Biology & Philosophy (2021) 36:30
https://doi.org/10.1007/s10539-021-09807-0

The math is not the territory: navigating the free energy principle

Mel Andrews[1]

Received: 8 October 2020 / Accepted: 28 May 2021

Abstract

Much has been written about the free energy principle (FEP), and much misunderstood. The principle has traditionally been put forth as a theory of brain function or biological self-organisation. Critiques of the framework have focused on its lack of empirical support and a failure to generate concrete, falsifiable predictions. I take both positive and negative evaluations of the FEP thus far to have been largely in error, and appeal to a robust literature on scientific modelling to rectify the situation. A prominent account of scientific modelling distinguishes between model structure and model construal. I propose that the FEP be reserved to designate a model structure, to which philosophers and scientists add various construals, leading to a plethora of models based on the formal structure of the FEP. An entailment of this position is that demands placed on the FEP that it be falsifiable or that it conform to some degree of biological realism rest on a category error. To this end, I deliver first an account of the phenomenon of model transfer and the breakdown between model structure and model construal. In the second section, I offer an overview of the formal elements of the framework, tracing their history of model transfer and illustrating how the formalism comes apart from any interpretation thereof. Next, I evaluate existing comprehensive critical assessments of the FEP, and hypothesise as to potential sources of existing confusions in the literature. In the final section, I distinguish between what I hold to be the FEP—taken to be a modelling language or modelling framework—and what I term "FEP models."

Keywords The free energy principle · Scientific modelling · Mathematical modelling · Epistemicvirtues · Demarcation problem

Mel Andrews
mel.andrews@tufts.edu

[1] The University of Cincinnati, Cincinnati, USA

Published online: 02 June 2021

Introduction

The questions most frequently—and most fervently—asked about the FEP are: Is it true? What is it true of? How do we know (empirically) that it is true? These questions, I argue, rest on a category mistake. They presume that the FEP is the sort of thing that makes assertions about how things are, cuts at natural joints, and can be empirically verified or falsified. I urge that before we can make serious headway on understanding the FEP and putting it to work in scientific practice, we must answer an entirely different set of questions: What sort of scientific object is the FEP? To what discipline(s) does the FEP belong? What role is it intended to play in relation to empirical research? Does the FEP even properly belong to the domain of science? The extant literature has largely begged, dodged, dismissed, and skirted around these questions, without ever addressing them head-on. To the extent that existing works have attempted to address these questions, all have proceeded under the—mistaken, or so I will argue—assumption that the FEP is, itself, fundamentally truth-apt. These questions must, I urge, be answered satisfactorily before we can make any headway on the theoretical consequences of the FEP. Empirical work with the FEP has proceeded in the absence of such a clarificatory project, but it has not been unencumbered by it. I take preliminary steps towards answering these questions in this paper, first by examining the historical path traversed by key formal elements of the framework and the implications they hold for its utility, and second, by offering a route to interpreting the FEP, and models built therefrom, in light of an abundant philosophical literature on scientific modelling.

Existing literature on the FEP invokes “the free energy principle” to refer indiscriminately to both the raw formal structure of the framework and to various models constructed therefrom. My novel proposal is that we reserve the term “free energy principle” to designate the model structure, which can be differentiated from distinct models composed from the combination of that structure with a scientist’s or theorist’s construal thereof. To this end, I first deliver an account of what is known as model transfer, which illustrates a phenomenon undergone by the FEP and helps us to see how models can be broken down into a structure and a construal. I also broach the topic of conceptual reification in modelling. Next, I trace out the history of the core mathematical elements of the FEP, illustrating the formal skeleton of the framework, sans interpretation. In section three, I tackle the claims made in existing critical assessments of the FEP, elucidating where these have gone wrong in their interpretation of the framework. From there, I look to the literature on scientific modelling once again, drawing conclusions about how to wield and interpret the various scientific models built from this formal foundation. My hope is that this text can serve as a fruitful starting place for philosophers and scientists looking to utilise the FEP formalism.

Model transfer, structure & construal

This section introduces the notions of model transfer, model structure, model construal, and model reification, which will enable us to better understand the FEP, along with its uses and misuses. According to several popular accounts of scientific modelling, a scientific model is composed of a structure and an interpretation. This breakdown is most applicable to abstract, formal models. It is perhaps easiest to see this distinction play out in cases where a model structure, originally designed for one modelling purpose, is exapted away from its original interpretation and lent a new one.

The lotka-volterra model

Take the Lotka-Volterra model. The structure, in this case, is a system of nonlinear differential equations. The Lotka-Volterra model originated in physics and chemistry. The equations were originally proposed by Alfred Lotka in 1910 as a model of autocatalysis—self-catalysing chemical sets. They describe the rate of change of chemical concentrations as a chemical system pushed out of equilibrium restores itself, by oscillations, back to steady state. A linearisation of the model produces a system similar to a harmonic oscillator.

A paradoxical trend was noted in fish populations of the Adriatic Sea surrounding World War I. Italian biologist Umberto d'Ancona measured the relative prevalence of fish of various species. He found that, during WWI, when fishing in the Adriatic Sea all but ceased, the predator population experienced a boom. In response, the prey populations diminished considerably. When fishing resumed after the war—an indiscriminate biocide—the prey population soared. Vito Volterra (1926) employed a system of nonlinear differential equations formally equivalent to Lotka's (1910) chemical model to explain how the removal of a constraint (fishing) increased the amplitude of predator population size, while the reimposition of this constraint—the return of fishing after the war—increased the amplitude of the prey populations. In 1956, Lotka independently proposed the same set of equations for the purpose of explaining predator-prey dynamics in his *Elements of Mathematical Biology*.

How does this exemplar help us to understand the free energy principle? In two ways. First, it is common for a relatively coarse-grained formal model initially utilised in one domain (in this case, physical chemistry) to be later imported into an altogether different discipline (in this case, population biology, ecology, and ethology). Second, the case renders intuitive the distinction between the structure and the construal of a model. The structure—a system of differential equations—remains the same for both the (1910) model of catalytic sets and the (1926, 1956) models of predator-prey dynamics. The two models differ in that, in the first instance, scientists interpret the equations to represent chemical concentrations, while in the second instance, scientists interpret the equations to represent population density.

