The Math is not the Territory: Navigating the Free Energy Principle

Mel Andrews

October 8th, 2020

Contents

1 Abstract

The free energy principle (FEP) has seen extensive philosophical engagement—both from a general philosophy of science perspective and from the perspective of philosophies of specific sciences: cognitive science, neuroscience, and biology. The literature on the FEP has attempted to draw out specific philosophical commitments and entailments of the framework. But the most fundamental questions, from the perspective of philosophy of science, remain open: To what discipline(s) does the FEP belong? Does it make falsifiable claims? What sort of scientific object is it? Is it to be taken as a representation of contingent states of affairs in nature? Does it constitute knowledge? What role is it intended to play in relation to empirical research? Does the FEP even properly belong to the domain of science? To the extent that it has engaged with them at all, the extant literature has begged, dodged, dismissed, and skirted around these questions, without ever addressing them head-on. These questions must, I urge, be answered satisfactorily before we can make any headway on the philosophical consequences of the FEP. I take preliminary steps towards answering these questions in this paper, first by examining closely key formal elements of the framework and the implications they hold for its utility, and second, by

highlighting potential modes of interpreting the FEP in light of an abundant philosophical literature on scientific modelling.

2 Introduction

With respect to the demarcation problem, I defend the position that the FEP can be taken to belong properly to the domain of science. I survey a number of philosophical accounts of various types of scientific models with similar use and epistemic status to the FEP. I do not insist on interpreting the FEP as a scientific model. It may be more aptly described as something like a modelling framework, or else a principle, akin to Hamilton's principle of least action. No account of such objects exists, however, within the philosophy of science literature, and I think the important takeaways from scientific models can easily be extended to modelling frameworks and principles if, indeed, this is what the FEP is. The FEP is certainly not anything like a theory, a law, a hypothesis, or a research programme, as these have classically been understood in the history and philosophy of science.

The only hint Friston himself has given us to go on with respect to the metatheory or philosophy of science of the FEP is that he labels it a normative theory or normative model, contrasting it with what he terms process models (Hohwy, 2020a, 2020b; Schwartenbeck, FitzGerald, Dolan, & Friston, 2013). Process models associated with the FEP include active inference, predictive coding, and various models falling under the more general label of predictive processing (Friston & Frith, 2015; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017). Three fundamental facts about the FEP follow from its status as normative model: 1. The FEP is not falsifiable and does not lend itself to the direct generation of predictions or hypotheses; 2. The FEP does not cut at joints of nature—which is to say that it does not illuminate the boundaries between natural kinds; 3. The FEP does not deliver mechanisms. The normative model–process model distinction is one drawn from mathematical psychology, and is thus specific to mathematical or computational models of cognition and action. Using this as a starting point, however, we can see how the FEP fits in with some of the more domain-generic literature on scientific modelling.

The literature on scientific models in general has tended to emphasise their practical utility as stand-ins or intermediaries for observation and manipulation of real-world systems (Downes, 2020; Weisberg, 2013; Wimsatt, 1987). This in itself makes the literature on modeling a good candidate for assessing the FEP. However, a few visions of modeling in particular are especially well-suited to the task. Thus I will briefly rehearse what has been written on normative models (Luce, 1995), exploratory models (Frigg & Hartmann, 2020), targetless models (Weisberg, 2013), and general and conceptual models (Barandiaran, 2008). These varieties of model are meant to serve as heuristics (Chemero, 2000, 2009), proofs of principle (Frigg & Hartmann, 2020; Gelfert, 2016), guides to discovery (Barandiaran & Chemero 2009; Chemero, 2000, 2009), and even as scaffolds for a kind of storytelling that should pave the way for fresh avenues of investigation.

In order to see how the literature on scientific models accords with the FEP, however, we will first need to establish a firm sense of what the FEP is—and what it is not. This will require dispelling a number of false assumptions that have been made about the framework. I accomplish this first by tracing out the historical buildup to the FEP, illustrating where the formalism has been derived from, and what it has come to signify. This serves the purpose primarily of showing how much of the mechanics of the FEP had physical meaning in its initial form, but has since come into a strictly formal use. Following this, I will demonstrate that the formalism is empty of any sort of facts or assertions about the state of nature that would allow it to draw taxonomical distinctions, to differentiate classes of natural systems, to explain their behaviour in terms of underlying mechanisms, or to bring forth testable hypotheses. In doing so, I hope to resolve some of the uncertainty surrounding the FEP.

3 The Free Energy Principle

3.1 History of the Formalism

The questions most frequently—and most fervently—asked about the free energy principle 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. 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 free energy principle, 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.

3.1.1 The Epistemic Turn in Statistical Mechanics

An important precursor to the FEP that seldom comes up in the literature is Jaynes' maximum entropy principle. 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 equations. 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 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 widest distribution, 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. To select just one example, MaxEnt has been met with particular appreciation amongst ecologists, in which it proves exceptionally good at picking out and predicting patterns in biodiversity and its distribution.

3.1.2 The Mean Field Approximation

Independently, a method 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 the free energy parameter by variations on a simplified candidate Hamiltonian to bring it into accord with the true Hamiltonian.[1] What is important to understand about Feynman's original formulation of the free energy minimisation technique is that it is 1. a formal trick for approximating otherwise intractable physical systems, and 2. that the free energy involved nonetheless refers to a physical quantity: Helmholtz free energy.

3.1.3 Free Energy 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 & van Camp, 1993; Hinton & 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 variational free energy, is minimised by successive iterations of a model, 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, Mattout, Trujillo-Barreto, Ashburner, & Penny, 2006; Hinton & van Camp, 1993; MacKay 1995a, 1995b, 1995c; Neal & Hinton, 1998). The ensemble density or approximate posterior density is the statistical equivalent of the mean field approximation (Friston et al., 2006). We can see that both the free energy parameter and the construct it is being leveraged to approximate referred 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).

[1] Think of the Hamiltonian of a physical system as the net kinetic and potential energies of all of the particles in the system.

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.

3.1.4 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. Computing technology enabled approximate Bayesian inference via Monte Carlo pairwise sampling methods. Instead of arriving at the target posterior density by manual marginalising, we iteratively hone in on the posterior by sampling it many thousands of times, which is possible using computers. By contrast, variational Bayesian methods toss out candidate probability distributions and acquires the Kullback-Leibler (K-L) divergence between candidate and target distributions. In many contexts, variational Bayes has the advantage over Monte Carlo methods in both accuracy and efficiency.

3.1.5 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 & Stephan, 2007; Friston, Trujillo-Barreto, & Daunizeau, 2008). This allows us to think of free energy minimisation simultaneously as a method of approximate Bayesian inference and as a flow.

