active inference (Friston et al., 2017), thus map free energy minimisation onto neural mechanisms by means of certain assumptions about these characteristics. For example, predictive coding assumes that the generative model is hierarchically structured, that probability distributions are Gaussian and encoded by their means and precisions (i.e., inverse variance), and that the approximate posterior factorises across hidden states both within and between levels of the model, such that non-adjacent levels of the hierarchy are conditionally independent (Friston, 2005). Further, it is associated with an evolving implementational theory in which this specific process of variational Bayesian inference (understood as hierarchical precision-weighted prediction error minimisation) involves canonical cortical microcircuits, cortical hierarchies, and the role of various neuromodulators in realising precision-weighting (Bastos et al., 2012).

Thus, the FEP does not logically imply any specific process theory. Rather, it defines a set of process theories where membership within the set is determined by a theory's consistency with the objective function and computational scheme of free energy minimisation. Any specific process theory within this set then involves assumptions about the shape of the distributions, the algorithmic approximations, and the mapping from this computational scheme onto the component parts, operations, and organisation of the brain's mechanisms (see Allen & Friston, 2018; Clark, 2017; Friston, 2019a; Hohwy, 2018). Because there is no canonical interpretation of the term "predictive processing" in the neuroscientific and philosophical literature, I will stipulate that it can be understood as that theoretical framework in cognitive science defined by its commitment to drawing exclusively from this set of process theories that conform to the FEP.[6]

[5] Process theories that attempt "to explain some aspects of underlying… mechanisms and how they give rise to behaviour" are thus classically distinguished from phenomenological models that merely "attempt to characterize aspects and patterns of behavior without asking about the underlying, internal mechanisms that give rise to the behavior" (Luce 1995, pp.1–3; see also Kaplan and Craver 2011).

[6] As Hohwy (2018, p.166) puts it, "The philosophical literature often uses the notion 'predictive processing' as shorthand for the suite of solutions offered by FEP.".

4 There is no high road to predictive processing

Summarising Sects. 2 and 3, the FEP seems to derive a cognitive-scientific framework from first principles concerning the nature of self-organisation in general. If successful, this is remarkable. As Buckley et al., (2017, p.74) note, it suggests that the FEP "draws conclusions about neurocognitive mechanisms from extremely general statistical considerations regarding the viability of organisms' survival in unpredictable environments." Specifically, it implies that the FEP derives a substantive *constraint* on theories of neurocognitive mechanisms—namely, that they must describe how such mechanisms implement free energy minimisation—from transcendental reflection on conditions of the possibility of existence for self-organising systems.[7]

In recent work, Friston (2019a) describes this chain of reasoning spanning first principles concerning self-organisation in general to a substantive cognitive-scientific framework as the "high road" to predictive processing, in contrast to a "low road" that takes more conventional empirical means. "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 "takes us on a top-down journey from near existential nihilism to the riches of predictive processing" (Friston, 2019a, p.175).[8]

I will now argue that this high road to predictive processing is illusory. Specifically, the claim that the FEP implies a substantive constraint on process theories in cognitive science—namely, that they must describe how the brain's mechanisms implement free energy minimisation—rests on a fallacy of equivocation. There are two importantly different ways of interpreting the claim that all self-organising systems—or, more precisely, all systems that can be described in terms of nonequilibrium steady states and Markov blankets in the manner outlined in Sect. 2—must minimise variational free energy. On what I will call the *Descriptive FEP*, it states that the existence of all self-organising systems can be redescribed *as if* it involves the minimisation of free energy. On what I will call the *Explanatory FEP*, it states that the computational scheme of free energy minimisation is implemented in the mechanisms *by which* all self-organising systems maintain their

[7] The reference to a "substantive" constraint here is intended to highlight that this constraint goes beyond the trivial constraint that process theories must describe how a system maintains its existence; instead, such theories must explain this capacity by appeal to some form of free energy minimisation.

[8] Unfortunately, it is not clear exactly what Friston means by "predictive processing" here, and it may be that he has an unconventional understanding of the concept of necessary entailment. Nevertheless, for the reasons returned to in Sect. 5. 2, the interpretation that I have assumed here—which is consistent with much else that is written about the FEP by Friston and others—seems the most likely.

existence. Although the Descriptive FEP plausibly does identify a condition of the possibility of existence for self-organising systems, it does not imply that mechanisms in the brain implement free energy minimisation. Although the Explanatory FEP would imply this if it were true, it does not identify a condition of the possibility of existence for self-organising systems. Thus, the only interpretation of the FEP on which it plausibly identifies a necessary imperative for all self-organising systems provides no reason for thinking that free energy minimisation is implemented in the mechanisms by which the brain works.

