Synthese (2021) 199:29–53
https://doi.org/10.1007/s11229-020-02622-2

NEUROSCIENCE AND ITS PHILOSOPHY

Self-supervision, normativity and the free energy principle

Jakob Hohwy[1]

Received: 5 August 2019 / Accepted: 6 March 2020 / Published online: 13 March 2020

Abstract
The free energy principle says that any self-organising system that is at nonequilibrium steady-state with its environment must minimize its (variational) free energy. It is proposed as a grand unifying principle for cognitive science and biology. The principle can appear cryptic, esoteric, too ambitious, and unfalsifiable—suggesting it would be best to suspend any belief in the principle, and instead focus on individual, more concrete and falsifiable 'process theories' for particular biological processes and phenomena like perception, decision and action. Here, I explain the free energy principle, and I argue that it is best understood as offering a conceptual and mathematical analysis of the concept of existence of self-organising systems. This analysis offers a new type of solution to long-standing problems in neurobiology, cognitive science, machine learning and philosophy concerning the possibility of normatively constrained, self-supervised learning and inference. The principle can therefore uniquely serve as a regulatory principle for process theories, to ensure that process theories conforming to it enable self-supervision. This is, at least for those who believe self-supervision is a foundational explanatory task, good reason to believe the free energy principle.

Keywords Self-organisation · Self-evidencing · Self-supervision · Unsupervised learning · Free energy principle · Normativity · Active inference · Principles · Predictive processing · Process theories

1 Introduction

The *free energy principle* (FEP) says that any self-organising system that is at nonequilibrium steady-state with its environment must minimize its free energy (Friston 2010; Friston and Stephan 2007). The principle is more than a limited characterization of self-organising systems because, ostensibly, it can be used to account comprehen-

✉ Jakob Hohwy
jakob.hohwy@monash.edu

[1] Cognition and Philosophy Lab, Philosophy Department, School of Philosophical, Historical and International Studies, 20 Chancellors Walk, Room E674, Monash University, Clayton, VIC 3800, Australia

sively for decision-making, action, perception, attention, and a wide range of mental and biological phenomena relating to self, mentalising, emotion, learning, memory, value, decision, and even life and morphogenesis (for introductions and discussions, see Clark (2016), Hohwy (2013); for recent accounts pertaining to foundational issues in biology, see Constant et al. (2018), Friston (2013), Kirchhoff et al. (2018)).

Due to this extreme ambition, there is considerable interest in—and fascination with—FEP. In discussion, this interest is often tempered with scepticism about the principle's ability to truly apply to all biological and mental phenomena, as well as doubts about the availability of empirical evidence in support of the principle. One central worry is that FEP may be unfalsifiable, which would, in the eyes of many empirical scientists, place it beyond scientific inquiry.

Alongside the growing interest in the free energy principle there is also considerable interest in the notion of 'predictive processing' (PP) (Clark 2013). Predictive processing has a long pedigree starting with Helmholtz's notion of unconscious inference through various formulations in the last century such as analysis by synthesis, epistemological automata, perception as hypothesis testing, and the Bayesian brain hypothesis (Clark 2013, 2016; Gregory 1968, 1980; Helmholtz 1867; Knill and Pouget 2004; MacKay 1956; Neisser 1967; Yuille and Kersten 2006). PP is the idea that many mental phenomena can be explained by appealing to the brain as harbouring a constellation of more or less purpose-built prediction mechanisms. PP leaves it an open question if other types of mechanisms can co-exist with predictive mechanisms, and if some mental phenomena, for example value processing or emotional valence, might be explained by non-predictive mechanisms.

Since FEP can be cast in terms of prediction mechanisms, it is natural to speculate about how FEP relates to PP. Some versions of PP are closely aligned with FEP (Hohwy 2013) but other versions of PP, and kindred types of approaches, are agnostic or silent on any principled relation between PP and FEP (Bar 2011; Clark 2016; Heeger 2017). Quite a few PP-type views predating FEP do not seem to explicitly anticipate the core ideas of FEP, though most are consistent with FEP given certain assumptions (for discussion of such views, see Friston 2010).

It may seem attractive to divorce PP from FEP, since PP appears to have considerable explanatory power independently of FEP, making it possible to circumvent the perhaps unrealistic ambitions of FEP and the issues surrounding its evidential support. Perhaps we should believe what PP offers, and leave FEP behind?

Here, I do not aim to show that FEP is true. Rather, I offer a more pragmatic argument: if one believes accounting for self-supervised (normatively constrained) learning and inference is a foundational task in philosophy and other cognitive science disciplines, then there is good reason to believe FEP. The paper therefore has three intersecting aims: (1) I first argue that FEP is best understood as offering mathematically enshrined conceptual analysis, and therefore not something in need of empirical evidence (Sects. 2, 3). (2) I then develop the view that, viewed like that, FEP delivers a new and important solution to the foundational problem about how organisms can display truly self-supervised, normatively constrained learning and inference. That is, I show that FEP ties self-organisation to self-supervision via the key notion of self-evidencing, that FEP thereby builds normativity into the conditions for existence of self-organising systems, and that this is what grounds self-supervision.[1] This sets FEP's notion of self-supervision apart from other approaches to self-supervision and normativity (Sects. 4, 5). (3) Therefore, I argue, it is attractive to use FEP as a regulatory principle in somewhat the same sense as some philosophical principles are: if FEP regulates predictive 'process theories' for perception, action and other cognitive processes, then we can view those theories as describing self-supervised systems (Sect. 6).

2 The free energy principle

FEP says that any self-organising system that is at nonequilibrium steady-state with its environment must minimize its free energy (Friston 2010; Friston and Stephan 2007). The technical notion 'minimising free energy' belongs in statistical physics and machine learning. It can be thought of generally as maximising the likelihood of (sensory) input to the system, conditioned on a model of how that input was generated; or, given certain assumptions (i.e., a mean field approximation and the Laplace assumption, explained below) as minimising the long-term average of prediction error (usually, in a hierarchical setting).

The reasoning that leads to FEP begins with analysis of a particular concept of a *self-organising system that is at nonequilibrium steady-state with its environment*. The concept of a self-organising system that can attain steady-state is of fundamental interest because it picks out biological systems that *exist*. That is, biological inexistence is marked by a tendency to disperse throughout all possible states in state space (e.g., the system ceases to exist as it decomposes, decays, dissolves, dissipates or dies). In contrast, to exist is to revisit the same states (or their close neighbourhoods).[2]

The states visited by the system can be thought of in terms of the causal effects on its boundary (e.g., cell membrane, or sensory organs including the skin). Thus,

[1] In machine learning, there is a recent focus on self-supervision as a crucial subtype of unsupervised learning, that is, learning that does not require labelled training data. In this research area, self-supervised learning proceeds by withholding some of the data and then letting the system predict it (cf. contextual cues); in this way data supervise the learning (such that self-supervision, in some sense, is a type of supervised learning too). In this paper, the notion of *self-supervision* is used in a quite general and generic sense, which captures the autonomy of self-supervision and its reliance on an internal model of causes of sensory input; it is intended to capture 'truly' unsupervised learning, namely where the system only relies on itself and its own exploration for normative constraints on learning and inference (this is pursued in Sects. 4, 5 below); the notion of unsupervised learning goes back to at least Barlow (e.g., Barlow 1989) and is now a textbook topic in machine learning.

[2] There might of course be other concepts of the existence of biological organisms. The current approach focuses on the idea of nonequilibrium steady state, which immediately suggests a statistical analysis. There is a long history connecting existence of biological systems with nonequilibrium steady state (or far from equilibrium states); (see, e.g., Ashby 1947; Nicolis and Prigogine 1977; Prigogine and Nicolis 1971; Schrödinger 1944; Von Bertalanffy 1950). More philosophically-oriented discussion can be found in Mark Bickhard's work, such as (Bickhard 2009), which draws on the dynamical systems approach to discuss both the nature of representation and topics close to FEP; there is an interesting project in exploring their affinities (see also Bickhard 2016). There are non-equilibrium steady state systems that are not biological in the common sense, such as non-biological adaptive systems, and perhaps phenomena such as tornadoes. There is debate about this issue of the scope of FEP; (e.g., Sims 2016). For this paper, I set aside a full discussion of scope issues and focus on the kinds of systems for which self-supervision and normativity are commonly discussed.

decomposition is characterised by exposure to effects the system is not prepared for, and that eventually destroys its boundary and disperses it (e.g., falling into a volcano or drowning). An existing system's 'states' can then usefully be described in terms of its *sensations,* which mediate the influence of the external world upon the system.

Now consider the states of an existing system in probabilistic terms and in the context of all the possible states in which it could, in principle, be found. If the system exists, it is more likely that it is found in some, and not others, of all the possible states. Simplistically speaking, if there are innumerable states, and the system tends to visit only states 17–21, then there is a low probability we find it in state 87, say. This is a trivial consequence of how we are here conceptualising existence. The states the system tends to visit define its *model*, and, given this model, its states (i.e., sensations) are divided into unsurprising and surprising (i.e., high and low probability states). Existence then entails that the system receives unsurprising sensations and that it avoids surprising states. More simply put, existence implies minimising surprise.[3]

Conceptual analysis of the concept of *existence of self-organising systems* therefore leads to the notion of surprise minimization.[4] If it is assumed that states are dynamic, that is, changing in a non-linear fashion over time, it becomes incumbent on such self-organising systems to *act* to visit unsurprising states. If they do not act, then, over time, they will be exposed to surprising states that will lead to dispersal. Hence, if a system exists, it must be able to, as it were, turn back from surprising states and head back into unsurprising territory.