3.2 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 free energy principle, 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. A three dimensional 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, Qian, and Qian (2004) and Zhang, Qian, and Qian (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.

Under the free energy principle, a system of interest is 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 $\dot{w}$—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 is the system's pullback, or random global attractor (Friston, 2019). All this means is that, statistically speaking, the system likes this region of its phase space—the way a cat likes a laptop computer or a ball likes to roll down hill. 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. For this reason, we can conceptualise the behaviours of the systems treated under the FEP as (latent) statistical models of the causes impinging upon them from their environments.

The literature on the free energy principle also rests centrally on the notion of a Markov blanket (Kirchhoff, Parr, Palacios, Friston, & Kiverstein, 2018). 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. Recall that systems are represented under the FEP as being subject to random fluctuations, which are responsible for their stochasticity. 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. As we have already discussed, 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 wants 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, Trujillo-Barreto, & Daunizeau, 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.

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 the 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.

We think of a system. We define the system by delineating certain variables and certain ranges of values for these variables. For nearly any system we could choose, there are values of, e.g., temperature and pressure at which the system would no longer be the system we had originally envisioned. Unless the system exists in a vacuum—and nature abhors a vacuum—it will be subjected to stochastic external influences. In order to remain the system that it is, the system will have to act so as to counteract these influences—at least in some very deflationary sense of the notion of action. This action can be conceived of as anticipatory.[2] Thus, representing a system under the FEP allows us to consider it as enacting a statistical model of its environment. The notion of model at play here is that of a generative model, of the kind familiar in machine learning contexts (Ramstead, Kirchhoff, & Friston, 2019). The fact that the NESS density—the high-probability region of the system's phase space—is a joint distribution over all of the system's variables licenses interpreting it as a generative model. The generative model is a formal expression of the good regulator theorem of cybernetics, which states that any subsystem capable of effectively controlling the system at large in which it is embedded must be isomorphic to the system at large (Conant & Ashby, 1970; Friston, Thornton, & Clark, 2012; Ramstead, Kirchhoff, & Friston, 2012). In other words, it must be a model, or a representation, of all of the variables of the system. As I suspect readers may have an inkling for at this point, applying this formalism to some unoccupied region of spacetime, or to a brick, is not going to get us much leverage to speak of on these systems. It is much more meaningfully applied to systems that take up a more active role in safeguarding their own identities: organisms, sub-organismal systems (cells, tissues, organs), and superorganismal systems (societies, social networks, ecosystems).

[2] The notion that the organisation of living systems, down to their very molecular makeup, is fundamentally *prospective*—that is, biological organisation is a manifestation of statistical structure absorbed from past experience towards the ends of future self-maintenance, is not unique to the FEP. For parallel accounts, see Baluška and Levin (2016), Bickhard (2001, 2016), Deacon (2011), and Pezzulo (2008).

4 Markov Blankets, Free Energy, & Generative Models

We have seen in the previous section where the various formal elements of the free energy principle come from. My hope in reviewing the history of the formalism is that this might have helped in building towards the intuition that the principle does not provide an account of how biological systems work, but merely a means by which to model them. In this next section, we will turn to the implementational details of the FEP, where it will be demonstrated that the formalism does not latch onto any features of the real world systems of interest.

4.1 Markov Blankets

For any chain or network of interactions, if the value of any given node is strictly determined by the value of its immediate predecessor or parents, the chain or network is said to exhibit the Markov property. The Markov property captures what is known, in probability theory and statistics, as conditional independence. Two events (or nodes) x and y are conditionally independent if and only if the probability of x given y is equal to the probability of x without knowledge of y.

A Markov chain is a chain of nodes that exhibit the Markov property. A Markov random network or Markov random field (MRF) is an undirected graphical model exhibiting the Markov property. If a Markov chain represents the temporal evolution of a system, there can be no causal interaction between past and future timesteps that bypass the current timestep. In a Markov random field, there can be no non-local interaction: only nearest-neighbors affect one another. The Ising model of ferromagnetism in statistical mechanics is the prototypical Markov random model. Bayesian networks are directed acyclic graphs (DAGs) in which the nodes or vertices can take on one or more numeric values. The most thorough exposition on graphical models of this type, and the text that introduced the Markov blanket and Markov boundary formalism, is Judea Pearl's (1988) *Probabilistic Reasoning in Intelligent Systems.*

A Markov blanket defined on a node (or set of nodes) within a Markov random field will be composed of the nearest neighbors of the node or set of nodes in question. However, on a Bayesian network, or directed acyclic graph, the Markov blanket for any node or set of nodes x are the parents, daughters, and parents of daughters of x. The Markov blanket of x encompasses all of the nodes whose states must be known in order to know the states of x and all of

x's daughters. A Markov boundary consists in the minimal Markov blanket of x; formally, it is the Markov blanket m of x for which no proper subset of m is also a Markov blanket of x. It is useful to differentiate Markov boundaries and Markov blankets because the Markov boundary is the minimal Markov blanket for any node x in question, while any Markov blanket can be expanded to form a greater Markov blanket, so long as some set z can be delineated such that x and x's daughters are conditionally independent of z, given the existence of the Markov blanket, ad infinitum.

Any real-world system or process can be represented as a node in Markov random networks. This says nothing, on the face of it, about the fidelity of representation in a Markov random network. If any real-world system can be represented as a node in a Markov random network, then any real-world system can be represented as possessing or instantiating at least a minimal Markov blanket. Marking the startling inclusivity of Pearl's original (1988) definition, we come to the conclusion that within any non-trivial network, any node ought to have a multitude of Markov blankets. Friston's employment of the Markov blanket construct differs from Pearl's original (1988) construction in the subdivision of the blanket into sensory states—nodes whose influence is directed towards the blanketed node x—and active states—nodes influenced by x. Friston's use of the construct also assumes sensory states to be shielded from any influence beyond internal states (our blanketed node, x). The FEP deals in Markov blanketed systems for which internal and active states parameterise a model of environmental states as indicated by sensory states. Beal (2003) invokes the Markov blanket and Markov boundary constructs within the context of Bayesian networks (directed, acyclic graphs). In a Markov random network, the Markov boundary of a node is just its nearest neighbors. Within a Bayesian network, a Markov blanket will include the parents, daughters, and parents of daughters of any nodes. On a directed graphical model, it also makes sense to speak of the influx and efflux of statistical influence. Influence can be directed inward towards a node shielded by a Markov blanket or boundary, or outwards from the node to the rest of the network, since the vertices or connections between nodes, on a Bayesian network, flow in a pre-specified direction. The directedness of a Bayesian network enables the subdivision of the Markov blanket into what Friston terms sensory and active states. Sensory states are influenced only by external states, while active states are influenced only by internal states. The FEP literature also shifts the Markov blanket from something which applies to only a single node in a graphical model to something which applies to sets of nodes. When the Markov blanket is invoked under the FEP, it is referring to the minimal Markov blanket for any node or set of nodes—what Pearl originally termed the Markov boundary.