In characterising these different interpretations of the FEP as descriptive and explanatory, I am embracing Kaplan and Craver's (2011, p.601) claim that "dynamical and mathematical models… *explain* (rather than *redescribe*) a phenomenon only if there is a plausible mapping between elements in the model and elements in the mechanism for the phenomenon" (my emphasis). Thus, although the Explanatory FEP and Descriptive FEP both target the same phenomenon—namely, the capacity of bounded systems to maintain a nonequilibrium steady state with their environment—they differ in their commitments concerning the mechanisms underlying this phenomenon. Specifically, the Descriptive FEP has no such commitments: it is silent on *how* self-organising systems maintain their existence over time. By contrast, the Explanatory FEP assumes the existence of a mapping between the computational scheme of free energy minimisation and the mechanisms underlying self-organisation. Specifically, it claims that the capacity of bounded systems to maintain themselves within the narrow range of states consistent with their existence is underpinned and maintained by mechanisms that implement the probabilistic models and variational inference involved in free energy minimisation.

In this section I clarify, expand upon, and defend these conclusions, before considering two objections in Sect. 5.

4.1 The explanatory FEP

According to the Explanatory FEP, the claim that all self-organising systems must minimise free energy implies that free energy minimisation is implemented in the mechanisms by which all bounded self-organising systems maintain a nonequilibrium steady state with their environments. That is, the explanatory FEP purports to explain—at least in an abstract and schematic way—how self-organising systems achieve this feat. Of course, as noted above (S3), variational free energy minimisation can take different forms at the level of specific probabilistic models and algorithms and such models and algorithms are in turn multiply realisable. In this sense the Explanatory FEP itself is silent on how the mechanisms of specific self-organising systems implement free energy minimisation. Nevertheless, it is committed to the claim that all such mechanisms do.

If the Explanatory FEP is true, it obviously would imply a substantive constraint on process theories in cognitive science. That is, if all self-organising systems maintain their existence by means of free energy minimising mechanisms, this must be true of the brain as well.[9] The Explanatory FEP is not true, however: it is not a condition of the possibility of existence for self-organising systems that they maintain their existence by means of mechanisms that implement free energy minimisation.

First, it is easy to conceive of bounded systems that persist over time by means of mechanisms that do not implement any form of variational Bayesian inference. Thus, there is no *conceptual* connection between free energy minimising mechanisms and self-organisation. Further, there are many actual systems for which the Explanatory FEP is false. Consider simple regulatory mechanisms to which the FEP is supposed to apply 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). *Nowhere in this simple mechanism is there anything involving the implementation of variational Bayesian inference* (Baltieri et al., 2020). Similarly, we can—and often do—build artificial intelligence systems that persist over time as bounded systems without implementing algorithms involving any form of variational Bayesian inference.[10]

This should not be surprising. The attempt to establish the FEP from first principles does not and could not establish the Explanatory FEP. The transcendental argument outlined in Sect. 2 is almost completely free of empirical content.[11] As Hohwy (2020b, p.8) puts it, its reasoning "moves a priori—via conceptual analysis and mathematics—from existence to notions of rationality (Bayesian inference) and

[9] An important qualification here is that this depends on whether the transcendental argument is correct to assume that self-organisation can be described in terms of Markov blankets and nonequilibrium steady states (see Colombo and Palacios 2021).

[10] Importantly, the point here is not that the FEP trivialises the notions of modelling and variational inference by extending them to extremely simple systems (see Van Es 2020). It is possible that some of free energy minimisation is implemented in the mechanisms by which all self-organising systems—from the simple to the most complex—self-organise. The point is simply that this possibility does not obtain. Further, I am also not arguing that some *specific* process theory associated with the FEP such as predictive coding is absent from the mechanisms by which certain systems work (e.g., Colombo and Wright 2018, p.20). Again, the point is that there are many systems—both possible and actual—that maintain themselves within those states consistent with their existence by means of mechanisms that do not implement *any* form of free energy minimisation.

[11] When establishing the a priori status of the transcendental argument, it is important to distinguish between two different claims: namely, the conditional claim that *if* a system can be described in terms of Markov blankets and nonequilibrium steady states, then it must minimise free energy; and the additional claim that all self-organising systems satisfy the antecedent of the conditional (i.e., can be described in terms of Markov blankets and nonequilibrium steady states). Although the justification of the conditional claim is a priori, one might question Friston's (2019a) assumption that the additional claim can be justified by a priori reflection on self-evident conditions that a system *must* satisfy to exist (see Colombo and Palacios 2021). I ignore this subtlety in the main text because my arguments hold even if one concedes that all self-organising systems *can* be described in terms of nonequilibrium steady states and Markov blankets. Specifically, even if it is ultimately an empirical fact that all self-organising systems can be described in terms of nonequilibrium steady states and Markov blankets, this fact alone is insufficient for establish anything about the mechanisms by which they come to be describable in this way.

epistemology (self-evidencing)" (Hohwy, 2020b, p.8). It thus purports to "derive a normative, a priori first principle from a provable definition of living systems" (Allen & Friston, 2018, p.2473). For this reason, proponents of the FEP are clear that the principle is unfalsifiable and "must *necessarily* be in play for any person or system we care to study" (Friston, 2019a, p.184; my emphasis; see also Hohwy, 2020b). By contrast, the Explanatory FEP is neither a priori nor necessary. It is a substantive empirical claim about the causal structure of the natural world. Given this, it is not the sort of claim that could be justified by such a priori reasoning.