The upshot of this conceptual analysis is then that existence (self-organisation in dynamic environments) entails surprise-minimising agency. Or, expressed in something that looks somewhat like FEP but does not yet mention free energy: any self-organising system that is at nonequilibrium steady-state with its environment must minimize surprise, given a model.

Note that minimizing surprise, given a model, is equivalent to maximising evidence for the model. This follows because surprise is the negative log marginal likelihood, where the marginal likelihood is known as model evidence in Bayesian statistics; namely, the probability of some data given the model of how those data were generated. Sensations that are not surprising given that model become evidence *for* the model. So an equivalent reformulation of the previous expression is: any self-organising system that is at nonequilibrium steady-state with its environment must maximise the evidence for its model. Existence thus implies *self-evidencing* (Hohwy 2016). There is then a

[3] Surprise is also known as *surprisal* and corresponds to Shannon's self-information: the negative log probability of some system states, conditioned that system or model. The average self-information of a system is entropy; suggesting that avoiding surprises places an upper bound on the entropy of a system's states.

[4] There is substantial debate about the very notion of conceptual analysis and the a priori in philosophy. I do not here rely on any particular approach but merely on a basic sense of conceptual analysis as given by grasp and elucidation of concepts, and the way in which that is a priori at least in the sense of not immediately requiring empirical investigation. It may be that our conceptual analysis is susceptible to empirical evidence, and this can lead to considerations about whether the initial analysis failed, or whether the concept has changed; it may also be that there are different conceptual analyses of the same terms (e.g., in different cultures), which in turn raises questions whether this is a case of different understandings of the same concepts or of different concepts (much of this discussion plays out in Frank Jackson's defence of conceptual analysis and the debate following Jackson (1998)). My argument here is subject to the eventual fate of these questions, together with other purported a priori conceptual analyses in philosophy.

conceptual link between existence and epistemological notions of evidence, which will matter for the argument below.

Next, on the route to FEP, is the question *how* self-organising systems can minimise surprise? Or, equivalently, how they can exist? How do they exist rather than not exist? The answer to the question cannot be what the conceptual analysis so far may seem to suggest: that systems exist by simply consulting an internal model from which they can read off what sensations would be surprising and then proceed to avoid those sensations. The reason this is not the answer constitutes the pivot point toward FEP; namely, that "a system cannot know whether its sensations are surprising" (Friston 2010: p. 128).

To see this problem, consider first that there is no absolute notion of surprise. The surprise of states, or sensations, is gauged only relative to a model that defines a set of expectations of sensations. When a system encounters a given sensation, directly gauging the surprise depends on knowing the model. But knowing the model would mean averaging over an infinite number of copies of oneself in all possible states, so as to construct a statistical model of which of those states are typical and which not. This kind of knowledge we cannot feasibly obtain. Put in terms of sensations, it cannot be that knowing the surprise of a sensation requires having sampled all possible sensations. And if the system cannot know what sensations are surprising, it cannot act to minimise surprise. Self-organising systems must be doing something else, if they exist. Note the character of this problem. Systems are finite (bounded) existences in time, and cannot, as it were, step outside themselves and assess their expected states from an unbounded, all-knowing, external point of view.

FEP demonstrates how to avoid the problem of not knowing the model. FEP accepts that systems cannot directly access surprise, and offers instead a method for implicitly minimising surprise, in virtue of quantities that the system *can* access. Though the mathematical reasoning here comes from variational calculus, the initial idea is simple: instead of trying to directly infer the model, the system assumes a *recognition* model,[5] q and then optimises this model in a way that guarantees the minimisation of surprise (approximately). The key move behind FEP is to minimise a free energy functional[6] F, of the model q and sensations, s, that furnishes an upper bound on surprise. F thus depends on how well the recognition model's expectations account for sensations, and how well the sensations can be made to fit with the model's expectations (F is called 'free energy' due to formal similarity with Gibbs free energy in thermodynamics).

The next question is how the assumed recognition model or belief, q, relates to the model of how sensory states were generated, m. The question can be stated in terms of the divergence between two probability densities q and p, where q is the belief distribution over the causes of sensations (i.e., external states) encoded by the internal

[5] The recognition *model* is sometimes described as a recognition *density*; namely, an approximate posterior probability density over unknown (external) states of the world causing sensory states. This model or density can be considered a *Bayesian belief* about something; namely, unknown states of the world (note that Bayesian beliefs are not propositional beliefs). Note also a distinction I set aside here: optimising a recognition density is different from optimising a model. In statistics, this is the difference between Bayesian model inversion—to produce an approximate posterior over unknown causes, given a model—and Bayesian model selection—0to produce an approximate posterior over competing models.

[6] A *functional* is a function of a function. The free energy is a functional because it is a function of a probability density; namely, q, which is a probability density function of external states.

states of a system and p is the true but unknown posterior density over the causes, given the sensory consequences. This divergence is captured by the Kullback–Leibler divergence, KL(q||p), which is the relative entropy of the two densities (such that the KL-divergence indicates the information lost when q is used to approximate p). KL(q||p) tells us how good q is as a stand-in for what the system should infer about external states beyond its boundary, given its sensations. The key twist here is that although p cannot be evaluated directly by the system, the KL-divergence can be evaluated. In brief, we can simply add this divergence to surprise; thereby creating a free energy functional that the system can minimise.

The task is now to make the divergence as small as possible, such that, from a probabilistic point of view, q and p become more similar, or identical: little or no information is then lost by using q to approximate p. By writing out the KL-divergence, re-arranging terms, and taking advantage of the fact that the KL-divergence is always equal to or greater than zero, we get the mentioned result that the free energy bounds surprise, where surprise is defined as the negative log probability of the sensory input, under a given generative model (for more formal introductions, see Bishop 2007; Bogacz 2017; Buckley et al. 2017; Friston and Stephan 2007). To illustrate simply, from KL(q||p), we arrive (simplifying and abbreviating) at an expression like this, F = KL(q||p) − logp(s|m). This equation means that, if the sensations (− logp(s|m)) are fixed and F is minimised, KL(q||p) (which is positive or 0) must necessarily get smaller; that is, the recognition density must approximate the true posterior density, given the sensations. So, merely by operating on the quantity, F, which the system can evaluate, the recognition density gradually becomes the true posterior density over the causes of sensations. In sum, the intractable problem of evaluating surprise can then be overcome, if the system minimises free energy.[7]

As q approximates p, the recognition density is revised, also known as Bayesian belief updating, giving rise to perceptual inference and learning.[8] In other words, the system's internal states are updated to represent the causes of sensory input in a Bayes optimal fashion. This creates a tight free energy bound on surprise.

Next, we add *active states* to the system's sensory and internal states, which endows the system with the capacity to act, that is, to sample sensations selectively. In this setting, the free energy relates to the *complexity* of the recognition density (i.e., the divergence between the approximate posterior and beliefs prior to updating); this complexity measures the degrees of freedom used to explain new sensations.[9] In this

[7] The implicit conversion of an intractable integration problem into a tractable optimisation problem is due to Richard Feynman, who introduced the notion of variational free energy via the path integral formulation of quantum electrodynamics. It was subsequently exploited in machine learning, where minimising variational free energy is formally synonymous with approximate Bayesian inference.

[8] Notice that the problem of self-supervised learning and inference relates to the general problem of learning where any learning system seeking to estimate a data generating process would have to minimize *risk* (expected value of some loss function). The problem is that the relevant joint probability distributions are unknown. So, the system has to minimize some proxy (e.g., empirical risk; Vapnik 1995). The FEP uses a proxy as well, minimizing (expected) free energy, to minimise surprise.

[9] There are various complexity measures in the literature. An influential approach belongs with the Akaike Information Criterion (Akaike 1974). Under FEP, complexity is conceived as the KL divergence between the prior and posterior distributions; a smaller divergence indicates that less complexity was introduced to account for new observations.