Conceptual reification is a common ailment of scientific modelling. It is particularly likely to occur in cases in which models have somewhat convoluted histories. Reification involves mistaking an aspect of a model—its structure, its construal, or the union of both—for an aspect of empirical data or the natural world; mistaking the math for the territory, so to speak. Reification also occurs when we take an analogical relationship to be a literal one, or when elements of a model's construal in its original domain of application get brought along parasitically into a novel domain in model transfer.

In the next section, we will examine the historical trajectory undertaken by several of the key formal elements of the framework; its legacy of model transfer. This historical exercise allows us to clearly separate the formalism from its interpretations in various domains and lends us insight into some of the reification that has led to confusions surrounding the FEP.

The free energy principle

History of the formalism

To understand the FEP, separated from the various construals attached to it in its various contemporary instantiations for theoretical or simulation purposes, and separated from various construals which may have attached themselves parasitically to the framework during its elaborate history of model transfer, it will be necessary to trace out this historical record. Many of the formal tricks embodied in the FEP originated in physics, some in machine learning, some in physics via machine learning. The FEP, however, is not a law of physics, nor is it a theory, model, or principle belonging to the physical domain. It is not a machine learning technology, though it draws upon some of the same underlying statistical techniques.

A relevant contingent of people concerned with the FEP take it to be, in one way or another, a *physical* description of natural systems. This has an obvious form: taking notions such as energy, entropy, dissipation, equilibrium, heat, or steady state, which play important roles in the FEP, in their senses as originally developed in physics. There is a more subtle form of this tendency, however, in which people begin with the assumption of an analogical relationship to physics, or a mere formal equivalence, but conclude that the formalism of the FEP nonetheless picks out real and measurable properties of natural systems, albeit perhaps more loosely and abstractly than its physical equivalents would. This, I argue, is a conceptual reification; a vestigial interpretation from the formal methods' origin in statistical physics.

The epistemic turn in statistical mechanics

An important precursor to the FEP that seldom comes up in the literature is Jaynes' maximum entropy principle.[1] The classical interpretation of statistical mechanics views the macroscopic variables of some physical system of interest—say, heat, volume, and pressure—as physical constraints on the microscopic behaviour of the system. This is a decidedly physical interpretation of the maths. Jaynes' (1957) critical insight was that we could give this all a subjectivist, epistemological reading, casting these macroscopic variables as knowledge about the system, with the lower-order details to be inferred. The principle of maximum entropy guarantees the maximum (information) entropy of a probability distribution given known variables. Maximising the entropy of the distribution guarantees that we are not building in any more assumptions than we have evidence for. This principle of maximum

[1] For a thorough overview of the FEP/MaxEnt connection, refer to Gottwald and Braun 2020.

entropy took the formalism of statistical mechanics and gave it an information-theoretic interpretation, turning the second law of thermodynamics into a sort of Occam's razor for Bayesian inference. This is because the maximum entropy principle brings us to adopt the probability density with the most widely dispersed probability density function, given the known variables, just as entropy will be maximised with respect to macroscopic variables in statistical mechanics. These are formally identical. Given the frequency with which the literature on the FEP makes reference to Jaynes, one might think it a rather inconsequential piece of the puzzle. In order to understand the FEP, however—and why it is closer to a statistical technique than it is to a falsifiable theory of biological self-organisation—it is important to see that there is a clear precedent for leveraging the maths of statistical mechanics as a method for Bayesian inference. Jaynes' maximum entropy principle (often referred to as MaxEnt) has had tremendous success as a tool for scientific modeling across the sciences. The free energy principle, much like the maximum entropy principle, takes the mathematical machinery of statistical mechanics and lends the formal tools therein a distinctly epistemic, inferentialist bent.

The mean field approximation

Independently, an approach known as mean field theory emerged in statistical mechanics at the beginning of the twentieth century that enabled physicists to study high-dimensional, stochastic systems by means of an idealised model of the system that would average out, rather than summing over, the interactions of elements within the system. Feynman (1972) introduced what are known as variational methods within the path-integral formulation of mean field theory. By exploitation of the Gibbs-Bogoliubov-Feynman inequality, one is able to achieve a highly accurate approximation of the energetics of a target system under a range of conditions. This is accomplished via minimisation of free energy by variations on a simplified candidate Hamiltonian to bring it into accord with the true Hamiltonian.[2] What is important to understand about Feynman's original formulation of the free energy minimisation technique is that it is 1. not to be taken as a literal representation of a target system but rather it is a formal trick for approximating otherwise intractable computational problems that arise in dealing with certain physical systems, and 2. that the free energy involved nonetheless refers to a physical quantity: Helmholtz free energy.

Free eenergy in machine learning

The method of variational free energy minimisation was adapted for statistics and machine learning towards the end of the twentieth century as *ensemble learning* or *learning with noisy weights* (Hinton and Van Camp 1993; Hinton and Zemel 1993; MacKay 2001). Thus free energy minimisation in statistics is a variational method for approximate inference where intractable integrals are involved. A quantity, termed

[2] We may think of the Hamiltonian of a physical system as the net kinetic and potential energies of all of the particles in the system.

variational free energy, is minimised, thus bringing the ensemble density or variational density—the approximate posterior probability density, on a Bayesian interpretation—into approximate conformity with the true target density (Friston et al. 2006; Hinton and Van Camp 1993; MacKay 1995a, b, c; Neal and Hinton 1998). According to Friston, this method of approximating the posterior density or ensemble density is a statistical analogue of the mean field approximation in statistical physics (Friston et al. 2006). We can see that both the free energy term and the construct it is being leveraged to approximate refer to energetic properties of the physical systems under study—Helmholtz free energy, and the system's Hamiltonian—as the method was originally purposed by Feynman (1972). The variational free energy and the variational or posterior probability density involved in the variational free energy minimisation technique as employed by Hinton and Van Camp (1993), however, are purely statistical constructs. Thus, although it may be tempting to lend the "free energy" under the FEP a physical interpretation, it is not meant to invoke a physical quantity. Variational free energy is not Helmholtz free energy, despite the formal similarity.