To illustrate this, consider the three stages of the transcendental argument: the first purports to formalise two self-evident conditions of existence for self-organising systems; the second draws on information theory to describe the satisfaction of these conditions in probabilistic terms; and the third draws on variational calculus to clarify how such conditions *as represented in probabilistic terms* can be satisfied in a way that is computationally tractable. As Friston (2012, p.2101) puts it, such reasoning thus "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 just that, however: a *description* of such systems, and not—or at least not necessarily—a feature of the systems themselves (Van Es, 2020). Thus, any connection between such probabilistic descriptions and variational approximations to Bayesian inference does not—and on its own could not—have any direct implications for our understanding of the mechanisms underlying their behaviour. To assume that it does is to confuse properties of a possible description of a system with properties of the system being described.

More generally, which features of a representation map onto its target is always an empirical question. Thus, even if—as the transcendental argument purports to establish—one can model self-organisation in terms of a probability distribution defined over an abstract state space, this does not imply that the properties, constraints, and implications relevant to this description will map onto the mechanisms by which such systems self-organise (Chater & Oaksford, 2000; Colombo & Wright, 2018, p.12). That is, the fact that one can *describe* the dynamics of a system in terms of free energy minimisation does not itself imply the existence of a mapping between this description and the concrete parts and operations that constitute the mechanisms that underlie and maintain such dynamics (Kaplan & Craver, 2011). Of course, such a mapping *might* exist in specific cases. For example, more limited versions of the Explanatory FEP might be true, such as those restricted to information-processing mechanisms in the neocortex (e.g., Friston, 2005). Nothing that I write in this article is intended to challenge empirical hypotheses such as this (see S6 below). The point is rather that such hypotheses *are* empirical: they are not implied or in any way justified by a priori reflection on conditions of the possibility of existence for self-organising systems.

This suggests a general lesson: insofar as the FEP is interpreted as a claim about all possible self-organising systems, it must be restricted to a claim about how such systems can be described or interpreted. That is, it suggests that the transcendental argument for the FEP at best establishes what I have called the *Descriptive FEP*.

4.2 The descriptive FEP

The Descriptive FEP does not explain, purport to explain, or constrain explanations of *how* bounded systems maintain a nonequilibrium steady state with their environment. Instead, it assumes the existence of this behaviour and contends that it can be *redescribed* as involving the minimisation of free energy (see, e.g., Gładziejewski, 2019; Klein, 2018).[12] At least in more recent presentations of the FEP, the Descriptive FEP often seems to be the intended interpretation. In describing the FEP, for example, Friston (2019b, p.24) writes that "the position taken here is not to ask *how* self-organisation emerges; rather, what properties do self-organising systems exhibit?" (my emphasis). Similarly, Kirchhoff et al. (2018, p.2) write of the FEP that "this teleological (Bayesian) interpretation of dynamical behaviour… *allows us to think* about a system that possesses a Markov blanket as some rudimentary (or possibly sophisticated) 'agent' that is optimizing something" (my emphasis).

Further, this interpretation is supported by frequent claims according to which all self-organising systems merely behave *as if* they minimise free energy. Thus, we are told 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, 2019b, p.179; my emphasis) and that "the states internal to a Markov blanket look *as if* they perform variational Bayesian inference" (Parr et al. 2020, p.11; my emphasis). Such claims bring the FEP in line with how other optimising "as if" principles are understood in the biological and social sciences: namely, to indicate that although a system's behaviour can be described as maximising some objective, the mechanism underlying the system's behaviour need not work by maximising the objective. For example, the individual-as-maximising agent principle in evolutionary biology holds that organisms can be described as if they seek to maximise fitness, where "as if" indicates that this principle is silent on the mechanisms underlying such fitness-maximising behaviour (Del Giudice, 2018, p.50). Similarly, rational choice models in the social sciences typically describe agents only as if they seek to maximise expected utility, where, again, "as if" is used to indicate that such models are silent on the mechanisms by which actions are generated (Chater & Oaksford, 2000).

Most importantly, if one interprets the FEP in terms of the Descriptive FEP, it plausibly does identify a necessary imperative for all self-organising systems. Specifically, although a priori reflection on conditions of the possibility of existence for self-organising systems cannot establish how self-organising satisfy these conditions, it is much more plausible that it can establish a way of redescribing them. In fact, as noted in the previous sub-section, this seems to be exactly how the transcendental argument works: it first redescribes the existence of self-organising systems in terms of nonequilibrium steady states and Markov blankets, it then redescribes a system's maintenance of a nonequilibrium steady state and Markov

[12] Klein (2018, p.2551) suggests that "talk about minimization of free energy and an organism's expectations is meant to be something like a description of how whole organisms behave," and Gładziejewski (2019, p.661) claims that the "FEP stands as an ingenious technical *redescription* of what adaptive or self-organising behavior is, rather than an *explanation* of it.".

blanket in terms of the avoidance of surprising sensory states, and then it redescribes the avoidance of surprising sensory states as if it involves the minimisation of free energy. Because such reasoning restricts itself to redescriptions of the fact that self-organising systems maintain their existence over time, it is intelligible how it could be justified by purely a priori reflection on this fact, and thus why the principle that results from this reasoning is unfalsifiable. That is, on this interpretation the reason that self-organising "systems that do not minimise free energy cannot exist" (Friston, 2013, p.2) is because minimising free energy is simply a redescription of the fact that they exist.