How, then, does this Markov blanket (or boundary) get operationalised under the FEP? If we project—or map—some system within a vector space, and this mapping undergoes a linear transformation, the eigenvector of that transformation is a vector that does not change dimensions, but only scales by some scalar λ. It may be helpful to consider that in the mechanics of rigid bodies, the eigenvectors of some transformation (motion of the body) are its principal axes of rotation. Matrices, thus, can be thought of as representing transformations that systems undergo. An adjacency matrix is a matrix representation of a graph, specifying the positioning and relation of the vertices. If we take the time evolution of some system, we can construct an adjacency matrix—call this A—representing the connection weights, interactions, or dependencies of the system for various components or variables over the course of the time for which we have data. We can construct a secondary matrix, B, from this first matrix, $B = A + A^T + A^T A$ (Friston, 2013; 2019). This is the Markov blanket matrix which encodes the parents, daughters, and parents of daughters. Note that the superscripted T is the transpose operation, such that A^T is the transpose of the adjacency matrix. If the node or set of nodes of interest to us is encoded by a binary vector $\chi_i \in 0, 1$, we determine those nodes belonging to the Markov blanket of χ_i by multiplying χ_i by our Markov blanket matrix B (Friston, 2013; 2019). The principal eigenvector of this new matrix $[B \cdot \chi_i]$ specifies the connectivity of each element of the vector. From there, we can select an arbitrary threshold of connection or interaction strength which separates states into blanket states, blanketed states, and blanket-external states.

This is how the Markov blanket construct is operationalised. The upshot is that the power to select systemic boundaries rests, at least in part, on the researcher's intuition. If the Markov blanket formalism were to independently track natural joints, we would need to equip it, at the outset, with some threshold value which would determine precisely the degree of conditional independence that would count—across the board—for the possession of a Markov blanket. If such a threshold existed, and were baked in to the Markov blanket formalism, then we could consider Markov blankets to be in some sense real features of real-world systems. We might discover them, measure them, and count them. We might meaningfully ask whether some existing system does or does not possess a Markov blanket. The threshold for conditional independence, however, is necessarily a post-hoc ascription, and an intuitively guided one, at that. If we can conceive of some *thing* as a discrete thing—as a coherent system—then it is possible to formally represent it as possessing a Markov blanket. There are no Markov blankets to be discovered in nature, and they are not in the business of illuminating natural joints.

4.2 Free Energy, Entropy

As we have seen, in order for a thing, or system, in question to remain a *thing*, it must continue to possess a Markov blanket. If it is to continue to possess a Markov blanket, then the states of its Markov blanket must minimise a statistical quantity termed variational free energy. This free-energy minimisation dynamic can also be interpreted as organisation to non-equilibrium steady state (NESS). In this sense, variational free energy, like the Markov blanket, tracks *thingness*—systemic cohesion. Critically, though, it only does this relative to the thing as we have already defined it. We have imbued the model with our intuitive grasp on what it means for a system to be the sort of system that it is—under what conditions the system continues to be *that system*, and the thresholds over which variables have gone out of bounds so that the system no longer exists as such.

In the mean field approximation, variational free energy was genuinely an approximation of the free energy of a system, viewed from a statistical-mechanical perspective. There is an unfortunate—though understandable—tendency for those first acquainting themselves with the FEP to interpret 'free energy,' and other, similarly confounding terminology from the framework, in physical terms. When we speak of *heat*, *energetics*, and *entropy* it can be difficult to shake the feeling that we are talking about some objective, measurable feature of a material system—particle motion, for example. I took the time at the outset of this paper to trace the history of the framework in Jaynes, Feynman, Hinton, and Beal because 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.

Finally, thermodynamic entropy and Shannon entropy are only equivalent under the generalised Boltzmann distribution, which, it has been argued, applies only at thermal equilibrium (Gao, Gallicchio, & Roitberg, 2019). Thus, in general, information entropy and physical entropy are distinct (Kondepudi, 2013).[3] Living systems are, by definition, far from equilibrium systems. Thus information entropy and thermodynamic entropy do not converge in the regimes of interest to us.

4.3 Generative Models

Early instantiations of the FEP had the FEP as a gradient flow on an ergodic density. Updated expressions of the framework reformulate in terms of a Nonequilibrium Steady State (NESS) density. The interpretation of a system—and its dynamics over time—as entailing a generative model rests on a dual interpretation of the NESS density. This is possible because the high-probability region of the system's phase space (the system's attracting set, or NESS density) is a joint distribution over all the system's variables, rendering it simultaneously a generative model. Under the FEP, the internal states of the system encode what is known as a recognition density, while the system's behaviour, over time, entails a generative model. How complex, how lifelike, and how cognitive a system appears will depend on the timescales of the trajectories that the generative model is implicitly solving (Corcoran, Pezzulo, & Hohwy, 2020). In predictive processing, this quality of the system is referred to as its temporal depth, or its counterfactual depth. It might, then, seem, that the FEP could perhaps be thrown at real world systems, and the degree of biological or cognitive complexity can be read off of features of the systems generative model. The FEP, however, does not dictate what we should write down for a given system's generative model. And once we have written down the generative model of the system under investigation, we have switched from the purview of the FEP to that of a process theory. Like the Markov blanket, the system's implicit generative model is a formal structure that the modeler constructs for the system;

[3] My thanks to Carlo Rovelli for alerting me to this point, and to P. Adrian Frazier for follow-up discussion.

the structure is not somehow emergent from (simulated) data. In some sense, the generative model will intuitively track joints of nature. But it neither finds them, by revealing structure from raw data or simulation, nor defines them, by pre-analytically revealing conceptual boundaries or transition points between classes of systems.

4.4 Recapitulation

To summarise, in this section we have run through a brief history of the key formal elements of the framework, and then examined its machinery to see whether we could pull any mind-independent truths from it. We could not. We found that the framework accommodated certain things very well: the Markov blanket does a very nice job of representing systemic boundaries, the temporal depth of the trajectories being solved by the latent generative models postulated under the FEP is a very elegant and informative representation of something like cognitive complexity, and the free energy parameter itself maps onto a system's attunement with its environmental context, its cohesion and internal consistency (among other things). We also found, however, that the FEP does not 'pick out' or 'discover' these aspects of natural systems—even in silico—but only provides a useful model of them.