Variational bayes

The finer points of the formulation of variational Bayes in use today were worked out by Beal (2003) and Attias (2000). Beal (2003) illustrates how conceiving of approximate Bayesian inference in terms of conditional probabilities can be facilitated via graphical models, such as Markov random networks, highlighting the import of the set of nodes that form the Markov blanket of the set of interest. An exact deployment of Bayes' theorem almost always leads to intractable integrals—the sort of calculus it would take an adept mathematician years to solve. By contrast, approximate variational Bayesian methods generate candidate probability distributions and assess the Kullback-Leibler (K-L) divergence between candidate and target distributions.

Innovations in friston's free energy minimisation

Karl Friston took the method of variational free energy minimisation and gave it a dynamical-systems interpretation, specifying the free energy minimisation dynamic in terms of the Fokker-Planck equation and, in particular, the solenoidal and irrotational flows that fall out of the Helmholtz decomposition thereof, of which the irrotational flow can be conceptualised as a gradient-ascent on an attracting set (Friston 2009, 2010, 2012, 2019; Friston and Stephan 2007; Friston et al. 2008). This allows us to think of free energy minimisation simultaneously as a method of approximate Bayesian inference and as a flow.[3]

[3] Friston notes that it is interesting that the formulation of free energy minimisation using gradient flows (otherwise known as gradient descent) was an important practical development for the data analysis tools commonly applied in neuroscience—for example, in dynamic causal modelling. In brief, this freed one from the analytic derivations of vanilla variational Bayes and the use of conjugate priors; enabling a generic variational scheme for modelling empirical data known as variational Laplace.

Jaynes' maximum entropy principle took the formalism of statistical mechanics and leveraged it to accommodate the process of inference given limited and noisy data. Friston's FEP goes a step further, borrowing mathematical tools from the physical sciences to enable the formal representation of processes of inference about the (stipulated) inferential dynamics of systems in nature.

Fundamentals of the FEP

The Fokker-Planck, or Kolmogorov Forward equation describes the time evolution of a probability density function. The Fokker-Planck equation originated in statistical mechanics, in which it described the evolution of the probability density function of the velocity of some particle, or its position, in which case it was known as the Smoluchowski equation. In the context of the FEP, the Fokker-Planck equation describes the evolution of the probability density function of the state of a system. As such it can be thought of as a trajectory through one abstract state space which is a probabilistic representation of some lower-order abstract state space representing what state a given system is in over some definite time window. Any vector field that satisfies the appropriate conditions for smoothness and decay can be broken down into solenoidal (curl) and irrotational (divergence) components. This is known as the Helmholtz decomposition; the fact that we can perform the Helmholtz decomposition is then known as the fundamental theorem of vector calculus.

The static solution to the Fokker-Planck equation is a probability density termed the Nonequilibrium Steady State density, or NESS density (Friston 2019; Friston and Ao 2012). The notion of nonequilibrium steady state is native to statistical mechanics, wherein it describes a particular energetic dynamic between a system and its surrounding heatbath. NESS is best understood as the breaking of detailed balance. Detailed balance is a condition in which the temporal evolution of any variable is the same forwards as it is backwards (the system's dynamics are fully time-reversible). Detailed balance holds only at thermodynamic equilibrium. In nonequilibrium steady state, balance holds in that none of the variables that define the system will undergo change on average over time, but there is entropy production, and there are flows in and out of the system. Jiang et al. (2004) and Zhang et al. (2012) have demonstrated that nonequilibrium steady state can be represented as a stationary, irreversible Markov process. This development paved the way towards a purely statistical rendering of the notion of NESS.

The literature on the FEP also rests centrally on the notion of a Markov blanket (Kirchhoff et al. 2018), an adaptation of Pearl's (1988) Markov boundary. A Markov blanket essentially partitions the world into a thing which can be conceived of as, in its very existence and dynamics, performing a kind of inference, and a thing it is inferring—on a yet more basic level, the Markov blanket allows us to partition the world into a system of interest, and all that lies outside of that system of interest. Systems are represented under the FEP as being subject to random fluctuations, which are responsible for the stochasticity of the systems involved. These fluctuations would result in the dissolution of the systems of interest, were it not for some balancing flow. In the absence of a counteracting flow, the system, as defined by its

Markov blanket, would cease to exist as such. If the set of states considered to be the system (internal states and their Markov blanket) are to resist this dissipative tendency, they must counteract it. This counteracting flow can be conceptualised in a number of ways. For one, we can think of the perturbations as causing the NESS density to disperse, and the irrotational flow under the Fokker-Planck equation as countering these fluctuations. We can also think of it as ascending the gradients induced by the logarithm of the NESS density. The system is hillclimbing on a landscape of probability. It seeks to ascend peaks of maximum likelihood and escape from improbable valleys. In fact, the FEP is a form of dynamic expectation maximisation, which is itself a maximum likelihood function (Friston et al. 2008). The flow of the system must also, moment by moment, minimise surprisal or self-information by gradient descent. Variational free energy constrains this activity by placing an upper bound on surprisal.

Under the FEP, a system of interest can be represented as being subject to random perturbations, which would induce dissipation were it not for some flow countering this dissipation. The Fokker-Planck equation encapsulates these random perturbations as *w*—the Wiener process, or Brownian motion. The curl-free (irrotational) dimension of the flow described by the Fokker-Planck under the Helmholtz decomposition will be seen to counter this flux, maintaining the integrity of the NESS density, which places high probability mass over the system's pullback, or random global attractor (Friston 2019). All this means is that, statistically speaking, the system prefers this region of its phase space—the way a cat likes a laptop computer or a ball likes to roll downhill. The NESS density can also be cast as a generative model, as the highest probability region of the system's phase space will be a joint distribution over all of the system's variables. By generative model, we mean here a joint probability distribution over external and blanket states. For this reason, we can conceptualise the behaviours of the systems treated under the FEP as statistical models of the causes impinging upon them from their environments. This follows from the complete class theorem which, in its Bayesian, statistical generalisation, states that any decision procedure operating according to a loss or risk function in a finite sample space is, under certain assumptions, Bayes optimal with respect to some prior. By extension, then, any dynamical system that minimises some loss or risk function according to some decision procedure is taken to be Bayes optimal under some generative model and priors.