One might object that this interpretation is at odds with the third stage of the transcendental argument. As described in Sect. 2, the imperative to minimise free energy is introduced in this argument on the grounds that evaluating and minimising surprisal directly is impossible. Variational free energy minimisation is then alleged to solve this problem because it approximates the minimisation of surprisal based on information that a self-organising system can access (see Hohwy, 2020a; Friston, 2010, 2013). This reasoning suggests that free energy minimisation is invoked not merely to redescribe the fact that self-organising systems exist but to explain this fact. Thus, Friston et al.' (pp.5–6) write that "the free-energy principle is not a surprise-principle. A principle of minimum surprise is a tautological truism—*the free-energy principle explains how that truism is realised*" (my emphasis).

I am not arguing that the Descriptive FEP is the interpretation of the FEP that its proponents intend to establish from first principles, however. Although some of the quotes highlighted above do suggest this intended interpretation, others—such as this one—do not.[13] My argument is rather that the Descriptive FEP is the only interpretation of the FEP *licensed* by the transcendental argument. Thus, to the extent that proponents of the FEP think that the transcendental argument licenses any claims about the causal structures that produce, underlie, or maintain self-organisation, they are mistaken. The fact that free energy minimisation provides a computationally tractable means of approximating the minimisation of surprisal cannot on its own establish that the mechanisms underlying self-organisation implement free energy minimisation for the reasons already enumerated: not only do many bounded systems maintain their existence by means of mechanisms that do not implement any form of free energy minimisation, but the imperative to minimise surprisal in the first place reflects a decision about how to describe such systems, and not necessarily a problem that they confront themselves (see Colombo & Wright, 2018, p.12).

There is a more charitable way of interpreting the third stage of the transcendental argument, however. To see this, it is helpful to distinguish between *how actually* and mere *how possibly models* in science (Kaplan & Craver, 2011). Whereas the latter "save the phenomena" and accurately describe and predict the system's dynamics, the former aim to capture the causal structure responsible for

[13] For example, the literature on the FEP contains ubiquitous references to the idea that the FEP "tries to *explain* the ability of biological systems to resist a natural tendency to disorder" (Friston 2012, p.2101; my emphasis; see also Friston 2013; Ramstead et al., 2018), all of which are strongly in tension with the Descriptive FEP.

producing and sustaining those dynamics. In Weisberg's (2007) terminology, mere how possibly models exhibit *dynamical fidelity* but not *representational fidelity*. To achieve the latter, there must be a mapping between the model and the causal structure responsible for producing its target phenomenon (Kaplan & Craver, 2011; Weisberg, 2007). As argued above, however, one cannot establish such a mapping a priori. Thus, the transcendental argument cannot on its own establish that self-organising systems maintain their existence by minimising free energy. Nevertheless, it plausibly can establish a schematic how possibly model. This is simply to return to the Descriptive FEP, however. On this reading, the transcendental argument establishes that self-organising systems can be redescribed *as if* they minimise free energy. It does not establish that this description maps onto the mechanisms by which they self-organise.[14]

In summary, to the extent that the FEP identifies a condition of the possibility of existence for self-organising systems, it must be interpreted in terms of the Descriptive FEP. Crucially, however, on this interpretation the FEP provides no reason to believe that free energy minimisation is implemented in the mechanisms underlying self-organisation. Thus, even if this version of the FEP is true, it provides no reason for thinking that process theories in cognitive science must describe how mechanisms in the brain implement free energy minimisation.

5 Two objections

I have argued that the only interpretation of the FEP on which it plausibly identifies a necessary imperative for all self-organising systems provides no reason for thinking that free energy minimisation is implemented in the mechanisms by which the brain—or any other self-organising system—works. Before exploring the implications of this conclusion, I will consider two objections: that the foregoing arguments rest on dubious assumptions about the nature of scientific explanation and a priori reasoning (S5.1), and that they neglect the crucial role of auxiliary assumptions in connecting the FEP to specific process theories in cognitive science (S5.2).

