The Literalist Fallacy & the Free Energy Principle: Model-building, Scientific Realism and Instrumentalism

Michael D. Kirchhoff[1], Julian Kiverstein[2] & Ian Robertson[3]

Affiliation:
1. University of Wollongong, School of Liberal Arts, Australia.
2. Amsterdam University Medical Centre, Department of Psychiatry, The Netherlands.
3. University of Wollongong, School of Liberal Arts, Australia

Emails:
1. kirchhoff@uow.edu.au
2. j.d.kiverstein@amsterdamumc.nl
3. ianrob@uow.edu.au

Abstract: Disagreement about how best to think of the relation between theories and the realities they represent has a longstanding and venerable history. We take up this debate in relation to the *free energy principle* (FEP) - a contemporary framework in computational neuroscience, theoretical biology and the philosophy of cognitive science. The FEP is very ambitious, extending from the brain sciences to the biology of self-organisation. In this context, some find apparent discrepancies between the map (the FEP) and the territory (target systems) a compelling reason to defend instrumentalism about the FEP. We take this to be misguided. We identify an important fallacy made by those defending instrumentalism about the FEP. We call it the *literalist fallacy*: this is the fallacy of inferring the truth of instrumentalism based on the claim that the properties of FEP models do not literally map onto real-world, target systems. We conclude that scientific realism about the FEP is a live and tenable option.

Keywords: Free energy principle; Scientific realism; Instrumentalism; Model building; Idealisation; Approximation

1. Introduction

Different kinds of problems beset scientific and philosophical inquiry. One concerns scientific realism. Scientific realism is the view that one reasonable goal of our currently best scientific theories and models is to offer literally true (or probably true, or approximately true) descriptions and explanations of what reality is like. Conversely, instrumentalism (or, scientific antirealism) is the view that scientific theories or models are nothing but instruments for the prediction and systematisation of target systems. Scientific models, on the latter view, *do* have explanatory power, and therein lies their utility. Yet, they are *not* truth-producing in the sense of representing parts of real-world, target systems (or, they do not have the aim of being truth-producing in the long-run).

In this paper, we take up this debate in relation to the free energy principle (FEP) - a mathematical framework in computational neuroscience, theoretical biology and the philosophy of cognitive science. We shall ask: what is the relationship between the scientific models constructed using the FEP and the realities these models purport to represent? Our focus will be specifically on the FEP and what, if anything, it tells us about the systems it is used to model. We call this issue *the map problem*: how does the map (theory, model) relate to the territory (real-world, target system) of which it is a map?

The FEP is a mathematical framework that postulates the characteristics any organism must have for it to exist (Constant 2021; Friston 2013; Kirchhoff et al. 2018). It states that any self-organising system that maintains a nonequilibrium steady-state with its environment must minimise its free energy (Friston 2010). The term ‘minimising free energy’ is a technical term taken from statistical physics and machine learning. It can be conceptualised as a way of maximising the likelihood of sensory input to a system or, equivalently, of maximising the evidence for a *model* - where the model is conceptualised as a phenotype (Friston 2012, 2013; Ramstead et al. 2018; Kirchhoff et al. 2018; Kirchhoff 2015). The FEP has been proposed as the basis for a grand unifying theory for the biological and cognitive sciences identifying mathematical principles that can be applied to model a variety of biological systems at all scales of organisation from cells and multicellular organisms to cognitive processes such as perception, planning, action and memory (Ramstead et al. 2019; Hesp et al. 2019).[1]

The FEP has received a lot of attention both in the sciences as well as in philosophy. The last two years have seen a spike in papers defending instrumentalism about the FEP. Instrumentalists about the FEP rely on the following kind of argument:

1. Scientific models introduce distortions into the representations of target systems via idealisation and approximation;
2. Scientific realism requires that models provide descriptions of target phenomena that are literally true.
3. Scientific models are not true and accurate representations of their targets.
4. *Therefore*, scientific realism about models is false.

[1] See Mann et al. 2021 for an introduction and user-guide to the FEP.

In this paper, we shall show that this kind of argument is problematic. We argue that inferring instrumentalism about the FEP on the basis of this kind of argument rests on a fallacy. We call it the *literalist fallacy*: this is the fallacy of inferring the truth of instrumentalism based on the claim that the properties of FEP models do not literally map onto real-world, target systems. We now turn to briefly highlight the main motivations for instrumentalism about the FEP and illustrate how they all commit the literalist fallacy.