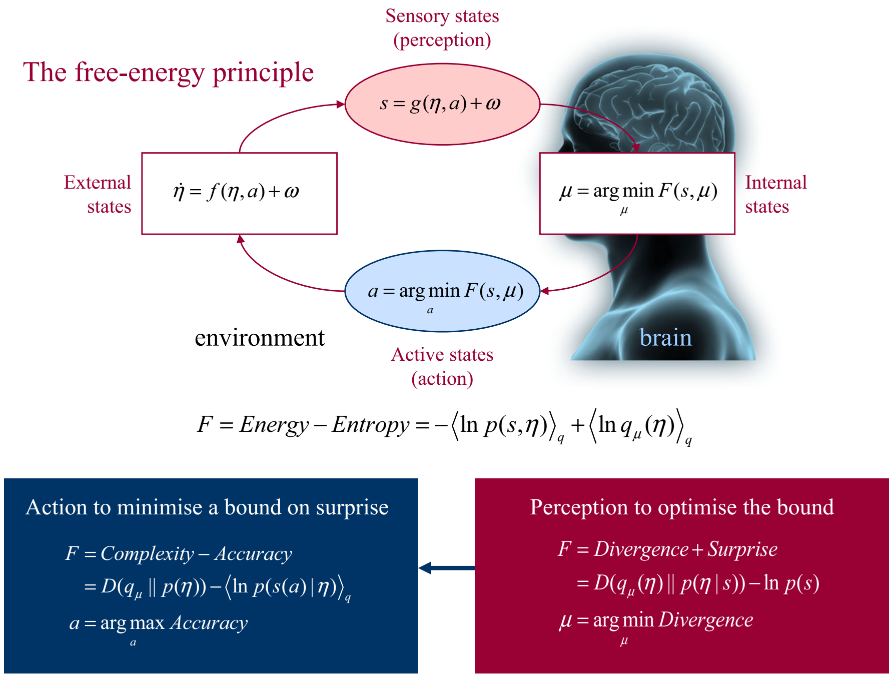
The free-energy principle
Sensory states (perception)
$s = g(\eta, a) + \omega$
External states
$\dot{\eta} = f(\eta, a) + \omega$
$\mu = \arg\min_{\mu} F(s, \mu)$
Internal states
$a = \arg\min_{a} F(s, \mu)$
environment
Active states (action)
brain
$F = Energy - Entropy = -\langle \ln p(s, \eta) \rangle_q + \langle \ln q_\mu(\eta) \rangle_q$
Action to minimise a bound on surprise
$F = Complexity - Accuracy$
$= D(q_\mu \,||\, p(\eta)) - \langle \ln p(s(a) \,|\, \eta) \rangle_q$
$a = \arg\max_{a} Accuracy$
Perception to optimise the bound
$F = Divergence + Surprise$
$= D(q_\mu(\eta) \,||\, p(\eta \,|\, s)) - \ln p(s)$
$\mu = \arg\min_{\mu} Divergence$

Fig. 1 Self-evidencing and the FEP: Upper panel: schematic of the quantities that define an agent and its coupling to the world. These quantities include the internal states of the agent μ (e.g., a brain) and quantities describing exchange with the world; namely, sensory input $s = g(\eta, a) + \omega$ and action a that changes the way the environment is sampled. The environment is described by equations of motion, $\dot{\eta} = f(\eta, a) + \omega$, that specify the dynamics of (hidden) states of the world η. Here, ω denote random fluctuations. Internal states and action both change to minimise free-energy or self information, which is a function of sensory input and a probabilistic belief $q_\mu(\eta)$ encoded by the internal states. Lower panel: alternative expressions for free-energy illustrating what its minimisation entails. For action, free-energy (i.e., self-information) can only be suppressed by increasing the *accuracy* of sensory data (i.e., selectively sampling data that are predicted). Conversely, optimising internal states make the representation an approximate conditional density on the causes of sensory input (by minimising a Kullback–Leibler divergence D between the approximate and true posterior density). This optimisation makes the free-energy bound on self-information tighter and enables action to avoid surprising sensations (because the divergence can never be less than zero). When selecting actions that minimise the *expected free energy*, the expected divergence becomes (negative) *epistemic value* or *salience*, while the expected surprise becomes (negative) *evidential value*; namely, the expected likelihood that prior preferences will be realised following an action (I am grateful to Karl Friston for helping to construct this Figure, which aims to facilitate translation of the simplified version of FEP given in the main text into the formal, more comprehensive and precise treatments presented in, for example Friston (2010) and Friston and Stephan (2007))

rearrangement, free energy becomes complexity minus *accuracy*, that is, the probability of sensations—solicited by active states—under the recognition density (Friston 2010). This means that when free energy is minimised (and holding complexity fixed), then actions selectively sample sensations that are expected under the current recognition model. That is, by avoiding surprise, given an assumed model, the system is able to maintain an accurate and simple account of its sensory samples, arguably in accord with Occam's razor (and Jayne's maximum entropy principle). This aspect of FEP is known as *active inference* (Fig. 1 gives a more formal schematic of these constructs).

In summary, the nature of variational free energy means that surprise must decrease as free energy is minimised by action (i.e., active states). In contrast, perceptual inference and learning alone cannot decrease surprise directly because, even though such a system can revise its (Bayesian) beliefs in the light of surprising sensations, it cannot change the sensory evidence for its model of the world. The notion of active inference is crucial because, as seen in the conceptual analysis above, it is by selectively sampling unsurprising sensations that the system can be said to exist. Perceptual inference is, however, also crucial for existence. It underwrites a tight bound on surprise, ensuring evidence for the model used to generate expected sensations for active inference. Without such a tight bound, active inference would be without appropriate constraints.

With these conceptual and mathematical arguments, the 'free energy' part is finally introduced and FEP is established. Rather than talking about minimizing surprise given a model the system cannot directly know, we now get that "any self-organising system that is at equilibrium with its environment must minimize its free energy" because by minimising free energy, surprise is implicitly minimised. FEP moves a priori—via conceptual analysis and mathematics—from existence to notions of rationality (Bayesian inference) and epistemology (self-evidencing). As will be discussed in Sects. 3–5, this a priori aspect is central to how we should assess FEP.

3 The free energy principle and its falsifiable process theories

FEP says organisms "must" minimise free energy. Notice however that this is not the 'must' of laws of nature in a standard Humean sense, for it is not a universal generalisation based on a limited set of empirical regular observations. Rather, it is a 'must' of conceptual analysis and mathematics, for that is all that was needed to arrive at FEP. FEP is therefore rightly called a 'principle' rather than a law of nature.[10] It describes behaviour that systems may or may not conform to, but the principle itself is not beholden to empirical verification. This is similar to Hamilton's principle of stationary action, which is based on mathematics and describes behaviour that may or may not apply to any given system.[11] Finding objects that do not conform to Hamilton's

[10] This is not to suggest that one should believe FEP merely because it is in some sense 'mathematical' (though there is perhaps a sense in which mathematical proof should be believed). Rather the point is that, when investigating what reasons we have for believing FEP, we should look for mathematical (and conceptual) reasons, not empirical evidence.

[11] Hamilton's principle states that a mechanical system develops in time such that the integral of the difference between kinetic and potential energy is stationary. There is some debate about the epistemic status of Hamilton's principle (see, e.g., Smart and Thébault 2015; Stöltzner 2009); in unpublished work the latter authors have argued that a Humean about laws can place Hamilton's principle as the most fundamental law, making it essentially empirical rather than a priori). I think most standard descriptions are consistent with the reading that it is not a law of nature in the usual sense but a mathematical principle for understanding the dynamics of a physical system in terms of a variational problem, given information about the system and the forces acting on it. My appeal to Hamilton's principle is not intended to establish complete parity between it and FEP; it may very well be that the former is not driven by conceptual analysis in the way I have argued FEP is. I appeal to the principle here to indicate that there is precedence in science for considering something a principle, which systems may or may not conform with, rather than a law. Another question for further discussion is whether Newton's laws stand to Hamilton's principle as process theories like predictive coding stand to FEP. Note finally that there is deep affinity between FEP and Hamilton's principle, such

principle would not falsify it; similarly, FEP stands whether or not systems conform to it. (We can then note that FEP is, in and of itself, an evidential principle (in virtue of the fact that surprise is negative log model evidence), such that it describes—but is not subject to—the notion of empirical or evidential falsification).

Consider again the scepticism about FEP's falsifiability, mentioned in the Introduction. The worry is that since there is no empirical evidence for FEP, we should not believe it—what we cannot experimentally test in the lab, we should not believe. This worry is misplaced because FEP is a priori and not the kind of thing for which empirical evidence can be sought. The reasons for believing FEP pertain, rather, to grasping a particular concept of *existence of self-organising (nonequilibrium steady state) organisms*, together with the mathematics of variational calculus. This does not however imply that FEP is an uninformative platitude beyond criticism; it is possible to challenge the conceptual analysis of existence, and there could be shortcomings in the mathematical derivations and proofs from variational calculus—or in their application to this particular notion of existence (for discussion, see Colombo and Wright 2018).

Indeed, rather than being an empirical generalisation, FEP is more similar to a piece of *philosophical* reasoning. Just like a philosopher like Immanuel Kant (1787) might ask "what are the conditions for the possibility of perception?", FEP can be construed as an answer to the question "what are the conditions for the possibility of existence?" The answer is arrived at through a process that can reasonably be considered a priori, by consideration of our conceptual grasp of the concepts involved, together with mathematics.

This may appear to make matters worse since it is now unclear how, if at all, FEP connects to empirical science, and why we should bother believing it or not. If FEP is a free-floating mathematical-conceptual construct, then it seems science is free to dispense with it. Is its promised great explanatory scope voided by its non-empirical epistemic status?