5 Reinterpreting the FEP

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. Looking at a few quotes and clarifying how—and how not—to interpret them may help to drive the point home.

Friston (2013) writes that "biological systems must minimise free energy" (p.2), and that "if systems are ergodic and possess a Markov blanket, they will—almost surely—show lifelike behaviour" (p. 11). These sorts of statements could easily lend the impression that free energy minimisation is an objective feature of natural systems and, further, that the framework might cut at some natural joint between more and less 'lifelike' natural systems. However, as we have seen in the previous section, none of the formal elements within the FEP map onto known features of natural systems, and none have the capacity to cut at natural joints.

Ramstead, Badcock, & Friston (2017) 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 nonequilibrium steady state (NESS). If we take the systems attracting set to be a NESS density, then its existence will entail a generative model. 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.

Friston, Da Costa, and Parr (2020) write that "the free energy principle asserts that any 'thing' that attains nonequilibrium steady-state can be construed as performing an elemental sort of Bayesian inference" (p.2). This could be read in one of two ways: It could, quite naturally, be read as an assertion that the free energy principle applies only to systems that fall within the physical regime of nonequilibrium steady-state. This would not, however, be a correct interpretation. For one thing, the notion of NESS at play here is a statistical one. For another, we can take this to mean that the FEP is usefully applied to systems in the physical regime of nonequilibrium steady state.

Physical systems that exist at thermodynamic equilibrium with their external milieus, systems that unresistingly dissipate to equilibrium, and systems that only fleetingly pass through a nonequilibrium state, are not well-captured by the framework. That is to say that it does not make sense to apply the formalism of the FEP to these systems, either because it is trivial or uninformative, or because doing so would prove intractable. A single hydrogen atom at rest in a vacuum is not meaningfully interpreted as performing approximate Bayesian inference over its environmental states. A timeseries of a mere (abiotic) self-organising system, say, a whirlpool, or a candle flame, would show it throwing off its Markov blanket and establishing a new one at every time step (Friston, 2013, 2019). Critically, this does not entail that the FEP cannot, in principle, apply to such systems. It means only that we have no reason to apply it to such systems; we have nothing to gain, epistemically, from applying it to such systems. If the FEP were unable, in principle, to apply to systems outside of NESS, we might expect it to articulate something of the essence of what it is to be a NESS physical system—of what it is to be alive. There has been a strong temptation in the literature to interpret the FEP in this way, as though knowledge about the phenomena of life or cognition could simply be read off from the framework. The math is not an expression of facts about biology, though, but a tool with which to investigate biology. The FEP may function as an aid to scientific work, if only indirectly, by inspiring novel hypotheses and via its process theories. In this way, if successful, the FEP should ultimately serve to reveal features of the systems it was constructed to accommodate. This knowledge, however, is not to be distilled from the framework antecedent to empirical work.

In fact, this precise argument has been made elsewhere, in an analysis of likewise heavily-idealised, normative models in ethology and evolutionary biology. Birch (2017) provides what I take to be the best existing assessment of the status of optimality models in biology and their relation to empirical work. His claim is that optimality models in biology were established to simulate real-world evolutionary dynamics and to generate testable hypotheses. The empirical results of the hypotheses that these mathematical models generate constitute knowledge of the systems in question. The models themselves do not, however, express knowledge about the natural world.

Much of the literature on optimality modelling in biology has been preoccupied with whether or not researchers have sufficient "theoretical justification" for the assumption that organisms maximise Darwinian fitness. Birch (2017) rightfully diagnoses this as a red herring. Theoretical justification only comes into play if we are bent on assessing such models in terms of truth value. As Birch shows, their value to science is not as expressions of truths about the world, but as inspiration for new avenues of empirical research. This is a job that they do quite well, in spite of the obviously false or nonsensical assumptions they rest on. The literature that takes a critical approach to the FEP has likewise fallen for red herrings, under the assumption that the FEP is an articulation of knowledge about the world. Optimality models are "among the most drastic simplifications tolerated anywhere in biology, and they are sometimes criticised for this reason" (Seger & Stubblefield, 1996, p.118). Yet this simplification "elicits questions that might otherwise go unasked" (Seger & Stubblefield, 1996, p.118). The purpose of the FEP is likewise to elicit questions that might otherwise go unasked. Fortunately, there is a rich literature in the philosophy of science to draw upon in characterising exactly these sorts of models.

6 Models

[A]ll models are wrong, but some are useful. – George Box, 1987, p. 424

6.1 The FEP as Scientific Model

The modelling literature lends us a number of plausible interpretations of what the FEP is and does. It may be that the FEP cannot be understood to represent a target system or systems at all, and that it is best leveraged as an analytic tool or studied on its own right as a formal system. If this is the case, then the epistemic value of the FEP cannot be derived from its representational properties, for it would not be understood to have any. Indeed, a unique asset of the modelling literature is that it offers an interpretation of the scientific method that prioritises utility over truth. This comes in two forms: first, the literature on scientific modelling has increasingly come to acknowledge the status of nonrepresentational (Downes, 2011) and non target-directed (Weisberg, 2007) models, and second, even under the presumption that all models represent target systems, the utility of a model is generally understood to stem from idealising, black-boxing, or coarse-graining away from inessential or distracting details of a target system (Wimsatt, 1987). In other words, deliberate misrepresentations or omissions make a good model, not fidelity.

If the FEP can be taken to represent real-world systems, it only does so at such a high level of abstraction as to be unfalsifiable. The elements of the framework do not map onto any known features of real world systems—at least not with any more granularity or specificity than the causal dynamics of such systems. The FEP may offer a proof of principle. It may illustrate conceptual relations between theoretical objects—life and mind, for example. The FEP might be understood not as modelling a specific target system, but as a generic model of a whole class of systems. As such, its usefulness may be in facilitating the unification of several phenomena under a single formal framework. Compare to Newton's work on gravitation: Newton's revolutionary contribution was not in detailing the mechanism underpinning the phenomena associated with gravitation—he remained perfectly agnostic as to the "physical seat" of gravity—but in showing that a range of distinct phenomena—celestial mechanics, the trajectory of a projectile launched from earth's surface, tidal patterns—could all be treated under a single mathematical framework.

While I remain agnostic as to the representational status of the FEP, I follow Birch's (2017) analysis of optimality models in evolutionary biology in arguing that the FEP ought not to be mistaken as constituting knowledge of natural systems.

Yet another role played by highly idealised models such as the FEP is that of a generator for more specific models, either by filling in details or by leveraging tradeoffs between generality and specificity. The FEP is also understood to place demands on the sorts of process theories consistent with it.