5.1 Scientific explanation and the a priori

I have argued that the a priori justification of the FEP cannot explain how self-organising systems maintain their existence but can at best establish a way of redescribing this phenomenon. An anonymous reviewer objects that this argument rests on dubious assumptions about the nature of scientific explanation and the limits of a priori reasoning. Specifically, they point to the fact that highly idealised explanatory models justified by a priori mathematical reasoning are ubiquitous in

[14] For this reason, simulations of self-organisation drawing on free energy minimisation (e.g., Friston 2013) are also of limited relevance: although simulations can demonstrate how a given phenomenon could be generated, on their own they cannot establish how a phenomenon is *in fact* generated (see Kaplan and Craver 2011; Weisberg 2007).

the sciences. For example, Fisher's principle seeks to explain why the sex ratio in most sexually reproducing species is 1:1 by drawing on mathematical (game-theoretic) reasoning to derive this ratio as the equilibrium point that fitness-maximising organisms will settle on for a wide range of initial conditions. According to some philosophers, the explanatory value of equilibrium explanations like this inherits not from capturing the actual causal processes that underlie, produce, or maintain the explanandum phenomenon in specific cases, but rather from establishing a universality class of systems that will exhibit the same behaviour (e.g., sex ratio) despite differences in their concrete physical details and initial conditions, which are thus shown to be explanatorily irrelevant (e.g., Rice, 2015, 2018).

This suggest two lessons: first, that some models can genuinely explain a phenomenon without informing our understanding of the mechanisms by which it is produced or underpinned in specific cases; and second, that the status of such models as successful explanations can be justified in substantial part by a priori mathematical reasoning alone (see Rice, 2015). If so, then perhaps the arguments advanced in the previous section rests on mistaken assumptions about scientific explanation and the irrelevance of a priori reasoning in scientific research.

This objection misconstrues my arguments. First, to the extent that the a priori reasoning involved in the transcendental argument for the FEP explains certain aspects of self-organisation without informing our understanding of the mechanisms by which it is achieved (see, e.g., Colombo & Wright, 2018), this concedes the very thing at issue: namely, that such reasoning does not support the claim that free energy minimisation is implemented in the mechanisms underlying self-organisation. That is, like the Descriptive FEP, this interpretation provides no reason to believe that process theories in cognitive science must identify how the brain implements free energy minimisation.

More importantly, the transcendental argument for the FEP is fundamentally different in its epistemic status to the justification of Fisher's principle and other optimality explanations in the special sciences. Although the game-theoretic reasoning behind Fisher's principle—as with almost all optimality and equilibrium explanations—is highly idealised and rests on all sorts of simplifying but strictly false assumptions (see Rice, 2015), it nevertheless seeks to explain *why* the sex ratio in most sexually reproducing species is 1:1, and its applicability to real-world populations depends on substantive and contingent assumptions about the natural world's causal structure—most obviously, that the reproductive strategies of organisms are selected in accordance with their contribution to fitness (Rice, 2015). The transcendental justification of the FEP is fundamentally different from this. It is more radically a priori, applying to all possible bounded systems that exist over time, and it does not purport to explain—even in an extremely schematic and idealised way—why or how such systems maintain their existence. As Hohwy (2020a, p.18) puts it, it aims rather "to *analyse* existence in terms of surprise minimization, rather than naturalistically explain one by appeal to the other" (my emphasis). Specifically, it restricts itself to redescribing the existence of self-organising systems, and redescriptions are not explanations.

Thus, the arguments advanced in Sect. 4 do not depend either on the view that scientific explanations must inform our understanding of mechanisms or on dubious assumptions about the irrelevance of a priori reasoning in science. Instead, their basis is more specific: namely, that the distinctive form of a priori reasoning involved in the transcendental justification of the FEP at best establishes a redescription of the capacity of self-organising systems to maintain their existence, and that redescribing this phenomenon as if it involves free energy minimisation provides no reason for believing that free energy minimisation is implemented in the mechanisms by which it is achieved.

5.2 Auxiliary assumptions and process theories

In Sect. 4, I quoted Buckley et al.'s (2017, p.74) claim that the FEP "draws conclusions about neurocognitive mechanisms from extremely general statistical considerations regarding the viability of organisms' survival in unpredictable environments." In a previous version of this article, an anonymous reviewer objected that I had left out the sentence that follows this one: "*Under certain assumptions*…it entails a hierarchical predictive processing model geared towards the inference and control of the hidden causes of sensory inputs…" (Buckley et al., 2017, p.74, my emphasis). The rebuke highlights a general objection: namely, that proponents of the FEP do not claim that it is the FEP alone that has important implications for our understanding of the mechanisms by which self-organising systems maintain their existence. Instead, it is only ever the FEP *in conjunction with certain auxiliary assumptions* about specific systems that carries such implications.