Ramstead et al. (2020) claim that "instrumentalist accounts in the philosophy of science suggest that scientific models are useful fictions: they are not literally true, but "true enough," or good enough to make useful predictions about, and act upon, the world." (2020, p. 4) This misconstrues the distinction between scientific realism and instrumentalism. The reason is that scientific realism is not - or, at least not necessarily - the claim that models must be literally true: they can be probably true, partially true, approximately true or probably, approximately true (Stanford 2003). To assume that scientific realism is the view that models must be literally true of their targets is an instance of the literalist fallacy. van Es & Hipolito (2020) state that "it remains disputed whether its [the FEPs] statistical models are scientific tools to describe non-equilibrium steady-state systems (which we call the instrumentalist reading) or are literally implemented and utilized by those systems (the realist reading)." (2020, p. 1) They conclude that since FEP models are not true and accurate descriptions of their target systems, instrumentalism about the FEP is the only option. However, accepting that organisms do not literally embody the mathematics of the FEP does not ground the claim that scientific realism about the FEP is false. To claim that it does is to commit the literalist fallacy. Bruineberg et al. (2021) focus on the ontological status of the Markov blanket formalism in the FEP.[2] They argue that much of the literature on the FEP implies that organisms literally instantiate the mathematical structure of Markov blankets. They argue that such a use of the Markov blanket formalism conflates a model with its target system. In part to avoid such a mistake they settle for an instrumentalist reading of the Markov blanket formalism as a potentially useful descriptive tool that does not have the aim of truthfully representing anything about target systems. This too is an instance of the literalist fallacy. Finally, Colombo & Palacios (2021) target the issue of ergodicity in the FEP; namely, that one can realistically model biological systems as having an ergodic density.[3] They deny that ergodicity captures properties of biological systems. However, modelling systems as ergodic is a modelling choice. It should not be

[2] Markov blankets are used to define statistical relations of conditional independence between nodes of a network. In the context of the FEP, the Markov blanket construct is used to define a boundary for a system where the relations of conditional independence allow for a distinction to be made between states that are internal to the system and those that are external. More on this in section five.

[3] Ergodicity refers to the time average of any measurable function of a system such that the system has a high probability of converging on this time average given sufficient time. If a system is ergodic then the probability of the system being in a state of a given value when observed at random is equivalent to the average amount of time the system occupies this state. More on this in section seven.

mistaken as evidence for the claim that systems return to the exact same states in their phase space. To insist otherwise would be to commit the literalist fallacy.

The structure of the paper is as follows: In section two, we take up the question of what kind of model the FEP is. In section three, we consider whether the FEP in virtue of being an approximate and idealised model is incompatible with scientific realism. These two issues we think are worthwhile getting some initial traction on before turning to the specific arguments for instrumentalism about the FEP. In the rest of the paper we discuss the central instrumentalist arguments for the FEP. Section four focuses on inferring instrumentalism about the FEP on the basis of variational free energy. Section five considers how the Markov blanket formalism has been used in the discussion over scientific realism. Section six looks at the topic of approximate Bayesian inference. Finally, section seven targets an argument suggesting that modelling biological systems as being ergodic fails to model anything biologically realistic about such systems. All of these arguments commit the literalist fallacy. We conclude that scientific realism about the FEP is a live and tenable option.

2. What Kind of Model is the FEP?

In what follows, we follow Weisberg's (2013) dual-aspect account of what comprises a scientific model: it has a particular (concrete or mathematical or computational) *structure* and an *interpretation* of that structure (see also Andrews 2021).

The question now is whether the FEP, given Weisberg's account, can be thought of as a scientific model. This turns out to depend on the *practice* of modelling involved in the FEP. Say one builds a model of a yet-to-be-built bridge in some appropriate scale. Here, the model would have a material or *concrete* structure (Weisberg 2013). The FEP does not offer anything comparable in terms of concrete structure. Instead, the FEP is best understood as a *mathematical* structure. It is a structure that expresses the dynamics of any system in terms of equations for random dynamical systems with a Markov blanket. On the basis of this, the dynamics of external states, internal states and blanket states are captured by a gradient ascent on what is called an 'ergodic density' (Palacios et al. 2020). All of this is mathematics and can be applied to any random dynamical system (i.e. any dynamical system whose equations of motion have an element of randomness due to noise). This would be part of the mathematical structure upon which the FEP is grounded (see Friston 2019 for full details).