This brings us back to the inferential interpretation of the dynamic described by the FEP. When we apply Bayes theorem to a problem of inference or belief updating, we want to maximise marginal likelihood. Marginal likelihood is the likelihood of some observation given our model; it is also termed Bayesian model evidence, or simply evidence. Surprise and evidence are inverse functions. When we minimise surprisal, we are maximising model evidence. Thus, systems under the FEP are said to be 'self-evidencing' (Hohwy 2016). Over time and on average, the minimisation of surprisal minimises information entropy. This effectively prevents a system's states from dispersing in a statistical sense—it keeps the values of certain key variables within certain existential (that is, definitive of the system) bounds. In minimising (an upper bound on) surprisal, we minimise (a lower bound on) model or marginal evidence, or simply evidence. This makes free energy minimisation equivalent to evidence lower bound optimisation (ELBO), an objective function that anyone with a background in machine learning or Bayesian statistics will find themselves familiar with.

Here we have traced the history of key formal elements of the FEP in Jaynes, Feynman, Hinton, Pearl, Beal, and Friston. Having a handle on this history is necessary in order to grasp the subtle turn away from statistical approximations of physical properties of physical systems to a pure, substrate-neutral method of statistical inference. When we speak of annealing a model in statistical mechanics, ratcheting the temperature of the system up and down in the hopes of bumping it out of local minima, this does not refer to an act of literally injecting energy into a physical system to increase the speed of particle motion. It is a statistical analogue of a physical process. Likewise, the energy and entropy of the FEP are formal analogues of concepts defined in thermodynamics and statistical mechanics with a long history of use in information theory, statistics, and machine learning, in which they have lost their correspondence to any measurable properties of physical systems.

Critical appraisals

Mine is not the first paper to attempt to get to the bottom of the FEP. There have been, to date, a number of attempts at comprehensive critical assessment of the FEP, including Colombo (2017) Colombo and Wright (2017, 2018), van Es (2020), Gershman and Daw (2012) Gershman (2019) Klein (2018), and Sims (2016). The nominal worries of these critical accounts include that the FEP lacks biophysical or cognitive realism, that it somehow contravenes experimental observation, that it is incapable of providing an all-encompassing account of brain function, or that it fails to make novel predictions (Colombo 2017; Colombo and Wright 2017, 2018; Gershman and Daw 2012; Gershman 2019; Klein 2018). The real worry, though, the worry that is only made explicit a few beers in to the spillover of the conference proceedings into some smoke-filled local tavern, the worry that is only put to words in the manuscripts that get rejected before they ever make it to print, is that the FEP is somehow empty; that it lacks all conceptual content. My claim is that the FEP is, indeed, empty in just this way, that that is not—or ought not to be—a secret, and that its contentlessness does not count as a mark against the framework.

Indeed, existing works in this genre all stack the solutions they arrive at against this conclusion. They all appear to beg the question. These critical accounts, much like the preponderance of positive accounts, fail to differentiate between the FEP as *formal model structure* and the various models built from or atop this structure, which themselves are often casually referred to as "the FEP" in the literature.

Colombo & Wright

Colombo and Wright (2018) do entertain the idea that the FEP is merely a formal modelling tool—indeed, they even diagnose the problem of reification in modelling, “the risk of conflating scientific models and their targets” (p. 12)—but they never take this hypothesis seriously. Colombo and Wright (2018) raise the matter as a serious question to be addressed: “should we understand FEP as a modeler’s tool to characterize and predict adaptive behavior, or should it be understood as an objective feature of target systems?” (Colombo and Wright 2018, pp. 15-16). As they proceed, however, they merely “assume that the probabilities involved in FEP aren’t simply modelers’ tools” (Colombo and Wright 2018, p. 17). It is unclear on what grounds they justify this assumption, beyond the further (unjustified) assumption that the aim of the modelling exercise must be realism and that “Friston (2013) seems to interpret the probabilities involved in FEP as objective features of real-world systems” (Colombo and Wright 2018). Ultimately, the hypothesis that the FEP is empirically contentless is swept aside: “what’s intended cannot be that FEP is unfalsifiable because it fails to be truth-apt” (Colombo and Wright 2018, pp. 22-23).

Gershman, Klein, & Williams

Gershman’s (2019) supposed critique of the framework does not actually countenance the FEP, but diverts its attention to process models, writing “we will be concerned with [the FEP’s] credentials as a theory, and therefore we will pay particular attention to specific implementations (process models)” (Gershman 2019, p. 2). Sims (2016) likewise conflates the FEP with associated corrolaries and process theories. For example, he writes: “In its form as a theory of cognition, the application of this theory to explain mental phenomena draws heavily upon the notion of expected precision” (Sims 2016, 970). Expected precision is a notion proper to predictive coding and various models that fall under the heading of predictive processing. In a similar vein, Sims writes that the “free energy principle makes certain non-trivial predictions about brain structure and function,” listing among these predictions a greater preponderance of top-down (feedback) connections than bottom-up, feedforward connections, and the organisation of the cortical hierarchy, likewise the purview of hierarchical predictive coding or predictive processing models. Klein (2018) argues that the FEP is either susceptible to the dark room problem or it is something like an idealised model. On this deflationary depiction of the FEP, it is taken as “a starting point from which one might develop explanations,” and its success (or failure) as a scientific tool ultimately rests on “the empirical adequacy of detailed models which spring from it” (Klein 2018, p. 2554). This, I think, is precisely the right mode of understanding FEP-based models. Thus, models built from the formal architecture of the FEP offer “a deliberate simplification, which buys scientific fruitfulness at the cost of literal truth” (Klein 2018, p. 2554). Notably, this is quite close to a Wimsatt (1987) view of the epistemic virtues of modelling. Williams (2020) delivers a description of the FEP that is diametrically opposed to Sims’ (2016) depiction: “the FEP does not advance a causal hypothesis. Specifically, it provides no information about how self-organization is causally generated and sustained in the systems that it applies to" (Williams 2020, p. 20).