It would be premature to side-line FEP as irrelevant for science simply because it is a priori. By that reasoning, Hamilton's principle should also be side-lined as irrelevant for science. But obviously it is not irrelevant; Hamilton's principle is a cornerstone of physics. So the issue needs to be considered in more depth.

The initial step is to recognise 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. Two process theories are prominent: for perceptual inference, predictive coding (Friston 2003, 2010), and for active inference, gradient descent on variational free energy (Friston et al. 2017). Predictive coding, for example, is a theory of perceptual inference—hypothesising that the brain minimises prediction error, such that it approximates hierarchical Bayesian inference over the long-term average. If the predictive coding scheme has precision-weighted prediction error minimisation then it conforms with FEP (given assumptions about the normality of probability density functions).[12] This theoretically and mathematically defined functional role for predic-

Footnote 11 continued
that FEP applies to systems for which Hamilton's Principle holds, hence, if the latter is not a priori then arguably FEP would not be either; for discussion of fundamental physics and FEP, see Friston (2019).

[12] Heuristically, if the underlying distribution is multimodal (i.e., non-normal, or not gaussian), then predictive coding can mischaracterise a given sample, which is close to one peak, as a large prediction error relative

tive coding is hypothesised to be realised through the familiar idea that the brain is hierarchically organised and passes predictive messages down the hierarchy and corresponding, precision-weighted prediction error messages upward. These ascending prediction errors lead to revision of each level's beliefs, in the light of the precision of the prediction errors (where precisions are tied to postsynaptic gain and plasticity). This is an empirical theory about brain function, which can be, and is being, tested in the lab (for a review, and a recent study, see Friston 2018; Stefanics et al. 2018). Evidence that fails to support the idea of precision-weighted prediction error minimisation is evidence against this type of predictive coding.

Process theories are not logically entailed by FEP. If they were, then FEP would be an empirical hypothesis since evidence against, for example, predictive coding would then be evidence against FEP. Rather, FEP can be considered a *regulatory* principle, "guiding" or "informing" the construction of process theories (e.g., Allen and Friston 2016). Several assumptions are needed to get from FEP to predictive coding. An initial, empirical assumption is "that the brain uses the [dimensionality-reducing factorisation resulting from the] mean-field approximation […] because it has evolved to exploit the ensuing computational efficiency" (Friston and Stephan 2007: p. 429; Parr et al. 2019; for discussion and anatomical considerations, see Schwöbel et al. 2018). And predictive coding relies on the Laplace assumption that posterior probability densities are normal (Gaussian), such that random fluctuations are dispersed about a single expectation and precludes posterior densities that are multi-modal. With this assumption in place the free energy can be viewed as the sum of long-term average prediction error, providing the link to FEP. These assumptions may or may not hold of a particular brain. This means that predictive coding is a hypothesis for any particular sort of brain, and is subject to falsification, for that particular sort of brain. The application of predictive coding schemes to the brain is guided by empirical beliefs about the hierarchical structure of the brain and the time scales of neuronal dynamics, and how these aspects of brain function map onto the message passing needed for perceptual inference and learning. Hence, FEP leads to the formal aspects of predictive coding only via certain assumptions, and predictive coding leads to testable hypotheses only via substantial empirical beliefs about brain function.[13]

Footnote 12 continued
to another peak; for discussion, see the Hierarchical Gaussian Filter developed in Mathys et al. (2014). For discussion of how this relates to empirical research on particular systems, such as human brains, see Friston (2009). In terms of evidence, Friston remarks that "there is no electrophysiological or psychophysical evidence to suggest that the brain can encode multimodal approximations: indeed, with ambiguous figures, the fact that percepts are bistable (as opposed to bimodal and stable) suggests the recognition density is unimodal" (2009: p. 298).

[13] Here the question of the scope of FEP is relevant. FEP is so general that it may apply to systems like single fat cells, which would not be sharing cognitive architecture with humans. The question what FEP implies for such systems then depends on the assumptions made for them. For FEP applied to really basic model systems, see Friston (2013). Interesting questions arise about the meaning of key notions, such as 'inference', 'representation', and 'computation' and how far and in what manner they might deviate from their literal senses, associated with symbolic representation etc. It seems likely that to get literal inference/computation/representation, we need to appeal to some subset of process theories and assumptions of particular systems of FEP, such as those arguably applying to humans. The thrust of my argument in this paper, to be unfolded in the next section, is that FEP entails approximation to Bayesian inference, and therefore a sense of normativity that seems relevant for what might be regarded basic notions of representation and misrepresentation (at least in the sense of genuine norm-violation). I think it is a substantial further

So, there is a possibility of obtaining empirical evidence that is in the vicinity of FEP. Still, the question remains: if much meaningful empirical science can be done on the process theories, why not dispense with the a priori FEP and just work on the process theories (see, e.g., Allen 2018)? Why should FEP guide theory construction? Indeed, versions of predictive coding existed before FEP was formulated, so it is possible to focus on such process theories rather than FEP (for review of predictive coding schemes, see Spratling 2017). In Sects. 4 and 5, I will build a case that FEP offers a solution to a foundational problem for any neurobiological, cognitive science, machine learning, or indeed philosophical, account of the mind, and that therefore FEP had better constrain process theories.

4 The problem of normativity

Recall Friston begins the argument for FEP by saying "A system cannot know whether its sensations are surprising". It is clear why finite, bounded systems confront this problem. It is natural to assume that solving this problem requires endowing systems with knowledge of their model, in order to allow them to adjudicate whether new sensations are surprising or not. However, FEP does not solve the problem by just endowing systems with knowledge of their own model. This is critical for appreciating FEP—and is expressed by the second half of Friston's formulation of the problem "A system cannot know whether its sensations are surprising *and could not avoid them even if it did know*" (Friston 2010: p. 128, emphasis added).

On a first interpretation, the problem Friston alerts to here is that simply endowing a system with a belief about its model does not explain how the system can also be a self-organising system capable of using that knowledge to actively avoid surprising states. The system also needs to generate—'read off'—the model's expectations and then act to maximise the probability it will sample just those expected states. As we saw above, this requires adding active states to the system, leading to active inference (that can then be associated with a process theory for active inference for inferring its own policies for minimising expected surprise in the long-term average (see Friston et al. (2017)).

However, deeper problems emerge from reflection on what it would mean to "endow" the system with knowledge of its own model. One approach for how to give the system knowledge of its model would be to present the system with labelled training data, informing it about which sensations are expected. This would not be adequate as it amounts to just exposing the system to more sensations (in the shape of training data) to assess for surprise, when the problem in the first place was to help it assess sensations. Put differently, the problem cannot be overcome by appeal to traditional supervised learning.

Footnote 13 continued
discussion how far and in what way the notions 'inference' and 'representation' used here deviate from literal (e.g., symbolic) inference and representation. My view is that, in so far as FEP ensures normativity, the use of those notions is justified to all systems for which FEP applies, since we often cash out those notions precisely in terms of normativity. I do also think it is likely that FEP will eventually lead to a recalibration of what we mean by 'inference' and 'representation'. However, the issues here are substantial and some aspects will need to be developed in subsequent research.

An alternative strategy could be to load information, structured in some way and embedded in some appropriate representational (neuronal) vehicle, straight into some storage circuit of the system from where the system can read out what is surprising and what not. To distinguish this strategy from the previous supervised learning strategy, this information cannot be treated as sensory input or as labelled in some way. Instead it must be treated as simply part of the causal structure of the system, that is, part of what causes its actual behaviour—in particular, part of what causes it to treat some sensation as surprising or not.

The problem with this strategy is that we can no longer distinguish between correct and incorrect application of its model. I will give a simple example to illustrate. Assume the system has a history of reacting to a certain constellation of input, in fact caused by dogs, with the inference "dog". From the outside, we would say it has a concept *dog* that applies to all and only dogs. On a new occasion, it reacts to a different input, in fact caused by a sheep, with the same inference, "dog". We want to say this is an incorrect inference. But we cannot rule out another interpretation, on which this is actually a correct inference according to the disjunctive concept *dog-or-sheep*. That is, perhaps the system always expressed this disjunctive concept with the inference "dog". Both interpretations are equally consistent with the behaviour, and we cannot appeal to a causal description of the system's internal, causal structure to determine which is right.

If there is no way to distinguish between correct and incorrect inference, then we undermine the *normative* aspect of what it is to be a model or representation of some state of affairs. Something can only be a model or representation if there are some satisfaction conditions for its application, which can in principle be violated. In other words, there has to be scope for a distinction between what the system *would* do, given its causal dispositions, and what it *should* do, given the representational content of its model. Under the causal strategy considered here, this distinction between 'would' and 'should' seems to disappear.

The simple example of "dog" repeats easily for other kinds of cases that involve conditions of satisfaction, truth, correctness or projection. For example, a typical example of self-supervised learning (in the machine learning sense) is context prediction where the relative positioning of a number of image patches is learnt after which one patch is sampled and the content of another is predicted and sampled. To illustrate, for an image of a dog, a context rule would be "if nose is sampled, then ear in upper right patch". In this case, the question is what constitutes error with respect to this prediction, ruling out disjunctive or disordered context representation. Another example would speak to the homeostatic grounding of organism existence, such as heart rate, glucose levels, body temperature; homeostatic belief might be a certain set point for heart rate and a corresponding policy of avoiding action that increases the rate above this point. The question here is what constitutes violation of the policy, ruling out that higher heart rates are surprising to the agent—what fact about the agent rules out "disjunctive beliefs" about heart rates?