This response constitutes a natural interpretation of Hohwy's (2018, 2020a, 2020b) recent work exploring the epistemic status and implications of the FEP. Specifically, Hohwy (2020a) acknowledges that the FEP is a priori and impervious to empirical disconfirmation. In this sense his interpretation conforms closely to what I have called the Descriptive FEP. Nevertheless, Hohwy also argues that the FEP should function as a "regulatory principle" for the construction of process theories in cognitive science, noting that "even if it is misguided to ask for empirical evidence for FEP itself, empirical evidence can be had for the process theories under FEP" (Hohwy, 2020a, p.9). Importantly, however, Hohwy (2020b, p.6) stresses that "process theories can be said to conform with FEP but *are not entailed by it* since… various *assumptions* are needed to get to process theories for particular systems" (my emphasis). For this reason, Hohwy (2020a, p.20, fn.24) explicitly denies that the "FEP itself implies cognitive architecture": "Notions of architecture," he writes, "will need to build on assumptions about the particular system in question."

As noted in Sect. 3, Hohwy is correct that auxiliary assumptions are needed to translate the generic computational scheme of free energy minimisation into process theories for specific systems. Nevertheless, this fact does not threaten the conclusion of this article. The issue is not whether the FEP implies the truth of a *specific* process theory in cognitive science, but whether there is an interpretation of the FEP on which it establishes both a condition of the possibility of existence for self-organising systems and a substantive *constraint* on process theories in cognitive science: namely, that such theories must describe how mechanisms in the brain implement free energy minimisation. For the reasons outlined in Sect. 4, I have argued that there is not. The only interpretation of the FEP on which it plausibly establishes a necessary imperative for all self-organising systems—namely, what I have called the Descriptive FEP—does not imply that free energy minimisation *of any form* is implemented in the mechanisms by which they self-organise. Appealing to auxiliary assumptions about how the generic scheme of free energy minimisation maps onto the mechanisms of a specific system does not address this argument. As far as the Descriptive FEP is concerned, there is no reason for assuming that any such mapping exists.

Hohwy might respond that the justification of auxiliary assumptions about how specific systems implement free energy minimisation is empirical and does not derive from the FEP itself. As noted above (S4.1), however, nobody denies or could deny that there could be an empirical justification of the hypothesis that some form of free energy minimisation is implemented in the mechanisms by which a specific system works. The question is what the FEP—understood as the claim that all self-organising systems *must* minimise free energy—adds. If I am right, it adds nothing: the fact that all self-organising systems can be redescribed as if they minimise free energy provides no reason for assuming that free energy minimisation is implemented in the mechanisms by which they self-organise. Specifically, it provides no reason for searching for auxiliary assumptions about how self-organising systems implement a specific form of free energy minimisation.

Thus, the appeal to auxiliary assumptions does not challenge the conclusion of this article. Perhaps Hohwy's appeal to auxiliary assumptions is not intended to challenge this conclusion, however. Specifically, perhaps the appeal to auxiliary assumptions is intended to concede that the FEP itself implies no interesting constraints on process theories in cognitive science. For example, an anonymous reviewer objects to the treatment of process theories in this article on the following grounds:

> "Although under the FEP it is true that "non-process theories", in so far as they are an accurate description of behaviour in self organising systems, will also act "as-if" they are minimising free energy, *this is not the point of process theories*. The point of process theories is to derive testable theories that describe dynamics that explicitly self-organise to a free energy minimum" (my emphasis).

This interpretation of the FEP and its relationship to process theories concedes the central thesis of this article: namely, that the fact that all self-organising systems must minimise free energy provides no reason to believe that process theories in cognitive science must describe how mechanisms in the brain implement free energy minimisation. That is, it concedes that the FEP is consistent with—and provides no basis for rejecting—process theories within which the concepts of free energy minimisation and variational Bayesian inference do not play any role, and thus it abandons the idea that there is a first principles justification of work in the cognitive sciences that draws on such concepts to explain psychological and neurocognitive phenomena.

This concession seems to be strongly at odds with much that is written about the FEP by its proponents in the cognitive-scientific and philosophical literature. For example, it is inconsistent with the idea that the FEP constitutes “an attempt to explain the structure and function of the brain, *starting from the very fact that we exist”* (Friston, 2009, p.293, my emphasis), with the idea that there is a “high road” to predictive processing, with frequent claims that process theories are “entailed” by the FEP (see Friston, 2010, 2019a), with the idea that “for an organism to resist dissipation and persist as an adaptive system…it *must* embody a probabilistic model of the statistical interdependencies and regularities of its environment” (Ramstead et al., 2018, p.2; my emphasis), and with Friston’s (2010, p.136) assertion that “if the arguments underlying the free-energy principle hold, then the real challenge is to understand how it manifests in the brain” (Friston, 2010, p.136). All such ideas presuppose that the FEP’s claim that all self-organising systems *must* minimise free energy provides a reason to believe that free energy minimisation must be implemented in the mechanisms by which they self-organise. Once one abandons this—that is, once one concedes that the fact that all self-organising systems must minimise free energy provides no reason for favouring process theories that posit a form of free energy minimisation—all such ideas must be rejected.

Beyond this dialectical point, however, why does such a concession matter? I take up this question in the concluding section.