It appears that we have the theoretical equivalent of binocular rivalry when it comes to depictions of the FEP's empirical commitments and epistemic status. How, then, do we resolve this conflicting vision? First, we ought to note that both Sims (2016) and Williams (2020), like most of their predecessors, take "the free energy principle" to refer to both the raw formalism and to various models predicated on that formalism which Friston and colleagues have described over the years. Additionally, Klein (2018) seemingly takes "the free energy principle" to refer to active inference and perceptual active inference, hierarchical predictive coding, and various predictive processing models, describing all as one and the same theory of cognition. The deflationary conclusion he reaches, however, maps onto the role played by various models constructed from the formal framework of the FEP: FEP-based models act as heavily-idealised, generic or targetless mathematical models. One core function such models serve in relation to scientific practice is as a stepping stone on the road to more fine-grained and empirically rich models. Williams (2020), on the other hand, adequately differentiates the FEP from various associated process models and adjacent theories within the Bayesian brain canon. The conclusions he reaches in regards to the FEP's explanatory scope and epistemological status fail to cohere to what the literature has, to date, *descriptively* used "the FEP" to denote. However, I contend that Williams (2020) assessments of the FEP are precisely on the mark as an appraisal of what the FEP ought, *normatively*, to refer to: namely, the formal structure, absent any interpretation relating it to a target.

Van Es

Van Es (2020) is the first to draw clear connections between FEP models and the literature on scientific modelling. Van Es's (2020) argument is fairly straightforward: 1. FEP models describe organism-environment interactions in terms of modelling. 2. It is unclear to what extent it is intended that an FEP model models organisms as though they were, themselves, engaged in a practice of modelling their environments, and the extent to which proponents of FEP models intend the models posited thereunder to be literal, in the world, and independent of scientists' modelling practice. 3. The existing literature utilising FEP models to address cognition seems to rest, fundamentally, on a conflation between the two. 4. Kirchhoff and Robertson (2018) argue that the sense in which organisms' dynamics mirror their environments, under FEP models, is not representational in nature, but merely covariational. 5. There is consensus in the literature on scientific modelling that scientific models are representational in nature. 6. The sole nonrepresentational account of scientific modelling on offer appeals to social practice of science to ground the representational features of models. 7. Models as leveraged by organisms, under FEP models, cannot recourse to the social practice of science to ground the representational features of their models. 8. Therefore, models posited under FEP models fail to count as models. 9. We must, then, according to van Es, adopt an instrumentalist stance on models under the FEP.

Van Es's argument, however, ignores the lengthy and abundant history of arguing over whether scientific models must be representational in nature, and whether their sole epistemic virtue must be their representational status. Downes' (2020) survey and introduction to the modeling literature supplies an overview of the debate between representationalists—those who hold that all models necessarily represent—and nonrepresentationalists—those who hold that not all models need necessarily represent. Downes (2011) argues for the position that not all scientific models represent. He notes here that the emerging consensus among philosophers concentrated on scientific modeling is that models serve a plurality of epistemic purposes for research. The characterisation of scientific modelling in van Es (2020) thus comes across as quite far-afield from the state of the literature. Van Es saddles the literature on scientific modelling with the task of demonstrating that models are, by necessity, representational. In fact, the scientific modelling literature shows quite the opposite: models need not represent, and representation is not the chief epistemic virtue of scientific modelling. Van Es points to Oliveira's (2018) pragmatist approach to modelling as the singular example in the literature of a nonrepresentationalist approach to understanding scientific models. In fact, the modelling literature has undergone something of a pragmatist turn in the 21st century, and nonrepresentational accounts abound.

These are not the most fatal flaws in van Es' argument, however. By far the more questionable premise is that the generative models embodied or entailed by organisms under various FEP models are models in precisely the sense in which scientific models are models. The assertion that all scientific models are generative models in the formal, statistical sense would be rejected by scientific modellers and philosophers of scientific modelling alike. It is clear that scientific models do not denote statistical models of joint probability distributions. Why, then, should we assume that the generative modelling stipulated under the FEP is precisely the same sort of modelling that scientists engage in, and subject to the same constraints?

While instrumentalism with respect to the mathematical constructs leveraged in FEP models is a meritorious position, I do not think it is novel—in fact, I think it is presumed throughout the literature, and the argumentative route van Es traverses to arrive at this position is a nonstarter. Van Es contends that the existing FEP literature erroneously conflates two senses of modelling. In fact, van Es' own paper demonstrates a conflation between various senses of the term. There are, in the first place, models as utilised by scientists to gain leverage over the natural world. There is a deflationary, statistical notion of a generative model as a statistical model of a joint probability distribution. Van Es (2020) seems to run the two together, and further conflates both with an unarticulated strawman notion of models, under which a brain may be said to "model" its environment in some cognitivist, representationalist sense. I believe the culprit here is a lack of fluency with both the philosophical literature on scientific modelling and the statistical techniques that undergird FEP models.

A tentative diagnosis

Whence such confusion? For one, we have seen the convoluted history of model transfer the FEP has undergone, with formal elements drawn from a number of disciplines and passing through multiple interpretations before achieving their current form and use. For another, it is not often pragmatically necessary in the practice or research or theorising to specify in painstaking detail what formal models we are drawing from and in what mixture and quantity. For another, at the risk of psychologising, when one's primary mode of relating to the world is not linguistic, but mathematical, the distinction between "maths" and "territory" makes little sense. In fact, I believe that a necessary precondition to being a good physicist is the loss of this distinction between literal description of the world and mere formal trick. This creates something of a barrier to interdisciplinary communication. Physicists or those with physics backgrounds are often known to say things which, to mathematical modellers in biology, cognitive science, pure maths, or machine learning often seem to mix metaphors. I view as symptomatic of this tendency Friston's overly literal descriptions of the formalism.