Andrews (2021) suggests that the FEP is an instance of what in the modelling literature is called a *targetless model*. In targetless modelling, the "object of study is the model itself, without [direct] regard to what it tells us about any specific real-world system. This type of modelling is most akin to pure mathematics." (Weisberg 2013,

p. 129) As we understand the proposal, Andrews draws this conclusion about the FEP because they think the FEP is *just its mathematical structure*. Were one to apply the definition of a scientific model provided by Weisberg (2013), this purely mathematical rendering of the FEP would discount it as a scientific model, in which case the map problem would not arise. (The map problem recall is the question of how the FEP as a scientific model relates to target systems it describes.)

There is more to be said about models in general, and the FEP especially. In both the concrete and mathematical case, the model structure must itself be *interpreted* by someone as, e.g., a scale model of a bridge. Or, in the case of the FEP, as a model of self-organisation in open systems. The FEP, as a mathematical structure, is a set of truisms. The truisms licence an interpretation of characteristic states of being (e.g., homeostatic states). Terms such as *free energy minimisation* and *maximisation of model evidence* can be given mathematical precision and used to describe how organisms are able to return to preferential states (e.g., maintaining core body temperature). Therefore, the FEP comprises both a mathematical structure and an interpretation of this structure. This would be consistent with Weisberg's view of scientific models. Moreover, it is not obvious that the FEP is a targetless model. Indeed, even when simulations of free energy minimising dynamics look constructed for the purposes of exploring the mathematics of the structure itself, there is almost always the ambition of saying something about real-world systems (see, e.g., Friston 2010). Andrews is correct in pointing out that the FEP falls under the larger grouping of approaches to modeling without a *specific* target (in the sense of Weisberg 2013). Yet, for the reasons here specified, the FEP is a better example of what Weisberg calls *generalised modeling*. The target in this form of modeling is not specific; it is more abstract. For example, a generalised model of sex reproduction will not say anything about sex reproduction in Australian wombats. There will not be any specific species modelled by a generalised model. Similarly, when the FEP uses the Markov blanket formalism to derive a proof of concept of how systemic states can be differentiated from environmental states (a necessary requirement for existence), the model one gets is a general model. It concerns a general (and statistical) delineation problem and does not refer to any specific system.

Models can have a generalised target; or no target. They can have multiple targets, and so on. In contrast to models without a target (e.g., a targetless model), a target-directed model has a specific target system it represents (Weisberg 2013). Or, the modeller has a specific target system in mind - e.g., maze navigation in rats or chemotaxis in Escherichia coli. The FEP becomes a target-directed model once implemented by a process theory - e.g., *active inference* (Friston et al. 2017). Unlike the mathematics of the FEP, the process theory provides a possible (mechanistic) story about how the FEP is implemented in real-world, target systems (see e.g. Tschantz et al. 2021 for an active inference model on bacterial chemotaxis). Once a process theory of the FEP is proposed, the map problem arises. Should we conceive of the process theory as a representation of the true behaviour of a target system? It is the application of the FEP under active inference that leads to the map problem.

3. Idealisation, Scientific Realism and the Free Energy Principle

In this section, we shall introduce several key issues that will inform the discussion of the instrumentalist arguments to come: idealisation and approximation. As we shall see, FEP models are both idealised and approximate models. What does this tell us about scientific realism with respect to the FEP?

Idealisation and approximation are seen by many as central to scientific progress. Yet, there is no consensus on how to understand the relation between the concepts of idealisation and approximation. What certainly *can* be said is that idealisation involves a deliberate simplification or distortion of some phenomenon of interest into a scientific model, or theory (e.g., variational free energy in the FEP). Approximation involves representing some target system inexactly (e.g., Bayesian inference in the FEP). Here we focus on idealisation (and return to approximation later).