A similar situation arises in Bayesian decision theory, where it is known as the complete class theorem, which means that the observed behaviour of any system—for any specified loss function—can be construed as Bayes optimal under some priors. Conversely, the observed behaviour of any system—for some specified priors—can be construed as Bayes optimal under some loss function (Brown 1981; Friston 2011). There is no way of separately identifying the prior and loss function; other than observing the system's behaviour. This means that there is a duality that is resolved in active inference by making the loss function and priors the same thing; that is, absorbing cost functions into prior beliefs. Apparent deviation from a past pattern of decisions is then best explained as a change of prior belief, ensuring the optimality of the decision but seemingly obviating the idea of incorrect or suboptimal decision.

Overall, this creates a dilemma for how the system can know its model. Either, information about the model is given as additional sensations, which pushes the problem back to the question of assessing the surprise of those sensations. Or, the information is hardwired into the causal structure of the system, which undermines the normative aspect of what it is to be a model of something.

As rehearsed here, the problem about how a system can know its model harks back to important foundational debates, which occur across cognitive science, machine learning and neurobiology, as well as philosophy of language and mind.

A central quest in cognitive science and machine learning is identifying the principles for *self-supervised* (or unsupervised) rather than *supervised* learning and inference.[14] If the system in question is self-supervised, then it must infer and learn from unlabelled sensory input. The difficulty here is to build a firm standard against which the system, by its own lights, can adjudicate if the categorisation and inference is in accordance with the encoded representation or not. If the system is supervised, then its learning samples are labelled by an external entity such as a programmer or teacher. The problem here is that the system's understanding of correct and incorrect application is parasitic on the understanding presupposed within the supervisor; hence there is no explanatory progress on the fundamental issue of how it came to learn and infer. The problem with which FEP begins is thus also a foundational problem for understanding or building self-supervised, truly intelligent systems. In cognitive science and theoretical neurobiology, a version of this problem arises when trying to describe cognitive function and behaviour from the perspective of the organism, rather than with the benefit of an all-knowing external observer's perspective. As we will see next, FEP then suggests that the problem of self-supervision is linked to the concepts of self-organisation and self-evidencing, in other words, that self-organisation is intrinsically normative.

In philosophy of language and mind, a version of this problem arises in discussions of the foundations of linguistic or mental *content*. The example given above, about

[14] I am here using these terms in a fairly generic, philosophical sense. In the fields of cognitive science, machine learning, and statistical learning, there is substantial treatment of the issue of supervision, using somewhat different understandings of the notion of supervision. In machine learning approaches, there are many unsupervised algorithms and many things that organisms do that involve supervised learning (including supervision by nature). In philosophy, supervised learning raises foundational problems of normativity, essentially related to the learner's grasp of the labels, which must be considered before supervised learning can truly be understood. I am here implying that self-supervised learning is, or should be, a (or perhaps the) quest of machine learning. Of course, valuable machine learning advances can come from devising robust supervised learning algorithms, but from a philosophical perspective, machine learning will only throw light on human intelligence (or approximate human intelligence) if it begins from a basis of self-supervision. This claim is based on the observation, versions of which stretch all the way back to Kant and beyond, that human intelligence must come about just by relying on sensory input and prior belief. See also footnote 1 for comments on my use of the notion of self-supervision.

ruling out odd concepts like *dog-or-sheep*, is a version of the *disjunction problem* (for review, see Adams and Aizawa 2017; for a review of the concept of normativity, see Glüer and Wikforss 2018). These are foundational problems that relate to Ludwig Wittgenstein's so-called *rule-following considerations*, where the notion of following a rule expresses the idea of applying concepts or words according to their content or meaning. Wittgenstein makes normativity central to a description of how content, and any kind of meaning, can be explained; as he famously remarks, "This was our paradox: no course of action could be determined by a rule, because any course of action can be made out to accord with the rule" (Wittgenstein 1953: §201). In Saul Kripke's influential interpretation of Wittgenstein (Kripke 1982), the problem is expressed in terms of *meaning-skepticism*—an inability to give a foundational, wholly naturalistic account of meaning (i.e., an account of the normativity of meaning (or content, representation, or belief) that does not rely on normative elements). The Kripke–Wittgenstein rule-following considerations are thus mirrored in the problems just rehearsed for how a system can come to know its own model.

What FEP needs to deliver is then a solution to some of the most recalcitrant problems across neurobiology, cognitive science, machine learning, and philosophy of language and mind. As I will argue now, FEP allows a new way to comprehend how a causal account of what the system *would* do can also be a normative account of what it *should* do.

5 The free energy principle's solution to the problem of normativity

All the elements for FEP's solution to the problem of normativity have been laid out earlier. The train of reasoning begins with the assumption that the system in question exists. For nonequilibrium steady-state systems to exist is for them to selectively sample low surprise states, that is, according to FEP, sample sensations that minimise free energy, or, as per active inference, the free energy expected following an action (Friston et al. 2015). A recognition model, q, is selected and brought to approximate p, the true posterior. The system here gets 'to know' its model in the sense that q's implicit inversion of the generative model amounts to the formation of accurate, evidenced beliefs about the causes of sensations. These beliefs encode the expected states of the system and thus the surprise of sensations. Under FEP, existence thus implies knowing the model.[15]

The processes that minimise free energy for a system such as a brain in perceptual and active inference can be characterised in causal terms—what the system would do. Given the system exists, there is a set of sensations that its causal states will make it sample. Certain sensations will cause certain changes in its internal states, which again

[15] Here and elsewhere in the paper, various typically personal-level terms are used ('know', 'believe', 'evidence' etc.). This is not to imply that these are personal level rather than subpersonal level processes or states. I am agnostic on how to draw that boundary and here simply use these terms more or less like they are used in the wider literature and in textbooks on machine learning and statistical learning. 'Knowing' is the appropriate notion to use here because the reasoning behind FEP leads to the idea that the model is inferred (and what a system infers it in some sense knows). There is a substantial, different debate to be had about the sense in which 'approximate inference' is 'inference', related to these issues, which is beyond the scope of this paper.

cause certain changes in its active states, in turn causing actions in the environment that lead to new, mostly unsurprising changes in its sensations. If, in contrast, those actions lead to surprising states, then the system is ceasing to exist.

The crucial task is then to show that what the system would in this causal manner do is also in an appropriate sense what it should do. The ‘should’ here will appeal to the epistemic ramifications of FEP described earlier: the self-evidencing that minimizes free energy and thereby characterises existence and that, critically for the present argument, at the same time approximates Bayesian inference in the long-term average (i.e., minimizes KL($q||p$)). This is an appropriate sense of what the system *should* do because Bayes’ rule can be regarded a paradigm of normativity: it prescribes optimal relative weighting of evidence and prior belief. Violations of the norm occur when too much or too little weight is given to the prior or to the evidence, leading to false inference. It follows that the internal belief revision and ensuing selective sampling that happens under FEP is normative, in this Bayesian sense—it approximates the optimal results a system would get by complying with the Bayesian norm.[16]

This normative FEP process is not afflicted by the problem of normativity. Exact inference is given by p, which specifies the satisfaction conditions for inferring causes (i.e., what evidence is needed). But, crucially, the system can use q to approximate p even though it is blind to p. This is the mathematical beauty of FEP’s use of the lower bound on surprise furnished by free energy: the KL-divergence decreases just by operations on q. So the normative process does not import normativity into the story by presupposing that the system already grasps the content of p. What is left is a purely causal story about what the system would do, given its causal set-up (its phenotype). This causal story can, for example, be cashed out in terms of variational inference where the system essentially varies the parameters of q one at the time, which is the way in which it minimises free energy (or long-term average of prediction error; for different neuronal inference schemes, see Parr et al. (2019)). Normative, self-supervised inference and learning then flows from the self-evidencing of the model’s beliefs. The constraints on this process come only from the system itself, through its self-organisation, or maintenance of its own boundaries over time, that is, its existence. This establishes the link from self-organisation to self-supervision, via the notion of self-evidencing.[17] With this, FEP provides a new type of account of self-supervised learning, including the problem of content in philosophy of mind and language (see Hohwy 2013: p. 181 for an earlier version of this view; see Kiefer and Hohwy 2018 for further discussion of normativity and the KL-divergence; see Piekarski 2019 for a discussion of normativity of prediction within a predictive processing framework).

The argument spells out ‘rule-following’ in terms of approximate inference, which is based on free energy minimization. One question here is if free energy minimization actually requires computation over probability distributions. Much here hangs on the appeal to approximate rather than exact inference. Exact inference would require

[16] I am not here providing a foundational defence of Bayesianism as such; I am relying on the fact that FEP implies an approximation to the exact Bayesian posterior, which is a good candidate for being a paradigm of normativity. For discussion of Bayesian optimality, see Rahnev and Denison (2018).