The organisation and dynamics of the living organism, the functional architecture of the brain, the structure of human social systems—these are the most complex systems known to exist. The sciences that study these systems are comparatively very young—and may never reach the maturity of the sciences oriented towards far more simple systems. The life, cognitive, neuro, and formal social sciences are still, in many respects, at a stage of trying to get a methodological foot in the door. Highly idealised models, such as normative or optimality models, assist in getting us traction on otherwise intractable phenomena.

6.2 Normative & Process Models

The FEP is introduced by Friston and colleagues as a normative model or normative theory, and contrasted with process models (Allen, 2018; Allen & Friston, 2018; Hohwy, 2020a, 2020b; Schwartenbeck, FitzGerald, Dolan, & Friston, 2013). Active inference, predictive coding, and specific instances of predictive processing are considered to be process models. The role of the FEP is to aid in the generation of such process models, as well as to place constraints on their viability.

6.3 Origins of the Distinction in Mathematical Psychology

Luce (1995) introduces four distinctions to the mathematical modelling of cognitive processes. Process models are contrasted to phenomenological models, normative models differentiated from descriptive models, dynamic models compared to static models, and a distinction drawn between noise and structure. Phenomenological models are similar to phenomenological approaches in physics; they capture gross behaviours and attributes without specification of internal structure or speculation into underlying causes or mechanisms. A process model, on the other hand, opens the black box of the mind. It attempts to understand some cognitive process in terms of the flow of information in the brain—though such information processing models come in greater and lesser degrees of neurobiological realism. Luce notes that "most mathematical modelers, although sometimes inspired by neural data, postulate mechanisms far more abstract and functionally defined than are found at the neural level" (1995, p.10). Under a normative model, it is presumed that reasoning should accord with formal logic, induction and beliefs with, e.g., Bayesian dictates for inference and credence, or decision-making with the results of optimising a utility function. A descriptive model, in contrast, represents the cognitive process of making choices, reasoning through problems, and drawing inferences as it is observed to happen—messy, sub-optimal, and irrational though it may be.

What, then, is the significance of taking the FEP to be a normative model, in Luce's sense? Friston and colleagues stress that the FEP, as process theory, is not falsifiable (Allen & Friston, 2018). It will not be possible to articulate a version of the FEP that can be held up against some real world process in such a way as to undermine or legitimate the model. The FEP will not directly generate predictions, tests, or hypotheses. Allen and Friston pose an intriguing question "if the FEP is unfalsifiable...is it uninformative?" (2018, p.2476). The answer they provide is that "[t]he FEP is uninformative" inasmuch as it can neither explain nor predict specific observations (Allen & Friston, 2018, p.2476). However, they emphasise that the FEP informs "the viability and sufficiency of...process theories" (Allen & Friston, 2018, p.2476). Allen (2018) also alludes to the role of the FEP in generating such process theories. On Allen's (2018) articulation, such process theories will fill in the mechanistic details that are lacking in the FEP. Not everyone, though, has so readily accepted that the FEP will prove useful in producing or weeding out process theories in this way. Friston, FitzGerald, Rigoli, Schwartenbeck, and Pezzulo (2017) note that "the enthusiasm for Bayesian theories of brain function is accompanied by an understandable skepticism about their usefulness, particularly in furnishing testable process theories" (Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017, p.2). In the section that follows, the philosophical literature on adjacent sorts of models will be reviewed, in the hopes that this will lend a sense of how it is that a highly-abstracted, unfalsifiable formal model such as the FEP can have scientific utility. Ultimately, of course, whether the FEP turns out to be useful in this way will be an empirical matter—in both senses.

In characterising the free energy principle, Friston draws a distinction between normative models and process models (Schwartenbeck, FitzGerald, Dolan, & Friston, 2013). The FEP, as Friston presents it, is a normative model; active inference, predictive processing, and predictive coding are process models. Friston's is an adaptation of a quadripartite distinction introduced by Luce in 1995, for the purpose of characterising the aims and scope of mathematical models in cognitive psychology. Per Luce's description, normative models characterise psychological processes in such a way as to render them in conformity with logical or rational norms. This is in contrast to descriptive models, which faithfully represent the messiness of human cognition. Phenomenological models provide an outward behavioural description, without 'opening the black box,' so to speak. Process models, on the other hand, provide specification of internal structure, mechanism, and information flow. Friston's employment of the distinction between normative and process models goes further than this, however. It implies a relationship of something like methodological supervenience between normative models and process models. The FEP is thus an umbrella-framework out of which predictive processing, predictive coding, and a version of active inference, its process models, fall with decreasing abstraction and increasing granularity. As a normative model, the FEP is intended to aid in the generation of process models, and to furnish constraints on viable process models. The FEP is, itself, however, not beholden to empirical data. Its virtues are not in its verisimilitude.

6.4 Models in Philosophy of Science

Very little can be said about the totality of scientific models, or the practice of scientific modeling as a whole. Perhaps the only uncontroversial generalisation that can be made is that models are scientific tools that are useful precisely because they are inaccurate or outright false (Wimsatt, 1987). Many scientists and philosophers of science interested in modeling practice have chosen to proceed by carving up the space of scientific modeling along multiple dimensions before attempting to pin down the relationship of models to theory, explanation, knowledge, data, and ultimately, real-world systems. In this section, I will draw from literature in the philosophy of science on varieties of scientific model that share informative resemblances with the FEP.

6.5 Exploratory Models

There is an obvious resemblance between the FEP, conceived of as a normative model, and what Frigg and Hartmann (2020) call textitexploratory modelling. They characterise exploratory models as "models which are not proposed in the first place to learn something about a specific target system or a particular experimentally established phenomenon" (Frigg & Hartmann, 2020). Frigg and Hartmann draw on Gelfert's (2016) depiction of exploratory modeling. Gelfert (2016) claims that exploratory modelling can serve four distinct epistemic aims. These include, in the first place, assessing the suitability of a target system, when the specific nature of the target system—and how to delineate it from other potential systems of study—is obscure. A second aim of exploratory modelling is proof-of-principle demonstration, which can involve either a conceptual demonstration of some mapping relation between model and target system or between phenomenon and underlying mechanism. An exploratory model might also generate a potential—in contrast to a necessary—account of some system of interest. This is dubbed a 'how-possibly' form of explanation. Lastly, exploratory models can further research aims in painting a picture of some class of natural phenomena in exceedingly broad brushstrokes, which will serve as a taking-off point for later investigation, often by way of more fine-grained models. An exploratory model, in Gelfert's words, is able to provide "conceptual clarity" while "staying largely clear of substantive ontological commitments with respect to the precise nature of the model's constituents" (Gelfert, 2016, p.88). Frigg and Hartmann's conception of an exploratory model is also informed by Massimi's (2019) work. Massimi (2019) argues for two kinds of exploratory model: models in which the target system is hypothetical, and models in which the target system is nonexistent. The chief epistemic virtue of exploratory modeling, on Massimi's account, is that it provides modal knowledge.