There are many places throughout the literature on the FEP in which the language used to describe the formalism can easily give rise to the misconception that the framework is a literal—perhaps physical—description of some measurable feature of natural systems, or cuts at natural joints. Ramstead et al. (2018), for example, write that "systems are alive if, and only if, there[sic] active inference entails a generative model" (p.33). Under the perspective of the FEP—that is, once we have elected to model biological systems using the formal tools the FEP provides us—any system we choose to model in this way will behave as the model dictates it must. Under the FEP, in order to be a system, certain mathematical assumptions must hold. In particular, we assume a weakly-mixing random dynamical system, a Markov blanket, and either ergodicity or (organisation to) nonequilibrium steady state (NESS). If we take the systems attracting set to be a NESS density, then its existence will entail a generative model. This is a result of the statistical generalisation of the complete class theorem. Thus in selecting to model a system under the FEP, we have presumed its dynamics to entail a generative model. This says nothing, however, about any empirically-ascertainable properties of living systems.

Both the literature forwarding the FEP and the literature critiquing it suffer from issues of translation: would-be interpreters of the framework face the burden of translation between linguistic and mathematical descriptions, between discrepancies in disciplinary standards and terms of art, and the long history of translation between various applications and interpretations which components of the mathematical framework itself have undergone. I urge that by unpacking these discrepancies in disciplinary conventions, scrutinising the history of the formalism, and divorcing it not only from meanings ascribed to it in past disciplinary settings and applications, but from meanings applied to it in what I term "FEP models," operationalisations of the framework, we can come to a considered understanding of what the FEP is in its own right.

Models & The FEP

The FEP as model structure

I propose here that we reserve “the free energy principle” to denote only the maths of which the FEP is composed. Models utilising this formalism to study natural systems bear also an interpretation, lending a means of interpreting the maths as *about* systems in nature. The papers that exist on the FEP, however, nearly all seem to leverage the FEP to address some issue, relating it to a target system. They are what we might call “FEP models.”

According to Weisberg’s (2007; 2013) account of models, a model’s structure—whether mathematical, computational, or physical—does not inherently relate to a target system in an epistemically fruitful way, e.g., by representing features of that target system. It is only with the addition of a scientist’s interpretation or construal that we derive a mapping between a model structure and the world, and a model is born. Once we have that model in hand, however, it can be tempting to say that we have knowledge of the natural world. The existence of models, their features, and the output of modelling work, however, do not constitute knowledge of nature over and above empirically-observed facts that we may have plugged into our models at the outset (if they are the type of models that incorporate measurements). To validate a model or its results and derive from it knowledge of natural laws, systems, or processes, we must match its predictions to experimental evidence.

We make a category mistake when we claim that a raw mathematical structure lends us predictions or places constraints on what can be observed in nature, and are guilty of reification: “there is no such thing as a solely mathematical account of a target system” (Nguyen and Frigg 2017, p.1). Likewise when we take the existence or qualities of a model to constitute knowledge of the natural world we make a category error and reify the model.

Models have been proposed utilising the formal structure of the FEP that take as their targets cortical structure, neuronal organisation (Friston, Fagerholm, Zarghami, Parr, Hipólito, Magrou, & Razi, 2021), the brain as a whole (Friston 2010), organisms acting in an environment (Bruineberg and Rietveld 2014; Bruineberg et al 2016, Buckley, Kim, McGregor, & Seth), morphogenesis of multicellular organisms (Kuchling et al. 2019; Friston et al. 2015), and even social structures and behaviours (Ramstead et al. 2016). These have been representationalist, FEP models (Kiefer and Hohwy 2018, 2019), cognitivist FEP models (Kiefer and Hohwy 2018, 2019), nonrepresentational, Gibsonian FEP models (Bruineberg and Rietveld 2014; Bruineberg et al 2016;), enactivist FEP models (Kirchhoff and Froese 2017Ramstead et al. 2019), as well as both dualist (Hobson and Friston 2014) and materialist (Friston et al. 2020; Kiefer 2020) FEP models. Each of these is referred to in the literature as “the free energy principle,” and the ontologies, epistemologies, and predictions borne of each deemed consequences of “the free energy principle.”

That one and the same “theory” can lend itself explanations of neuronal, cellular, and social organisation seems puzzling; that it can lend support for both neurocentrism and extended cognition even more so. I propose that we can resolve the source of this error signal by denying, in the first place, that the FEP is a theory—or even a model. Instead, we ought to use "the FEP" to denote the model structure: the raw formal framework, sans interpretation. In combination with the numerous construals that exist in the literature, this structure becomes "FEP models." The FEP itself, then, lacks all empirical or conceptual content. It is an empty formalism. The many models built thereon, however, can be seen to differ with respect to their content (thus different targets, conflicting interpretations of the same target). If we continue to take the FEP to refer to all of the above, then the conclusion we must reach (and that critics of the framework have reached) is that the FEP must either be vacuous or internally inconsistent. I say we bite the bullet on vacuousness, but restrict the term to be used only in reference to the formalism.

To illustrate briefly what I mean here: Ramstead et al. (2018) write that: "The FEP is a mathematical formulation that explains, from first principles, the characteristics of biological systems that are able to resist decay and persist over time," (p. 2) and that it "asserts that all biological systems maintain their integrity by actively reducing the disorder or dispersion" (p. 3). I wish to urge that claims to the effect that the FEP "explains" or "asserts" anything are misguided. Rather it is only FEP based models which can assert or explain and, indeed, relate at all to the world. Compare this to Kiefer (2020) who more rather more carefully describes his work as a "conjunction of the free energy principle...and the identity thesis" (p.1). Here I am not making a prescriptive case that we ought to always and only ever refer to "the FEP" as the maths in the absence of any interpretation and to the models constructed therefrom as "FEP models," though admittedly this might go a long ways in clearing up some of the misapprehensions in the literature. Rather it is my hope that this distinction will equip the casual reader with the conceptual tools necessary to parse the FEP literature, as notoriously dense and befuddling as it is. The "FEP" as it has been addressed in the literature thus far is something of an impossible figure: A pure formalism, empty of all content. A precise theory of neuronal signalling and transient ensembles. A tautology. A transcendental argument. It is both unfalsifiable and yet makes precise predictions about the physical or physiological systems and dynamics capable of instantiating it. Both materialist and dualist. Both representational, cognitivist, and neurocentrist and, at the same time, ecological, enactive, and extended. In this section, I have argued that the only way around interpreting the FEP as overcommitted to conflicting claims is to distinguish the formal framework from various models composed therefrom. We ought, then, to think of the FEP as a mathematical structure alone, free of conceptual content, predictions, or representations of worldly systems. Models add to a structure an interpretation, or construal, which lends them a mapping function to worldly systems. Models composed from the formal structure of the FEP may or may not make predictions, or place constraints on the varieties of systems capable of realising the dynamics they specify. This is likely to differ from model to model. Attempting to saddle the formal architecture of the FEP itself with falsifiable predictions, however, is simply a category error.