[17] There have been previous suggestions linking self-organisation and self-supervision. Ashby, for example, argued that systems can be both self-organised and also display determinate behaviour (Ashby 1954). The current proposal makes this link via the notion of self-evidencing inherent in FEP.

computation over probability distributions but it is less clear that approximate inference does, given it only relies on free energy minimization (or long-term average prediction error minimization). As indicated earlier, I think a system can engage in free energy minimization without explicit computation over probability distributions (e.g., the causal processes involved in organising its internal states to anticipate sensory input). The key point is that approximate inference is not just like exact inference except with approximate values, rather the ‘approximation’ is the minimization of the KL-divergence: an approximation of the states of the system to the states that it would have if it were indeed computing exact inference. This means that it is correct to say that such a system is literally doing ‘approximate inference’ even if it does not literally do ‘inference’ in the sense of ‘computations over probability distributions’. In sum, the point is to show that the approximation to exact inference ensures a normative dimension to the causal processes involved in the free energy minimization that underwrites self-organisation of the system.[18]

One might immediately object that here the system is following the basic-level rule “minimise free energy!” (or, “approximate Bayes!”), which is just as problematic as following any other rule: it also has satisfaction conditions, which are equally subject to the problem of normativity. But FEP’s solution undercuts the problem for this basic-level ‘rule’ by noting that ‘following’ this particular rule is equivalent to existence. In this way, the solution offered by FEP is not an ill-fated reduction of the ‘should’ to the ‘would’, or a mere exercise in presupposing normative elements. Rather, FEP’s conceptual analysis allows us to see how existence (analysed as self-organisation) is at the same time both causal and normative. In a sense, it is a reduction of the ‘would’ and the ‘should’ jointly to the ‘is’ of existence.

Returning to the philosophical debate about the rule-following considerations, FEP can then be taken to provide a new and formal interpretation of Wittgenstein’s gnomic observation that “If I have exhausted the justifications [for how I follow a rule], I have reached bedrock and my spade is turned. Then I am inclined to say: ‘This is simply what I do’.” (Wittgenstein 1953: §217). In the optics of FEP, when Wittgenstein exclaims “This is simply what I do” he expresses the notion of existence equivalent to selectively

[18] In this paper, the focus is on self-organising systems, which is what FEP is formulated for. Such systems can act to maintain themselves in their expected states. There are some very substantial questions about where to put the boundary between self-organised systems in this sense and any other system in the broadest sense (e.g., mechanical systems, or any system indeed that physics can describe). In some iterations, the notion of free energy minimization is so general that it literally applies to every thing (Friston 2019); (see also fn 11). In other words, something is needed to distinguish mere causal mechanisms from self-organising biological organisms (and to distinguish between things and non-things). One distinction is between things that can model expected free energy and infer policies on this basis in order to engage in active self-evidencing, and things that cannot. Mechanical systems cannot do this, if their action repertoire is pre-set (by a designer who has performed the active inference for the mechanism, for example; for self-supervised artificial intelligence mechanisms, the discussion veers into the substance of this paper). Hence, the question how the system can minimize surprise, if it can’t know its model a priori, pertains to self-organising systems conceived in this way. Further research is needed to fully discuss the question what if anything FEP and kindred approaches imply about non-self-organising systems. In particular, there is the question what, if anything, the notions of perceptual and active inference come to in these kinds of cases where it is less clear that they apply; for example, a projectile is described by Hamilton’s principle but does not, in any sense of the word, “compute” its stationary points of action.

sampling a limited set of unsurprising states, given a model. With FEP, we can see that this is normative in the sense that it approximates Bayes.[19]

The FEP-solution to the problem of normativity does not imply that *everything* the system actually does at any given instance is also what it should do. This is important because the crux of the problem is that it must be possible in principle for the system to make mistakes in perceptual and active inference, relative to the given norm. The key element here is that free energy must decrease on average over time, but that this does not disallow transient increases.[20] The free energy trajectory will fluctuate around exact Bayesian inference and only over time and in stable situations will it get closer and closer to it (for review, see Bogacz 2017). The parameters of the recognition model will not correspond to the exact model's parameters at the outset, and the system will try different settings for them as it finesses its model (cf. variational inference and learning). In the course of this, what the system actually does at any given time can deviate from what the norms of exact inference would dictate at that time. In a toy example, even if the system should infer "dog" for all and only dogs, and its causal set-up makes this what it would do (in the sense of its free energy minimisation), it is possible for the system in a given situation to actually infer "dog" upon being confronted with a sheep—it can engage in false inference in the sense of inference that in fact violates the norms of exact inference. There is a further question here about veridicality: does approximation to exact inference guarantee that the recognition model's representations are truthful, and is violation of the norm the same as misrepresentation? An affirmative answer would not be unreasonable here, since exact Bayesian inference is likely to lead to truthful representation. But for the purposes of this paper, we can be agnostic about that step of the overall account; the main focus here is on the possibility of normativity, for which 'misrepresentation' in the sense of 'norm-violation' suffices (see discussion, and an initial framework for thinking about this issue in Kiefer and Hohwy (2018)).[21]

[19] These comments may need some clarification, to set them in the context of computational neuroscience and machine learning approaches. The argument here is not that, a priori, all 'bottom-up' approaches fall short of converging on Bayes' rule; and there are bottom-up learning methods (from single-layer perceptrons onwards) for which there are convergence proofs. Rather, the argument is that some of these bottom-up methods rest on supervised learning (e.g., perceptrons), which raises the problem of normativity in focus here, and suggesting they do not conform to FEP. If the bottom-up method does in fact conform to FEP, then it can potentially form a process theory (subject to assumptions and anatomical plausibility) for which the problem of normativity will not arise, and thereby a good starting point for the quest for truly self-supervised learning. In this light, convergence does not suffice for normativity because the underlying process should also be self-supervised in a manner that does not evoke the problem of normativity; conformity to FEP demonstrates that both constraints are satisfied. When it comes to the 'top-down' approaches more commonly endorsed by FEP, the claim is that they are normative in the sense of converging on Bayes; a good place to explore the mathematical grounds for this claim is in discussion of the Hierarchical Gaussian Filter (Mathys et al. 2011) that sets precision-weighted prediction error minimization in the context of approximate inference and reinforcement learning (with a dynamic learning rate).

[20] The time scale over which free energy is assessed is important. Here I just consider it the appropriate time scale for the organism in question but there is a substantial further question to address here. The critical point here is that approximation is not instantaneous.

[21] An interesting question is whether this approach will outlaw "weird" concepts like, indeed, a regular concept *dog-or-sheep*, which itself has satisfaction conditions. FEP could allow such concepts if they had conceivable free energy minimization properties. If not, and if such concepts are regarded as *bona fide* concepts nevertheless, then FEP would only be a partial solution. The example given here is semantic but

Notice finally that this account is distinct from any *instrumentalist* or *teleosemantic* notion, which would be the idea that systems minimize surprise in order to achieve the (known) goal of continued existence (or surviving, or achieving intermediate goals and rewards); put differently, the account is focused on letting systems infer their own goals, which cannot be known a priori in the sense presupposed by teleosemantics.[22] The account is also distinct from notions on which some systems have an *evolved* capacity for minimising surprise. Such accounts contradict the conceptual analysis above because they leave it conceivable that self-organising systems could exist without minimising surprise and self-evidencing. The FEP account is fundamentally different from such other attempts at explaining meaning and content because the first step in its explanation is to *analyse* existence in terms of surprise minimization, rather than naturalistically explain one by appeal to the other.

In sum, the system gets to know the model, and normativity emerges within the system's causally described behaviour, based on its existence. FEP is therefore a compelling solution to the problem of how a system can know its own model, which speaks to foundational issues about self-supervised inference and learning, and the possibility of meaning and content.

6 The philosophy of FEP

At the end of Sect. 3, the question arose why we should not just dispense with FEP and instead work only with distinct process theories. We are now in a position to offer an answer.

Earlier, I observed that FEP is much like a piece of philosophical reasoning, using conceptual analysis (together with mathematics) to elucidate the conditions for the possibility of existence. Through the discussion of FEP's solution to the problem of normativity, FEP can be seen to rather uniquely tie existence to normativity.

Footnote 21 continued
the account extends to action, that is, the problem of inferring a specific policy, for example, for avoiding an environment that is too cold. Active inference sets out how expected free energy is minimised, given an internal model of the environment's states (including the agent's own states). That is, a precise policy for action (conceived as a series of control states) is inferred, which is expected to maintain the system in its expected (not too cold) states (e.g., putting on a coat). The system can rank and execute specific policies but only given an internal model governed by FEP, which cashes out normativity by allowing a KL-divergence to be minimised between states achievable given a policy and states the system expects to occupy. Policies in active inference (for expected surprise) are then like priors in perceptual inference (for actual surprise). A policy is inferred by assessing the surprise expected under different policies. On the basis of the model's inferred policy, an expectation of sensory input is generated, which is minimized through action. As such, policies are part of the generative model and help specify the expected states of the system. The cost function itself is absorbed into the priors, and policies can also be updated and there can be model selection (based on complexity considerations/Bayesian model evidence). In this sense, just as a system's internal model can have a fine-grained set of priors that describe its beliefs about the world, it can have a fine-grained set of policies that describe beliefs about how it acts in the world. Active inference thus furnishes an answer to challenges about decision-making and inference of specific policies, such as (Klein 2016).