6 Conclusion: Summary and implications

I have argued that there is no first principles justification of the claim that mechanisms in the brain implement free energy minimisation. The idea of such a “high road” to predictive processing—that is, a derivation of a substantive constraint on process theories in cognitive science from transcendental reflection on conditions of the possibility of existence for self-organising systems—rests on a fallacy of equivocation. The FEP can be interpreted as a redescription of the capacity of self-organising systems to maintain their existence, or as a causal explanation of this capacity. Although the former interpretation plausibly can be derived from first principles, it does not imply that free energy minimisation is implemented in the brain. Although the latter interpretation would imply this if it were true, it cannot be derived from first principles. Thus, the only interpretation of the FEP on which it plausibly establishes a necessary imperative for all self-organising systems provides no reason for thinking that free energy minimisation is implemented in the mechanisms by which the brain works, and thus no reason for favouring process theories in cognitive science that draw on the concepts and formal tools of free energy minimisation or variational Bayesian inference to explain psychological phenomena.

Interestingly, the idea that there is a distinction between models that posit free energy minimisation as *explanations* of worldly phenomena and the principle that all self-organising systems *must* minimise free energy was acknowledged in an early article by Friston et al., (2006, p.71):

> “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 principle that free energy minimisation is mandatory in biological systems has a more “fundamental” status than mechanistic explanations of biological capacities, it has a fundamentally different epistemic status. If I am right, free energy minimisation can only be viewed as mandatory in this way—that is, as a condition of the possibility of existence for biological systems—if interpreted as a means of redescribing the existence of biological systems. When viewed in this way, however, it ceases to inform or constrain our understanding of the mechanisms underlying the behaviour of such systems. Specifically, the only connection that that it bears to models that posit free energy minimisation as causal explanations of worldly phenomena is a formal connection in the mathematics used.

Why does this matter? In some important respects, it does not. For example, nothing that I have written in this article is intended to challenge the utility of using the FEP as a modelling framework in the cognitive sciences (see Andrews, 2021). The evolving formal apparatus surrounding the FEP has proven highly fecund when it comes to generating process theories describing psychological phenomena such as learning, perception, sensorimotor control, and decision-making (see Hohwy, 2020b). My thesis is simply that the fact that all self-organising systems must minimise free energy provides no support for this research programme. Its success—as with the success of all research programmes in the cognitive sciences—is ultimately an empirical matter.

In at least two other respects, however, the conclusion of this article does matter. First, extraordinary claims are advanced on behalf of the FEP in cognitive science and philosophy: for example, that it provides a “unified brain theory” (Friston, 2010) and “grand unifying principle for cognitive science and biology” (Hohwy, 2020a) that is allegedly “widely recognised in neuroscience as a unifying theory of the brain and biobehaviour” (Ramstead et al., 2018, p.1) and that establishes that “*all* neuronal processing (and action selection) can be explained by… minimising variational free energy” (Friston et al., 2017, p.1; my emphasis). Although such claims might be justified in part by appeal to empirical evidence for the process theories formally associated with the FEP (see Clark, 2016; Hohwy, 2020a), they are often connected with the idea that process theories drawing on variational free energy minimisation are somehow distinctive in the cognitive sciences in their connection to first principles concerning life or existence more generally (see, e.g., Friston, 2010, 2019a; Friston et al., 2017; Hohwy, 2016; Ramstead et al., 2018). Thus, Buckley et al., (2017, p.74) write that the FEP

> “is an ambitious project, spanning a chain of reasoning from fundamental principles of biological maintenance essential for sustainable life, to a mechanistic brain theory that proposes to account for a startling range of properties of perception, cognition, action and learning.”

If I am right, such a claim is extremely misleading: there is no chain of reasoning—no “high road”—that takes one from first principles concerning life or existence to a mechanistic brain theory or framework for generating such theories. Once one recognises this, it is reasonable to ask what could justify even a moderate degree of confidence in some of the most ambitious claims associated with the FEP of the sort highlighted above. Even focusing just on the limited domain of perception, for example, the empirical evidence for variational free energy minimisation as an explanation of how the brain works remains highly controversial (see Walsh et al. 2020), and critics have raised worries that many FEP-based models in cognitive science involve post hoc “just so” stories that—like the Descriptive FEP more generally—redescribe psychological and neurocognitive phenomena as involving free energy minimisation without validating such models against competing explanations (Litwin & Miłkowski, 2020). These worries might not be as pressing if such work had a principled theoretical justification of the sort provided by the “high road” to predictive processing. If the conclusion of this article is correct, however, there is no such justification. Thus, I hope that the present article adds additional pressure on some of the most ambitious claims advanced on behalf of the FEP and related ideas in the cognitive sciences.