Conclusion

A recent and abundant literature concerns itself with the FEP, though its claims and status are contested. The FEP as it is addressed in this literature—both for and against—is a mathematical structure applied to the modelling of various phenomena across the social, cognitive, neuro-, and life sciences. Attempts to secure the precise nature of the FEP, its utility for scientific practice, just how it represents systems in its respective domains of application, how precisely it is bolstered or refuted by existing empirical evidence, what constraints it places on process theories and lower-order models, have thus far been foiled. I suspect that this is the case because the questions we have been asking of the framework make implicit category errors, attempting to saddle it with attributes that fail to apply to the framework in virtue of the sort of thing that it is. What are the FEP's theoretical commitments? What empirical support does it receive? There are as many answers to these questions as there are papers on the FEP, and continuing to ask them of the modelling framework as a whole, without regard to the distinct forms it assumes, will remain a fruitless undertaking.

In this paper, I have argued the case that the free energy principle be considered not a theory, a law, a hypothesis, a paradigm, or a model, but a formal modelling structure. One immediate consequence of this conclusion is that questions of the epistemic status of FEP models, their empirical content, the predictions they do or do not make, and their precise relation to various corollary models and process theories will have to be assessed piecemeal, for each FEP model in its own right. This seems to pass the explanatory buck. In this respect, the state of the FEP mirrors the state of the modelling literature at large: there is very little that can be said evaluatively of models as an undifferentiated whole. There are, though, broad-brushstrokes appraisals to be made of the sorts of models that emerge from the FEP. Following this, there are more exacting claims to be made about specific instantiations of the framework. These, however, will have to be the subject of a later paper.

Another, more positive consequence, however, is that, having separated the formal essence of the FEP from various interpretations thereof, we are now free to build theoretical or empirical models with it without worrying about its theoretical commitments or empirical support, for it has none. The only barrier to utilising the FEP is understanding the maths and understanding how to relate it to an open question or target system of interest. We did not enquire as to the theoretical commitments or empirical support of the evolutionary algorithm or Conway's game of life; we simply played with them. Our approach to the FEP ought to be the same.

That no silver bullet will vanquish—or vindicate—the spectre of the FEP once and for all may come as a disappointment to some. Have all our attempts to nail it down been in vain? I, for one, think not. It is my sincere hope that one positive result of this exercise will have been to disabuse a few scholars of outmoded conceptions of the scientific method. Scientists are not everywhere all the time dealing in literally true direct representations of natural systems. Arguably, much of what we are doing as scientists—especially now, especially in the younger disciplines—is far more heuristic and bottom-up than the theory-driven, hypothetico-deductive, falsificationist frames of yore would have us believe.

It is perhaps natural—at least not uncommon—to think that without hard and universal desiderata for what differentiates scientific pursuits from other intellectual projects, for what separates good and bad science, that we will slip into everything-goes relativism or pluralism. An effort is only scientific, and hence, worthwhile, if it makes concrete, falsifiable predictions. If we adopt this frame, most of today's scientific methods are not scientific at all. In fact, many of our historical bastions of the scientific method have failed to live up to these rigorous standards. It should not be considered a failing, however, that Galileo, Newton, and Darwin's work failed to be paradigmatic, failed to be theory-driven, failed to uncover mechanisms, or failed to conform to a hypothetico-deductive model. Model-based philosophy of science embraces the messiness and pluralism of scientific practice. It has, at times, done so at the cost of being overly-permissive. Pluralism as a thesis, however, does not absolve us of the responsibility of distinguishing good and bad science, working methodologies from those that have become enmired. That we cannot dismiss the FEP outright for failing to put forth a testable hypothesis—or accept it because it purports to explain a great many things—should not cause us to throw our hands in the air. It should be, rather, a call to arms, an impetus to get ever more exacting.

Acknowledgements This paper has benefited tremendously from thoughtful feedback and lengthy discussions with many people, most notably Liam Bright, Tony Chemero, Andrew Corcoran, Carolyn Dicey Jennings and the Philosophy Lab at UC Merced, Krzysztof Dołega, Thomas van Es, Karl Friston, Cecilia Heyes, Bryce Huebner, Alex Kiefer, Michael Kirchhoff, Maxwell Ramstead, and Dalton Sakthivadivel.

Compliance with ethical standards

Declarations The author declares no conflicts of interest.

References

Attias H (2000) A variational Bayesian framework for graphical models In Advances in neural information processing systems (pp 209-215)

Beal MJ (2003) Variational algorithms for approximate Bayesian inference. Doctoral dissertation, University College London

Bruineberg J, Kiverstein J, Rietveld E (2016) The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. Synthese 195(6): 2417–2444

Bruineberg J, Rietveld E (2014) Self-organization, free energy minimization, and optimal grip on a field of affordances. Frontiers Human Neurosci 8:599

Colombo M (2017) Social motivation in computational neuroscience Or if brains are prediction machines, then the Humean theory of motivation is false. In: Kiverstein J (ed) Routledge handbook of philosophy of the social mind. Routledge, London, pp 320–340

Colombo M, Wright C (2017) Explanatory pluralism: An unrewarding prediction error for free energy theorists. Brain Cogn 112:3–12

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

Downes SM (2011) Scientific models. Philosophy. Compass 6(11):757–764

Downes SM (2020) Models and Modeling in the Sciences: A Philosophical Introduction. Routledge van Es T (2020) Living models or life modelled? On the use of models in the free energy principle Adaptive Behavior, 1059712320918678