[22] There is interesting discussion from the dynamical systems perspective, but kindred to FEP topics, in Bickhard (2009).

In this sense, FEP is on a par with standard philosophical analyses such as, for example, the *functionalist* analysis of mental states in terms of input–output and internal states; the *liberal* analysis of fairness as the distribution of resources selected behind the veil of ignorance; or the *compatibilist* analysis of free will in terms of unimpeded actions. There are rather divergent opinions about the role and worth of such philosophical, conceptual analyses, but they can arguably function as important *regulatory principles* for science. For example, compatibilism about free will is the position that free will and determinism are compatible, based on the conceptual analysis of free will, which says that to act freely is to act in accordance with one's own causally determined beliefs and desires. This can serve as a regulatory principle for the neuroscience of free will, by ruling out experimental approaches that operate with competing, incompatibilist analyses (this is, for example, relevant for determining if Libet's famous experiments of willed action are relevant for free will (Libet 1985; Libet et al. 1983)).

This is analogous to how FEP should be considered: as a regulatory principle for the distinct process theories explaining perception, action, attention and other mental phenomena. We should believe FEP if we believe it offers a cogent analysis of the concept of existence of nonequilibrium steady-state systems (and to the extent we are convinced by the soundness of the underlying mathematics). But, in particular, if what I have argued above is correct, we should believe FEP because adopting it as a regulatory principle for process theories would ensure that those process theories can implement a new type of solution to the problem of normativity—it ensures that a causal process of self-evidencing can also be conceived as a normative process. Thus, the argument is that we are better off pursuing process theories that conform to FEP because conformity is sufficient to ensure those process theories allow a distinction between correctness and incorrectness.[23] For example, a pure bottom-up process theory of perception in the human brain, according to which the sensory signal is progressively decoded rather than being assessed relative to an internal generative model will likely not conform to FEP (modulo the assumptions for a type of particular system) and for such a system it will then be an open question if it can itself distinguish correct and incorrect representation, that is, it does not benefit from the assurance with respect to self-supervision that derives from conformity to FEP (of course, there might be an alternative to FEP that also confers such benefits).[24] Similarly, process theories

[23] An interesting question arises here about artificial systems running active inference algorithms but which are not easily seen as self-organised or autonomous systems: do they display normativity? A full answer is beyond the scope of this paper partly because it touches on issues around emulation, which may undermine true self-supervision. It may be that some artificial systems can be considered truly normative, in the FEP sense, and therefore self-organising.

[24] Notice that here care is taken not to imply that FEP itself implies cognitive architecture. Notions of architecture will need to build on assumptions about the particular system in question, which will constrain processes for message passing structure. It is a topic for further discussion how assumptions play this role, and what assumptions may look like in various non-human systems. This invokes a larger issue in philosophy of science concerning the ways in which principles constrain process theories (or laws). For FEP, the starting point for this issue is the idea of the addition of assumptions about particular systems to the principle. However, further discussion is needed of what this exactly means: is the relation something on a spectrum between derivation (given assumptions) and more informal notions of coherence, for example? There is a significant body of literature on these questions, which is still to make contact with the specific status of principles versus process theories (see, e.g., Craver 2005; Zednik and Jäkel 2016).

that rely on supervised learning will not conform to FEP, since the objective function guiding the system will not be limited to the free energy but in addition quantities the system must be gifted by the supervisor.[25]

From the perspective of philosophy of language and mind, we should be anxious to adopt a regulatory principle, like FEP, which helps us avoid process theories that lead to the problem of normativity. Theories afflicted by this problem cannot help explain why perception, action and communication, and other mental processes, are meaningful. So, if we are taken by the problem of normativity, then we would do well to adopt FEP as a regulatory principle for the formulation of our process theories. This is a pragmatic argument for FEP, not conclusive proof. I am saying that, given compelling constraints on the types of explanations cognitive science pursues, we better believe the free energy principle.

To situate FEP as a compelling regulatory principle for cognitive science, machine learning and theoretical neurobiology, a brief final comment may be helpful. At times, the alleged problem that there is no evidence for FEP is dealt with by aligning FEP with the theory of evolution by natural selection (e.g., Hohwy 2015). This theory is sometimes (controversially) said to be unfalsifiable because at the highest level of abstraction, the theory seems to imply the platitude that to be fit is to reproduce is to be fit. This comparison then implies that FEP is formulated at such a level of abstraction from actual self-organising systems that it is also a platitude. The discussion above suggests otherwise: FEP is a far-from-trivial piece of conceptual-mathematical analysis providing a new conception self-organising systems by using the notion of self-evidencing to relate existence to normativity and self-supervision. The theory of evolution by natural selection in contrast seems like a causal mechanism sketch that abstracts away from the actual processes of natural selection in actual biological organisms. As such, that theory would be falsified if there was no evidence of natural selection in concrete biological organisms (e.g., if, contrary to actual fact, there were evidence of creationism). Of course, there are mountains of evidence for natural selection and by extension for the theory of evolution. In contrast, no empirical evidence can undermine FEP, even if evidence could undermine process theories that conform to FEP. FEP is therefore more similar to Hamilton's principle and to philosophical reasoning than to the theory of evolution by natural selection.

This line of reasoning suggests that other naturalistic approaches than those that conform to FEP fall short of explaining inference and learning (e.g., teleosemantics, briefly discussed in Sect. 5). To substantiate this suggestion, a full comparative treatment of alternatives is need, which is beyond the scope of the current paper (see Kiefer and Hohwy 2018 for some initial discussion). For present purposes, there are three relevant observations: first, other naturalistic approaches to normativity, whether predating FEP or not, may or may not conform to it; my suggestion is that conforming to FEP is an important yardstick for assessing such theories since FEP provides (I have argued) assurance that the particular systems in question are both normative and

[25] There are many proposals for unsupervised learning, which are not designed to conform to FEP (e.g., Zheng et al. 2018). The claim is not that such approaches cannot deliver what they promise. It may be that they are in fact conforming to FEP, or they may establish normativity in their own right (perhaps conforming to some other principle). Here, the focus is FEP and the claim is that it has a philosophically appealing approach to normativity in terms of self-supervised (as that term is used here).

self-supervised. That is, there is a valuable research project in assessing whether such naturalistic approaches, conceived as process theories, conform to FEP or not. There are also alternative accounts of self-organising systems, which do not seem to rely on FEP (and which do not obviously provide solutions to the problem of normativity) (Kauffman 2019; Varela et al. 1974); such theories may or may not conform to FEP and thus may or may not describe self-supervised systems in a way that satisfies the constraints in the philosophical debate on normativity. Second, there are accounts aligned with predictive processing that pre-date FEP (see Sect. 1). Clearly, consulting FEP was not vital for guiding the construction of these earlier accounts. However, the key question is to consider them candidate process theories and analyse if they, post hoc, can be seen to conform to FEP. Third, it does not follow from my argument so far that FEP is the only principle that could ensure self-supervision. Perhaps there are other such principles, however, I am not aware of any competitors developed with a similar level of conceptual and formal depth and detail.

7 Concluding remarks

The free energy principle, FEP, implies that perception, action and attention, and other mental phenomena all arise in the minimization of free energy that characterises self-organising (nonequilibrium steady-state) systems. FEP may seem infeasibly ambitious, and appears beyond empirical confirmation, and so it can seem tempting to treat FEP as a curiosum that does not and should not inform the search for less abstract process theories of mental phenomena. On this reasoning, we are not at all compelled to believe FEP.

However, FEP is best understood as a piece of a priori philosophical-mathematical reasoning. This means it is misguided to require empirical evidence for it. It is an important principle because, like some other philosophical principles, it can work as a regulatory principle in the construction and exploration of worthwhile process theories.

FEP is, I argued, a substantial and new type of solution to the problem of normativity concerning how machines and organisms can infer a model of themselves and their world and apply them in their behaviour in a rule-governed, autonomous manner. This speaks directly to the quest for the principles of self-supervised inference and learning, as well as debates about the nature of content and meaning in philosophy of language and mind. As a regulatory principle for process theories, FEP would therefore help ensure efforts are focused on process theories that would have a good chance of describing biologically plausible and philosophically tenable accounts of human and machine perception, attention, decision-making and action. Process theories that are not guided by FEP risk accruing serious philosophical problems, making a mystery of the existence of contentful mental states.

Many older and several contemporary approaches to mind and cognition now emphasise or allow predictive processing (PP, mentioned in the Introduction) in some guise or other. Most of these either predate FEP (Ashby 1947; Helmholtz 1867), disavow or criticise it (Block 2015; Colombo and Wright 2018), or are somewhat agnostic (Clark 2016). These PP approaches are instead seen as constellations of process theories that merely happen to have predictive processing as a characteristic (e.g., as an evolutionary solution to processing bottle-necks), and which can in principle be combined with non-predictive processes (for example, for valence and emotion). These more ecumenical PP approaches, which are not adhering to FEP as a regulatory principle, are at a disadvantage because they will be at risk of the problem of normativity. Adherents of predictive processing may thus benefit from working under the regulatory principle given by FEP.