6.6 Modelling with and without Specific Targets

Weisberg (2013) has likewise laid out a useful taxonomy for scientific models: he differentiates concrete, mathematical, and computational models. Cutting across this tripartite distinction is a division between what he calls target-directed modeling and modeling without a specific target. In this latter category, we find generalised modeling, minimal modeling, hypothetical modeling, and targetless modeling. A target-directed model is one built for the purpose of predicting, controlling, or explaining the behaviour of a specific system under specific conditions. For example, we might have data pertaining to the foraging behaviour of a particular species of ant across a number of colonies observed in the amazon rainforest, and we might construct a model for the purpose of discovering something about the ethology and ecology of this particular ant species. Then again, we might want to think about foraging behaviour much more broadly and abstractly—divorced from the minutiae of particular populations of particular species in particular ecological settings. Considering the problems posed by foraging in a patchy habitat as faced by organisms in abstract terms, we might perceive a resonance between foraging and economic models that deal with optimising decisions in scenarios in which we face diminishing returns. We might come up with a heavily-idealised, organism and context-generic model of foraging. This latter approach is what Weisberg dubs *modeling a generalised target.*

6.7 Targetless Models

Targetless modeling is another form of modeling without a specific target. The one distinguishing feature of targetless modeling, as Weisberg portrays it, is that the model is never brought in direct contact with the results of empirical science: "[t]he only object of study is the model itself, without regard to what it tells us about any specific real-world system" (Weisberg, 2013, p.129). The targetless model is itself never used in the generation of hypotheses or predictions, and never fitted to data. The construction and analysis of targetless models is "most akin to pure mathematical analysis" (Weisberg, 2013, p.129). Frequently, such models are used to scaffold explanations of a very general form of natural phenomena at a very high level of abstraction. If successful, these models often give rise to corollaries, which themselves will interface with data. But the power of the targetless model lies in the fact that it is untethered from real-world systems. It still, like the generalised model, represents some broad class of natural phenomena. Unlike the generalised model, though, the targetless model is intended to be studied as an object of scientific intrigue in its own right, independently from empirical results. Weisberg presents a class of computational models known as cellular automata as exemplary of targetless models. In particular, he focuses on a model of cellular automata known as Conway's game of life.

6.8 Models & Simulations

Barandiaran's work is especially well tailored to our needs in assessing the FEP because Barandiaran has focused his career—and his analysis of the science of simulations—on theories and models falling under the umbrella of what are called life-mind continuity approaches: that is, approaches that seek to discover, represent, or evaluate dynamical principles common to the simplest of biological organisation and complex cognitive systems. Life-mind continuity approaches exist on the fringes of both theory and empirical investigation. They occupy something of an uncomfortable territory, being at once both highly specific and extremely abstract. Developing an explanation of colour vision in mantis shrimp or a model of locomotion in motor proteins proceeds relatively straightforwardly, and can be evaluated relatively straightforwardly. There is often a clear and direct connection between the theories and models scientists come up with and the data they collect. Life-mind continuity approaches, however, being high-level and existing, as they do, in empirically uncharted waters, rely heavily on simulation work. These approaches—and the technologies that enable them—are also relatively new. For this reason, philosophers seeking to critically evaluate the nature and results of work in this area—like Barandiaran—have had to construct their own accounts of how this work proceeds.

6.9 Generic & Conceptual Models

Barandiaran (2008; Barandiaran & Chemero 2009) has developed an in-depth analysis of computer simulation in the life and cognitive sciences—in particular, artificial life, or alife models. He offers us a taxonomy for these, which includes functional models, mechanistic-empirical models, generic models, and conceptual models. These models are classified, in part, by what they are meant to represent, and how they are meant to be evaluated. The first two—mechanistic-empirical and functional models—are empirically evaluated, that is, evaluated against some data collected from some real-world system. The second two—generic and conceptual models—are evaluated theoretically, meaning that they are evaluated against theory or formalism. Generic models "serve to discover or classify generic properties of complex systems" (Barandiaran, 2008, p.53). Conceptual models, on the other hand, explain by means of their resonance with theoretical concepts. "Abstract conceptual models," writes Barandiaran "are used to formalise or compare definitions of generic concepts (such as emergence, complexity or hierarchy) while domain specific conceptual models are used to explore the role and interaction between more specific concepts (such as learning, plasticity, evolvability, etc.)" (Barandiaran, 2008, p.54). I propose that we can develop insights for evaluating the FEP from both conceptual and generic models.

6.10 Generic Models

Barandiaran refers to generic models as "computational constructions or template[s]" which make "no particular reference to any specific object of study, but whose formal structure has been selected in virtue of their resemblance with a wide range of natural phenomena" (Barandiaran, 2008, p.58). Deployment of generic models stands to benefit research in revealing "generic abstract properties of complex systems" (Barandiaran, 2008, p.58). Examples of generic models include cellular automata, neural networks, dynamical systems models, and domain-general models of self-organisation. Usually a generic model will have started its life as a domain-specific empirical model. When some feature of the model is seen to generalise beyond its original domain of application, the model is rinsed of its target-specific details and rendered generic. We have seen that this is the case for many of the formal elements of the FEP: they originated in empirical domains with specific usage, and were rendered general by the transition to a purely statistical or information-theoretic usage. The FEP, however, is a model sewn-together from many such elements. In this respect, it differs from Barandiaran's characterisation of generic models.

6.11 Conceptual Models

Conceptual models are "a tool to question and reorganise theoretical assumptions and concepts" (Barandiaran, 2008, p.59). They further scientific understanding by allowing us to explore the relationships between the notions at play in our theories. The FEP allows us to do this with notions such as life, identity, health, prediction, cognition, perception, and the like. The relevant relationship of similarity in conceptual modeling is not between the formal structure of the model and some target system, but between the formal structure of the model and a conceptual structure. Conceptual models are an aid to theory construction, to definition building, and to establishing proofs of principle. But, as Barandiaran notes, prominent evolutionary biologist John Maynard Smith decried A-life models, referring to the approach as "science without facts" (Quoted in Horgan (1995)). This gets at another aspect of this array of models: scientists and philosophers are often loath to acknowledge their scientific merit until they have allowed us to make headway in some domain. For example, Hohwy (2020) argues that the FEP is best understood as offering “a conceptual and mathematical analysis of the concept of existence of self-organising systems” (p.1).