Feynman RP (1972) Statistical mechanics: a set of lectures Reading. W A Benjamin, Mass

Friston K (2009) The free-energy principle: A rough guide to the brain? Trends Cognitive Sci 13(7):293–301

Friston K (2010) The free-energy principle: A unified brain theory? Nat Rev: Neurosci 11(2):127–138

Friston K (2012) A free energy principle for biological systems. Entropy 14(11):2100–2121

Friston K (2013) Life as we know it. J Royal Soc Interface 10(86):20130475

Friston K (2019) A free energy principle for a particular physics arXiv preprint arXiv:190610184

Friston K, Ao P (2012) Free energy, value, and attractors Computational and mathematical methods in medicine

Friston KJ, Fagerholm ED, Zarghami TS, Parr T, Hipólito I, Magrou L, Razi A (2020) Parcels and particles: Markov blankets in the brain arXiv preprint, arXiv:200709704

Friston K, Frith CD (2015) Active inference, communication and hermeneutics. Cortex 68:129–143

Friston KJ, Wiese W, Hobson JA (2020) Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. Entropy 22(5):516

Friston K, Levin M, Sengupta B, Pezzulo G (2015) Knowing one's place: a free-energy approach to pattern regulation. J Royal Soc Interface 12(105):20141383

Friston K, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W (2006) Variational free energy and the Laplace approximation. Neuroimage 34(1):220–234

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

Friston K, Trujillo-Barreto N, Daunizeau J (2008) DEM: a variational treatment of dynamic systems. Neuroimage. 41(3):849–885

Gershman SJ (2019) What does the free energy principle tell us about the brain? Neurons, Behav, Data Anal, Theory 2(3):1–10

Gershman S J, Daw N D (2012) Perception, action and utility: The tangled skein Principles of brain dynamics: Global state interactions, 293-312

Gottwald S, Braun DA (2020) The two kinds of free energy and the Bayesian revolution. PLoS Comput Biol 16(12):e1008420

Hinton GE, Van Camp D (1993) Keeping the neural networks simple by minimizing the description length of the weights In: Proceedings of the sixth annual conference on Computational learning theory (pp 5-13)

Hinton G E, Zemel R S (1993) Autoencoders, minimum description length and Helmholtz free energy. Advances in neural information processing systems (pp 3-10)

Hobson JA, Friston KJ (2014) Consciousness, dreams, and inference: The Cartesian theatre revisited. J Consciousness Stud 21(1–2):6–32

Hohwy J (2016) The Self-Evidencing Brain. Noûs. 50:259–285

Jaynes ET (1957) Inform Theory Statist Mech Phys Rev 106(4):620

Jiang D Q, Jiang D, Qian M (2004) Mathematical theory of nonequilibrium steady states: on the frontier of probability and dynamical systems (No 1833) Springer Verlag

Kiefer AB (2020) Psychophysical identity and free energy Journal of the Royal Socety. Interface 17:20200370

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

Kiefer A B, Hohwy J (2019) Representation in the Prediction Error Minimization Framework. In: Sarah K Robins, John Symons Paco Calvo (Eds), The Routledge Companion to Philosophy of Psychology: 2nd Edition London, UK: pp 384-409

Kirchhoff MD, Froese T (2017) Where there is life there is mind: In support of a strong life-mind continuity thesis. Entropy 19(4):169

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

Kirchhoff M, Robertson I (2018) Enactivism and predictive processing: A non-representational view. Philoso Explor 21:264–281

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

Kuchling F, Friston K, Georgiev G, Levin M (2019) Morphogenesis as Bayesian inference: A variational approach to pattern formation and control in complex biological systems Physics of life reviews

Lotka AJ (1910) Contribution to the theory of periodic reactions. J Phys Chem 14(3):271–274

Lotka AJ (1956) Elements of Mathematical Biology New York. Dover Publications, New York

MacKay DJ (1995a) Developments in probabilistic modelling with neural networks–ensemble learning In Neural Networks: Artificial Intelligence and Industrial Applications. Springer, London, pp 191–198

MacKay D J (1995b) Ensemble learning and evidence maximization In Proc Nips (Vol 10, No 154, p 4083)

MacKay DJ (1995c) Free energy minimisation algorithm for decoding and cryptanalysis. Electron Lett 31(6):446–447

MacKay DJ (2001) Local minima, symmetry-breaking, and model pruning in variational free energy minimization Inference Group. Cavendish Laboratory, Cambridge, UK

Neal RM, Hinton GE (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan MI (ed) Learning in Graphical Model

Nguyen J, Frigg R (2017) Mathematics is not the only language in the book of nature. Synthese, 1-22

de Oliveira G S (2018) Representationalism is a dead end. Synthese, 1-27

Palacios ER, Isomura T, Parr T, Friston K (2019) The emergence of synchrony in networks of mutually inferring neurons. Sci Rep 9(1):1–14

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Elsevier

Ramstead MJD, Badcock PB, Friston KJ (2018) Answering Schrödinger's question: A free-energy formulation. Phys Life Rev 24:1–16

Ramstead MJ, Kirchhoff MD, Friston KJ(2019) A tale of two densities: Active inference is enactive inference. Adaptive Behavior, 1059712319862774

Ramstead MJ, Veissiére SP, Kirmayer LJ (2016) Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention. Front Psychol 7:1090

Sims A (2016) A problem of scope for the free energy principle as a theory of cognition. Philos Psychol 29(7):967–980

Volterra V (1926) Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Memoria della Reale Accademia Nazionale dei Lincei 2:31–113

Weisberg M (2013) Simulation and similarity: Using models to understand the world. Oxford University Press

Weisberg M (2007) Who is a Modeler? Br J Philos Sci 58(2):207–233

Williams D (2020) Is the Brain an Organ for Prediction Error Minimization? PhilSciArchive Preprint

Wimsatt WC (1987) False models as a means to truer theories. In: Nitecki M, Hoffmann A (eds) Neutral models in biology. Oxford University Press, Oxford, pp 23–55

Zhang XJ, Qian H, Qian M (2012) Stochastic theory of nonequilibrium steady states and its applications. Part I Phys Rep 510(1–2):1–86