Idealisation looks to present a problem for any kind of scientific realism. Consider this formulation of scientific realism by Godfrey-Smith (2003): one “actual and reasonable aim of science is to give us accurate descriptions (and other representations) of what reality is like. This project includes giving us accurate representations of aspects of reality that are unobservable." (2003, p. 176) Idealisation does not look like a strategy for delivering accurate representations of the kind the scientific realist demands. However, Weisberg (2007) draws our attention to the fact that idealisation is consistent with scientific realism. One type of idealised model Weisberg discusses is what is known as ‘Galilean idealisation’. He says of this kind of idealised model that it is the most straightforward type of idealised model compatible with scientific realism. Accordingly:

> “...the Galilean idealizer does aim to give complete, non-distorted, perfectly accurate representations. In order to accommodate the possibility of Galilean idealisation, scientific realists need to understand that achieving accurate representations of complex phenomena is an ongoing process. Even when the short term practice involves the willful introduction of distortion, the long-term aim can still be to give an accurate representation of what reality is really like. Thus scientific realism is perfectly compatible with Galilean idealisation, if the realist aim is understood to be long term or ultimate.” (2007, p. 657)

Klein (2018) has suggested in passing that the FEP is best thought of as Galilean idealisation. He describes the FEP as being:

> “...literally false, but with some understanding gained via over-simplification. Such idealizations can often be elaborated to be true of particular systems, and those elaborated models—which often look very different from the original model—can have considerable explanatory power. But if this is the case, then it is worth keeping in mind that FEP is a starting point from which one might develop explanations, and that its defence would ultimately rest on the empirical adequacy of detailed models which spring from it. Simplicity does not count in its favour, for FEP is simple in the way that friction-free planes and infinite populations of bunnies are simple: that is, a deliberate simplification, which buys scientific fruitfulness at the cost of literal truth.” (Klein 2018, pp. 2253-2254)

Idealisations may involve a distortion of something to be understood in order to gain a better understanding of it. A Galilean idealisation is the simplification or distortion of something into a model in order to turn a computationally intractable problem into one that is tractable (McMullin 1985). We pay special attention to this in section four, highlighting how variational free energy is introduced as an upper bound on surprise - turning a computationally intractable problem into a tractable one.

Is Klein right to say of Galilean idealisations, in general, and with respect to the FEP, that such idealisations are literally false? He is only partly right. Specifically, it does not follow that if a model is a Galilean idealisation it is *literally false*, in the sense of being *completely* false. Galileo, in his work on projectile motion or the diurnal rotation of the Earth made use of models with *mixed claims*: some entities should be taken as literally true (e.g., projectiles, the Earth, and so on), while other entities are more abstract and fictional (e.g., frictionless planes). Galilean models idealise by positing abstract entities to solve particular problems. The FEP does the same. However, as with Galileo’s models, the FEP is not literally false. FEP models idealise but they are not exhaustively idealisations in the sense of being complete distortions of target systems. We return to this central point in sections four, five and six.

Klein goes on to claim that Galilean idealisations ‘can often be elaborated to be true of particular systems, and those elaborated models—which often look very different from the original model—can have considerable explanatory power.’ This strikes us as correct. Indeed, this is what an *implementation* of the FEP aspires to with active inference. Klein’s final point is well-taken; namely, the “FEP is simple in the way that friction-free planes and infinite populations of bunnies are simple: that is, a deliberate simplification, which buys scientific fruitfulness at the cost of literal truth” (Klein 2018, p. 2254). However, we shall argue that literalism is not required for scientific realism to be true of the FEP. Techniques like idealisation and approximation are perfectly fine bedfellows for the scientific realist.

A different reason for holding Galilean idealisation to be consistent with scientific realism is that Galilean idealisers aspire, in the long run, to provide a more accurate description of target systems. There is no reason to think this is not the case for researchers in the FEP literature. The aspiration to construct realistic models must be appreciated to be an ultimate or long-term goal of a theory. The FEP need not be truth-producing just yet, or it need not be completely truth-producing yet. Here it is worth mentioning a common qualification about scientific realism; namely, that scientific theories such as the FEP need not be literally true to admit of a realist interpretation.