6.12 Alternative Epistemic Virtues

What these various models of modeling share in common is the understanding that certain types of models are useful not in virtue of accurate—or even inaccurate—representation of features of some real-world system, but in virtue of epistemic virtues seemingly orthogonal to truth. We have seen that such models can proceed by offering proofs of principle. Such models can also provide knowledge of counterfactual scenarios—means of exploring the results of manipulating certain systems in ways that we are unable to manipulate such systems in real life, or in-principle explanations of phenomena whose mechanisms are unknown. Models of this sort can also serve as an abstract or analogical touchstone, or an entry point, into an unexplored domain. This will serve as the basis for more targeted, more fine-grained modelling work, or empirical investigation, later on. Models in this genre also aid in the process of theory construction, by allowing exploration of the nature and interrelation of conceptual objects, and by inspiring theorists to draw connections and to ask questions that would otherwise go undrawn or unasked. Such models offer leverage where other scientific methods stop short: systems and dynamics far too complex, or too new, to be treated under standard approaches.

6.13 Guides to Discovery

One way of encapsulating the diverse functionality of this genre of modelling, as outlined above, is by thinking of these models as *guides to discovery.* This is a notion developed in Chemero (2000, 2009), though phrased perhaps most succinctly and articulately in Barandiaran and Chemero (2009): “A guide to discovery is some means for a scientific research program to advance by making predictions for future experimentation, or extending the reach of the program to new phenomena, or solving conceptual problems within the program, or casting empirical findings in a new light, and so forth” (Barandiaran & Chemero, 2009, p.288). This is a function that models like the FEP are uniquely equipped to perform.

6.14 Takeaways from the Modelling Literature

We have seen in this section that, at least according to prominent contemporary work in philosophy of science, there are many ways for the FEP to serve as an aid to scientific work without constituting falsifiable assertions about the state of nature. It is capable of serving as a proof of principle demonstration, as a tool for conceptual analysis—for example, of the notions of organismality, of systemic identity, or of cognition—it is capable of unifying a number of phenomena that have hitherto been investigated separately under a single formal framework, which may pave the way for future empirical investigation of commonalities that run between the phenomena involved. The FEP can play an essential a role in generating more concrete models, and as a means for evaluating their viability.

A very important question, however, remains unsettled: is the FEP merely a mathematical truism, or does it make assertions about nature, albeit at a very high level? The modelling literature I have included above, and my discussion of it, has been deliberately agnostic with respect to this issue. I would like to draw again from a parallel discussion on formal models in evolutionary biology. Mathematical biologist Martin Nowak has been arguing for decades that a mathematical model known as Hamilton's rule does not describe a synthetic, biological truth, but rather an analytic, mathematical truth; one that would hold true of any dataset due to the nature of regression coefficients (Nowak, McAvoy, Allen, & Wilson, 2017). In a 2018 lecture on the topic, Nowak analogised use of Hamilton's rule with standing over the shoulder of a laboratory scientist and repeating over and over again that two plus two equals four.

Apart from their entertainment value, Nowak's comments pose an intriguing puzzle for philosophers and scientists concerned with the menagerie of mathematical models employed by scientists, and the exceptional diversity among them. How do we distinguish models that make assertions about nature from those with no empirical strings attached? What makes a model a scientific model, and what sets it apart from 'mere math?' Plenty of mathematical models used by scientists are representative of states of affairs in the world which are, at least in principle, open to empirical verification or falsification—although models, insomuch as they rest on facts or assumptions about the natural world, are in general not in the game of falsification. Some of these representational models express only facts that have already been empirically verified; others rest on gambits, speculations, useful fictions, or idealisations. Then there are mathematical models which, like '2 + 2 = 4,' express only mathematical truths. The power law, $\frac{1}{r^2}$, most famously detailed in Newton's *Principia*, was clearly derived, and several times over, at that, from calculations based on astronomical observations. But from the perspective of modern physics, it seems as though this formula might be mathematically necessary.

Our understanding of the status of scientific models shifts over historical time. And there are a plethora of models that exist in a sort of superposition between mathematical truism and contingent fact or gambit. Hamilton's principle of least action—Friston often compares the status of the FEP with the LAP—is neither a mathematical truism nor a statement of fact. It makes reference to worldly systems, yet is unfalsifiable because it holds necessarily for all systems—possible and actual.

In this paper, I have remained agnostic with respect to the issue of whether the FEP is merely a mathematical truism or a representation of contingent states of affairs in nature in the first place because I think that the results are not yet in. More work needs to be done on the framework before it can be known whether it applies universally and necessarily, or only under certain conditions. In the second place, I avoid coming down on this issue because I believe it to be wholly within the range of possibility that we should, in the end, discover that the FEP is something of an intermediary between pure math and contingent representation of nature. In this respect, the existence of the FEP may well come to push the boundaries of what the literature on scientific models has heretofore dealt with.

7 Conclusion

I believe that the FEP can provide a powerful framework for modelling living systems across scales. It can do this in demonstrating proofs of principle, by placing demands on the sufficiency of process theories, and by serving as a structure or template from which narratives are woven—the sort of narratives that enable entrance into hitherto empirically uncharted territories. Already the FEP has been leveraged in the construction of more granular models of specific phenomena: from the behavioural profile of autism spectrum disorders (Perrykkad & Hohwy, 2020a, 2020b) to morphogenesis and regeneration in amphibians (Friston, Levin, Sengupta, & Pezzulo, 2015; Kuchling, Friston, Georgiev, & Levin, 2019) to the phenomenal experience of selfhood under the action of psychedelic drugs (Carhart-Harris & Friston, 2019; Deane, 2020). The FEP has also enabled important conceptual work on the nature of life (Ramstead, Badcock, & Friston, 2018), cognition (Corcoran, Pezzulo, & Hohwy, 2019), and the continuities between them (Kirchhoff & Froese, 2017).

Figuring out what the FEP is, and what use it holds for scientists, is a worthwhile project in and of itself. How we come down on the matter—and what route we take in getting there—will have important implications for philosophy of biology, philosophy of cognitive science, and the philosophy of science in general. The dialogue unfolding in the literature on the FEP raises important questions about the relationship between science and philosophy of science, between theory, model, and data, about the scientific method and the aims of science, and even—perhaps most especially—about what counts as science in the first place.

8 Acknowledgements

This paper has benefited tremendously from thoughtful feedback and lengthy discussions with many people, most notably Tony Chemero, Thomas van Es, Karl Friston, Bryce Huebner, Alex Kiefer, Michael Kirchhoff, Thomas Parr, and Maxwell Ramstead.