Second, there is currently a large and growing literature exploring the epistemic status and scientific and philosophical implications of the FEP (Bruineberg et al., 2018; Colombo & Wright, 2018; Friston, 2019a; Hohwy, 2020a; Van Es, 2020). I hope that this article demonstrates the importance of disambiguating between different interpretations of the FEP in pursuing these questions (see also Andrews, 2021). Although the hypothesis that information-processing mechanisms in the human brain implement a process of variational Bayesian inference (Friston, 2005) and the claim that all self-organising systems must minimise free energy are often referred to as “the FEP” (e.g., Friston, 2010, 2013; Ramstead et al., 2018), they are in fact radically different claims not just in terms of their scope but in terms of their theoretical commitments. Failure to acknowledge this fact will result in inevitable confusion. For example, Ramstead et al., (2018) write that “the FEP… *describes*, formally, the… dynamics of all living systems” (p.8, my emphasis), and also that the FEP “has been extended beyond the brain to *explain* the dynamics of living systems, and their unique capacity to avoid decay” (p.1, my emphasis). Such claims illustrate the confusions that I have sought to highlight in this article. The claim that all living systems *must* minimise free energy is not an extension of the claim that mechanisms in the brain work by minimising free energy but a completely different kind of claim, and a formal description of the capacity of living systems to avoid decay is not an *explanation* of that capacity.

Similarly, there is currently a large and growing literature exploring the implications of the FEP for foundational questions in the philosophy of mind and cognitive science, including the representational and inferential status of psychological phenomena such as perception (e.g., Bruineberg et al., 2018; Kiefer & Hohwy, 2018; Van Es, 2020; Williams, 2018). In addressing such questions, it is crucial not to draw lessons about free energy minimisation-based *explanations* of psychological phenomena (e.g., Friston, 2005) from consideration of the principle that all self-organising systems *must* minimise free energy, and vice versa. The empirical hypothesis that variational Bayesian inference is implemented in the mechanisms by which the brain works will have radically different theoretical and philosophical implications to the claim that all self-organising systems can be redescribed as if they minimise free energy. Thus, one cannot simply generalise lessons extracted from one proposal to the other. I hope that the current article illustrates and emphasises this point, and therefore contributes to clarifying future discussions about the theoretical and philosophical importance of the FEP.

Acknowledgements For helpful comments, corrections, and criticisms, I would like to thank several anonymous reviewers, Andrew Corcoran, Stephen Gadsby, Marcella Montagnese, Matteo Colombo, Maxwell Ramstead, and Jakob Hohwy.

Funding This work was supported by an Early Career Research Fellowship at Corpus Christi College, University of Cambridge.

Declarations

Conflicts of interest No conflicts of interest or competing interests.

References

Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese, 195*(6), 2459–2482.

Andrews, M. (2021). The math is not the territory: Navigating the free energy principle. *Biology & Philosophy, 36*(3), 1–19.

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.

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.

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

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese, 195*(6), 2417–2444.

Buckley, C. L., Kim, C. S., Mcgregor, S., & 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. (2016). *Surfing Uncertainty*. Oxford University Press.

Clark, A. (2017). Predictions, precision, and agentive attention. *Consciousness and Cognition, 56*, 115–119.

Colombo, M., & Wright, C. (2018). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese, 19*(S14), 3463–3488.

Colombo, M., & Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biology & Philosophy*, *36*(5), 1–26.

Es van, T. (2020). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior, 29*(3), 315–329.

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. (2013). Life as we know it. *Journal of the Royal Society Interface, 10*(86), 20130475.

Friston, K. (2019b). A free energy principle for a particular physics. *ArXiv, 1906*, 10184.

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

Friston, K., Da Costa, L., & Parr, T. (2020). Some interesting observations on the free energy principle. *ArXiv Preprint, 2002*, 04201.

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

Gershman, S. J. (2019). What does the free energy principle tell us about the brain? *ArXiv Preprint, 1901*, 07945.

Gładziejewski, P. (2019). Mechanistic unity of the predictive mind. *Theory & Psychology, 29*(5), 657–675.

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

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

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

Hohwy, J. (2020b). Self-supervision, normativity and the free energy principle. *Synthese, 10*, 1–25.

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

Hohwy, J. (2018). Prediction error minimization in the brain. In M. Sprevak & M. Colombo (Eds.), *Handbook to the Computational Mind* (pp. 159–173). Routledge.

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.

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese, 195*(6), 2387–2415.

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of The Royal Society Interface*, *15*(138), 20170792.

Klein, C. (2018). What do predictive coders want? *Synthese, 195*(6), 2541–2557.

Litwin, P., & Miłkowski, M. (2020). Unification by fiat: arrested development of predictive processing. *Cognitive Science, 44*(7), e12867.

Parr, T., Da Costa, L., & Friston, K. (2020). Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A*, *378*(2164), 20190159.

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2018). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews, 24*, 1–16.

Rice, C. (2015). Moving beyond causes: Optimality models and scientific explanation. *Noûs, 49*(3), 589–615.

Rice, C. (2018). Idealized models, holistic distortions, and universality. *Synthese, 195*(6), 2795–2819.

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.

Weisberg, M. (2007). Who is a Modeler? *The British Journal for the Philosophy of Science, 58*(2), 207–233.

Williams, D. (2018). Predictive processing and the representation wars. *Minds and Machines, 28*(1), 141–172.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.