The solution to the problem of normativity offered by FEP is new and attractive. It rests just on conceptual analysis of the concept of existence (self-organising systems persisting in dynamic environments) together with the mathematics of variational calculus. In an a priori manner, FEP bases what systems would do and what they should do on the assumption of existence, while allowing the possibility of misrepresentation (or norm-violation). Thereby it can use the notion of self-evidencing as a bridge from self-organisation to self-supervised inference and learning, avoiding the meaning-sceptical problem of normativity.

Acknowledgements Thank you to Stephen Gadsby, Karl Friston and Thomas Parr for comments and suggestions on earlier versions; and to anonymous reviewers for several suggestions. Thank you for comments and suggestions to participants in workshops at Macquarie University and University of Cambridge, including Dan Williams who organised the latter, where some of this material was presented. Thank you to the members of the Cognition and Philosophy Lab for comments. This research is supported by the Australian Research Council DP190101805.

References

Adams, F., & Aizawa, K. (2017). Causal theories of mental content. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (summer 2017 edition)*. https://plato.stanford.edu/archives/sum2017/entries/content-causal/.

Akaike, H. (1974). *A new look at the statistical model identification*. Paper presented at the IEEE Transactions on Automatic Control.

Allen, M. (2018). The foundation: Mechanism, prediction, and falsification in Bayesian enactivism: Comment on “Answering Schrödinger’s question: A free-energy formulation” by Maxwell James Désormeau Ramstead et al. *Physics of Life Reviews, 24,* 17–20. https://doi.org/10.1016/j.plrev.2018.01.007.

Allen, M., & Friston, K. J. (2016). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, *195*, 2459–2482. https://doi.org/10.1007/s11229-016-1288-5.

Ashby, W. R. (1947). Principles of the self-organizing dynamic system. *The Journal of General Psychology, 37*(2), 125–128. https://doi.org/10.1080/00221309.1947.9918144.

Ashby, W. R. (1954). *Design for a brain*. New York: Wiley.

Bar, M. (2011). *Predictions in the brain: Using our past to generate a future*. Oxford: Oxford University Press.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation, 1*(3), 295–311. https://doi.org/10.1162/neco.1989.1.3.295.

Bickhard, M. H. (2009). The biological foundations of cognitive science. *New Ideas in Psychology, 27*(1), 75–84. https://doi.org/10.1016/j.newideapsych.2008.04.001.

Bickhard, M. H. (2016). The anticipatory brain: Two approaches. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 261–283). Cham: Springer International Publishing.

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

Block, N. (2015). The puzzle of perceptual precision. In T. Metzinger & J. M. Windt (Eds.), *Open MIND*. Frankfurt am Main: MIND Group.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology, 76*(Part B), 198–211. https://doi.org/10.1016/j.jmp.2015.11.003.

Brown, L. D. (1981). A complete class theorem for statistical problems with finite sample spaces. *The Annals of Statistics, 9*(6), 1289–1300.

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. https://doi.org/10.1016/j.jmp.2017.09.004.

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

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York: Oxford University Press.

Colombo, M., & Wright, C. (2018). First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese*. https://doi.org/10.1007/s11229-018-01932-w.

Constant, A., Ramstead, M. J. D., Veissière, S. P. L., Campbell, J. O., & Friston, K. J. (2018). A variational approach to niche construction. *Journal of the Royal Society Interface*, *15*, 141. https://doi.org/10.1098/rsif.2017.0685.

Craver, C. F. (2005). Beyond reduction: Mechanisms, multifield integration and the unity of neuroscience. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(2), 373.

Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, *16*(9), 1325–1352.

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. (2011). What is optimal about motor control? *Neuron, 72*(3), 488–498. https://doi.org/10.1016/j.neuron.2011.10.018.

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

Friston, K. (2018). Does predictive coding have a future? *Nature Neuroscience*. https://doi.org/10.1038/s41593-018-0200-7.

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

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214. https://doi.org/10.1080/17588928.2015.1020053.

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

Glüer, K., & Wikforss, Å. (2018). The normativity of meaning and content. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Spring 2018 Edition ed.).* https://plato.stanford.edu/archives/spr2018/entries/meaning-normativity/.

Gregory, R. L. (1968). Perceptual illusions and brain models. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 171,* 179–196.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B, 290,* 181–197.

Heeger, D. J. (2017). Theory of cortical function. *Proceedings of the National Academy of Sciences, 114*(8), 1773–1782. https://doi.org/10.1073/pnas.1619788114.

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

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

Hohwy, J. (2015). The neural organ explains the mind. In T. K. Metzinger & J. M. Windt (Eds.), *Open MIND*. Frankfurt am Main: MIND Group.

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

Jackson, F. (1998). *From metaphysics to ethics*. Oxford: Oxford University Press.

Kant, I. (1787). *Kritik der reinen Vernunft*. In Königlichen Preußischen Akademie der Wissenschaften (Ed.), *1900–, Kants gesammelte Schriften*. Berlin: Georg Reimer.

Kauffman, S. (2019). *A world beyond physics: the emergence and evolution of life*. New York: Oxford University Press.

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese, 195*(6), 2387–2415. https://doi.org/10.1007/s11229-017-1435-7.

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. https://doi.org/10.1098/rsif.2017.0792.

Klein, C. (2016). What do predictive coders want? *Synthese, 195*(6), 2541–2557. https://doi.org/10.1007/s11229-016-1250-6.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719.

Kripke, S. (1982). *Wittgenstein on rules and private language*. Oxford: Oxford University Press.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *The Behavioral and Brain Sciences, 8,* 529–566.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious initiation of a freely voluntary act. *Brain, 106,* 623–642.

MacKay, D. M. C. (1956). The epistemological problem for automata. In C. Shannon & J. McCarthy (Eds.), *Automata studies* (pp. 235–251). Princeton, NJ: Princeton University Press.

Mathys, C., Daunizeau, J., Friston, K., & Stephan, K. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2011.00039.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., et al. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2014.00825.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Nicolis, G., & Prigogine, I. (1977). *Self-organization in non-equilibrium systems*. New York: Wiley.

Parr, T., Markovic, D., Kiebel, S. J., & Friston, K. J. (2019). Neuronal message passing using Mean-field, Bethe, and Marginal approximations. *Scientific Reports, 9*(1), 1889. https://doi.org/10.1038/s41598-018-38246-3.

Piekarski, M. (2019). Normativity of predictions: A new research perspective. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.01710.

Prigogine, I., & Nicolis, G. (1971). Biological order, structure and instabilities. *Quarterly Reviews of Biophysics, 4*(2–3), 107–148. https://doi.org/10.1017/S0033583500000615.

Rahnev, D., & Denison, R. N. (2018). Behavior is sensible but not globally optimal: Seeking common ground in the optimality debate. *Behavioral and Brain Sciences, 41,* e251. https://doi.org/10.1017/S0140525X18002121.

Schrödinger, E. (1944). *What is life?*. Cambridge: Cambridge University Press.

Schwöbel, S., Kiebel, S., & Marković, D. (2018). Active inference, belief propagation, and the bethe approximation. *Neural Computation*. https://doi.org/10.1162/neco_a_01108.

Sims, A. (2016). A problem of scope for the free energy principle as a theory of cognition. *Philosophical Psychology, 29*(7), 967–980. https://doi.org/10.1080/09515089.2016.1200024.

Smart, B. T. H., & Thébault, K. P. Y. (2015). Dispositions and the principle of least action revisited. *Analysis, 75*(3), 386–395. https://doi.org/10.1093/analys/anv050.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition, 112,* 92–97. https://doi.org/10.1016/j.bandc.2015.11.003.

Stefanics, G., Heinzle, J., Attila Horváth, A., & Enno Stephan, K. (2018). Visual mismatch and predictive coding: A computational single-trial ERP study. *The Journal of Neuroscience*, *38*, 4020–4030. https://doi.org/10.1523/jneurosci.3365-17.2018.

Stöltzner, M. (2009). Can the principle of least action be considered a relativized a priori? In M. Bitbol, P. Kerszberg, & J. Petitot (Eds.), *Constituting objectivity: Transcendental perspectives on modern physics* (pp. 215–227). Dordrecht: Springer.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Dordrecht: Springer.

Varela, F. G., Maturana, H. R., & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *Biosystems, 5*(4), 187–196. https://doi.org/10.1016/0303-2647(74)90031-8.

Von Bertalanffy, L. (1950). The theory of open systems in physics and biology. *Science, 111*(2872), 23–29.

Wittgenstein, L. (1953). *Philosophical investigations*. Oxford: Basil Blackwell.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends Cogn Sci., 10*(7), 301–308.

Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese, 193*(12), 3951–3985. https://doi.org/10.1007/s11229-016-1180-3.

Zheng, D., Luo, V., Wu, J., & Tenenbaum, J. (2018). *Unsupervised learning of latent physical properties using perception-prediction networks*. Retrieved from arXiv arXiv:1807.09244.

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