9 References

1. 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.

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

3. Badcock, P. B., Friston, K. J., & Ramstead, M. J. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of life Reviews, 31*, 104-121.

4. Baluška, F., & Levin, M. (2016). On having no head: cognition throughout biological systems. *Frontiers in Psychology, 7*, 902.

5. Barandiaran, X. (2008). Mental life: A naturalized approach to the autonomy of cognitive agents. Unpublished Dissertation, Universidad del País Vasco, San Sebastian, España.

6. Barandiaran, X. E., & Chemero, A. (2009). Animats in the modeling ecosystem. *Adaptive Behavior, 17*(4), 287-292.

7. Beal, M. J. (2003). Variational algorithms for approximate Bayesian inference (Doctoral dissertation, UCL (University College London)).

8. Bickhard, M. H. (2001). Function, anticipation, representation. In *AIP Conference Proceedings* (Vol. 573, No. 1, pp. 459-469). American Institute of Physics.

9. Bickhard, M. H. (2016). The anticipatory brain: Two approaches. In *Fundamental issues of artificial intelligence* (pp. 261-283). Springer, Cham.

10. Birch, J. (2017). Fitness Maximization. In Joyce, R. (Ed.), *The Routledge Handbook of Evolution and Philosophy.* Routledge.

11. Box, G. E., & Draper, N. R. (1987). *Empirical model-building and response surfaces.* John Wiley & Sons.

12. Carhart-Harris, R. L., & Friston, K. J. (2019). REBUS and the anarchic brain: toward a unified model of the brain action of psychedelics. *Pharmacological Reviews, 71*(3), 316-344.

13. Chemero, A. (2000). Anti-representationalism and the dynamical stance. *Philosophy of Science, 67*(4), 625-647.

14. Chemero, A. (2009). *Radical Embodied Cognitive Science.* Cambridge: The MIT Press.

15. Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science, 1*(2), 89-97.

16. Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2019). From Allostatic Agents to Counterfactual Cognisers: Active Inference, Biological Regulation, and The Origins of Cognition. *Biology & Philosophy, 35(32).*

17. Deacon, T. W. (2011). *Incomplete Nature: How Mind Emerged from Matter.* W. W. Norton & Company.

18. Deane, G. (2020). Dissolving the self. *Philosophy and the Mind Sciences, 1*(I), 1-27.

19. Downes, S. M. (2011). Scientific models. *Philosophy Compass, 6*(11), 757-764.

20. Downes, S. M. (2020). *Models and Modeling in the Sciences: A Philosophical Introduction.* Routledge.

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

22. Frigg, R., & Hartmann, S. (2020). Models in Science. Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.*

23. Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*(7), 293–301.

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

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

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

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

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

29. Friston, K. J., Fagerholm, E. D., Zarghami, T. S., Parr, T., Hipólito, I., Magrou, L., & Razi, A. (2020). Parcels and particles: Markov blankets in the brain. arXiv preprint, arXiv:2007.09704.

30. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. Neural computation, 29(1), 1-49.

31. Friston, K., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex, 68*, 129-143.

32. Friston, K., Levin, M., Sengupta, B., & Pezzulo, G. (2015). Knowing one's place: a free-energy approach to pattern regulation. *Journal of the Royal Society Interface, 12*(105), 20141383.

33. Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *Neuroimage, 34*(1), 220-234.

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

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

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

37. Gelfert, A. (2016). *How to Do Science with Models: A Philosophical Primer* (Springer Briefs in Philosophy). Cham: Springer International Publishing.

38. Gao, X., Gallicchio, E., & Roitberg, A. E. (2019). The generalized Boltzmann distribution is the only distribution in which the Gibbs-Shannon entropy equals the thermodynamic entropy. *The Journal of chemical physics, 151*(3), 034113.

39. Hinton, G. E., & 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).

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

41. Hohwy, J. (2016). The Self-Evidencing Brain. *Noûs, 50*, 259-285.

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

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

44. Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review, 106*(4), 620.

45. 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.

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

47. 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.

48. Kondepudi D. (2013) Chemical Thermodynamics. In: Runehov A.L.C., Oviedo L. (eds) *Encyclopedia of Sciences and Religions.* Springer, Dordrecht.

49. 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.*

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

51. MacKay, D. J. (1995b). Ensemble learning and evidence maximization. In Proc. Nips (Vol. 10, No. 1.54, p. 4083).

52. MacKay, D. J. (1995c). Free energy minimisation algorithm for decoding and cryptanalysis. *Electronics Letters, 31*(6), 446-447.

53. Massimi, M. (2019). Two Kinds of Exploratory Models. *Philosophy of Science, 86*(5): 869–881.

54. Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. Jordan MI (ed.) In Learning in Graphical Model.

55. Nowak, M. A., McAvoy, A., Allen, B., & Wilson, E. O. (2017). The general form of Hamilton's rule makes no predictions and cannot be tested empirically. *Proceedings of the National Academy of Sciences, 114*(22), 5665-5670.

56. Palacios, E. R., Isomura, T., Parr, T., & Friston, K. (2019). The emergence of synchrony in networks of mutually inferring neurons. *Scientific reports, 9*(1), 1-14.

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

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

59. Perrykkad, K., & Hohwy, J. (2020a). Fidgeting as self-evidencing: A predictive processing account of non-goal-directed action. New Ideas in Psychology, 56, 100750.

60. Perrykkad, K., & Hohwy, J. (2020b). Modelling me, modelling you: the autistic self. Review Journal of Autism and Developmental Disorders, 7(1), 1-31.

61. Pezzulo, G. (2008). Coordinating with the future: the anticipatory nature of representation. Minds and Machines, 18(2), 179-225.

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

63. Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2019). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, 1059712319862774.

64. Ramstead, M. J., Veissière, S. P., & Kirmayer, L. J. (2016). Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention. *Frontiers in psychology, 7*, 1090.

65. Seger, J., & Stubblefield, J. W. (1996). Optimization and adaptation. In: Rose,, M. R., & Lauder, G. V. (Eds): *Adaptation*, pp 93–123. San Diego: Academic Press.

66. Schwartenbeck, P., FitzGerald, T., Dolan, R., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in psychology, 4*, 710.

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

68. Weisberg, M. (2007). Three kinds of idealization. The journal of Philosophy, 104(12), 639-659.

69. Wimsatt, W. C. (1987). False models as a means to truer theories. In M. Nitecki and A. Hoffmann, (Eds.), *Neutral models in biology*, 23–55. Oxford: Oxford University Press.