A slightly different concern is that model building *simpliciter* implies instrumentalism. It would suggest that model building is ultimately a matter of convention; not discovery. Yet, following Williamson, we think:

> “that would be a very naïve conclusion to draw … If we are investigating a complex reality out there, it is not at all surprising that it is sometimes best to use a sophisticated, indirect strategy, to ask questions quite subtly related to the overall aims of the inquiry. To build a model is just to identify by description a hypothetical example which we intend to learn about in hope of thereby learning about the more general subject matter it exemplifies. Nothing in that strategy is incompatible with a full-bloodedly realist nature for the scientific inquiry.” (Williamson 2017, pp. 3-4)

So, in summary, scientific models - in order to be compatible with scientific realism - need not necessarily be veridical models; nor do they always need to truthfully describe their targets. Even if a scientific model deliberately introduces distortions into its model of some target system, scientific realism need not be ruled out. Indeed, even if a scientific model adds into its structure fictional elements for instrumental gain, this still does not get scientific realism of any sophistication off the table (Godfrey-Smith 2009).

We now turn to assess each of the arguments for an instrumentalist interpretation of the FEP in detail.

4. Argument One: Variational Free Energy

It is reasonably straightforward to think of scientific models as abstract entities. The centre of mass of the solar system, a frictionless plane, a path of least action are all examples. So is *variational free energy*. Variational free energy is an information-theoretic construction used to *compute* how organisms are able to resist decay by positing variational free energy as an upper bound on entropy. Under the FEP, for an organism to exist it must keep its states (e.g., homeostatic states) within certain bounds. Or, it must maintain a low conditional entropy over its internal states (Corcoran et al. 2021). The FEP invokes an information-theoretic term to explain how this is achieved: *surprise*. This term allows a quantification of the *improbability* of some outcome (e.g., some sensory data). An important aspect of surprise is that it is conditional on the phenotype of an organism. A standard example is that the quantity of surprise goes up and down relative to whether a fish samples a sensory state on land or in its natural aquatic milieu. The FEP states that by minimising the surprise associated with particular sensory states, an organism is able to keep the entropy of its states low (and effectively survive), since entropy converges with long-term surprise.

Here is the important point for instrumentalists about the FEP (e.g. Friston 2019; van Es 2020; Ramstead et al. 2019, 2020). Surprise is widely recognised as being *computationally intractable*. The main reason for this has to do with having to summarise over the joint probability distribution involving how external states cause sensory states. Computing the surprise associated with sensory observations would require complete knowledge of the external dynamics resulting in sensory input (Friston 2010). Such a computation is intractable. This is why variational free energy is relevant. It is associated with a proxy for the quantity of surprise elicited by sensory data, and defined as a *functional*: it is a function of the function of sensory and internal states. It can be employed to reduce surprise precisely because it is defined as always being equal to or greater than surprise (Friston 2019). Variational free energy is thus an *abstract entity* enlisted into the FEP to provide a mathematical description of how organisms are able to maintain a low entropic distribution over constituent states.

Friston says that the FEP is “a mathematical formulation of how adaptive systems (that is, biological agents, like animals or brains) resist a natural tendency to decay,” (Friston 2010, p. 1). Crucially, according to Ramstead et al. (2020), “this means that internal and active states will look as if they are trying to minimise the same quantity; namely, the surprisal of states that constitute the thing, particle, or creature.” (2020, p. 21) Or, as Ramstead et al. (2019) also put it, “any system that avoids surprising exchanges with the world (i.e., surprising sensory states) will look as if it is predicting, tracking, and minimising a quantity called variational free energy, on average and over time.” (2019, p. 320) This suggests that what is implied is that “the system does not actually predict, infer, track or minimize a quantity called variational free energy, but it merely looks *as if*. The probabilistic model merely tracks certain real statistical relations in the organism-environment system.” (van Es 2020, p. 321) It is this ‘*as if*’ formulation of the FEP that leads Friston (2019), Ramstead et al. (2019, 2020), van Es (2020), and others, to conclude that systems do not literally minimise variational free energy. Since systems do not literally reduce this quantity, they conclude that scientific realism about the FEP is false.

Variational free energy is an abstract entity in FEP models because it is not literally true that target systems minimise this quantity. Nevertheless, it raises some quite interesting modelling issues and questions. For example, it can be used to “quantify and simulate self-evidencing [i.e., maximisation of model evidence].” (Friston 2019,

p. 85) It prompts the question of whether “self-organisation approximate Bayesian inference – or does Bayesian inference approximate self-organisation? (Friston 2019, p. 85) These are modelling questions that may lead to new discoveries about target systems. We have more to say about this in section six. Here we want to focus on the following: that the status of variational free energy as an abstract entity *has lent false credibility* to instrumentalist claims about the FEP.

The first reason is a confusion concerning scientific realism about models. This is especially evident in the following quotes from Ramstead et al. (2020). They claim to be advancing an instrumentalist account of the FEP *and* that their account “is … coherent with, but rests on distinct assumptions from, the realist position.” (2020, p. 1) They also describe their position as a kind of “nuanced realism” (2020, p. 5). This is confusing, since a commitment to such a tandem of positions is incoherent. Finally (as we highlighted in the introduction), they say of instrumentalists accounts in the philosophy of science that they “are useful fictions: they are not literally true, but ”true enough,” or good enough to make useful predictions about, and act upon, the world.” (2020, p. 4) Here is the crux of the issue: if taken literally, the FEP cannot be true. Alternatively, the FEP may not be understood literally. However, if the FEP (or scientific theories more generally) should not be understood literally, then scientific realism about the FEP (or scientific theories more generally) is false. Or, so it seems to those favouring instrumentalism about models.

This dilemma rests upon a misunderstanding of scientific realism. Instrumentalism about the FEP does not follow from the fact that FEP models do not provide descriptions of target systems that are literally true. Or, differently put, one cannot justify instrumentalism about the FEP because FEP models are idealised models - models that introduce distortions into models of target systems. In fact, scientific realists have always allowed that sophisticated models are highly partial and idealised, and yet that their predictive prowess constitutes prima facie grounds for concluding that they function to accurately describe aspects of target phenomena. The FEP comprises an indirect representation of a target system, even if it involves the use of abstract entities. It is an idealised approach to representing complex or unknown processes in the world. It is standard practice to view scientific models as *indirect representations* of real-world, target systems (Weisberg 2006). Furthermore, most theoretical models are composed of a set of *mixed claims* (Psillos 2011). This means that the model will posit, if true, the presence of “both OK-entities (such as electrons and their ilk) and supposedly non-OK-entities (such as numbers … or theoretical ideals.” (Psillos 2011, p. 6) The FEP is such a model. It puts forward OK-entities (such as neurons, reflex arcs, hierarchical structures in the brain, and so on) and non-OK-entities (such as the variational free energy notation), where non-OK-entities must be understood in terms of not being literally true of target systems.

This raises an important question: what is the status of such *abstract entities* such as variational free energy. Some proponents of the instrumentalist reading of the FEP say that the claim that organisms minimise free energy is a *useful fiction* - it is an ‘as if’ story; not a literally true story. It is tempting to think that if a claim is not literally true of a target system, but is nevertheless explanatorily useful, then this claim must have the status of a fiction. It is the *fictional status* of variational free energy that seems to motivate instrumentalism about the FEP.

Taking variational free energy to be a useful fiction is not incompatible with a scientific realist interpretation of the FEP. We noted above that implicating abstract entities in model building is part and parcel of the process by which descriptions of models are constructed - infinitely large populations in biology, ideal gases, mass-points, state space, attractor points, perfectly isolated systems, and so on, are examples. Variational free energy falls within the same class of abstract entities.

Godfrey-Smith (2009) notes three important properties of abstract entities in science. First, *hypothetically*, were such entities to exist, they would be entities located in space-time. They would allow for actual, not hypothetical, interventions. Second, abstract entities have *investigative* properties; they are the common property of a community of researchers - e.g., the FEP community, which inherits many of its notations from other areas such as statistical physics and machine learning. This means they can be “investigated collaboratively [and] surprising properties might be uncovered by one investigator after being denied by another.” (2009, p. 102) Finally, given their hypothetical *and* investigative properties, “their status, though not their role, [...] seem analogous to the fictions of literature.” (2009, p. 102) The crucial point here is that by means of theorising with abstract entities, one may learn about the world - both with respect to observable entities and unobservable structures. This is enough for scientific realism, and it is sufficient vis-a-vis the FEP: it posits abstract entities in its modelling of target systems in order to learn about such systems.

We end this section with a concern. It might be objected that works of fiction do not tell you anything about the world. So fictions cannot do the needed work to ground scientific realism - not even in the case of the FEP. Here, again, the work on the relation between models and fictions in science by Godfrey-Smith (2009) is constructive. Briefly, since we will return to this issue in section five and six, one might say that certain parts of a theoretical model are fictional or abstract entities (e.g., variational free energy), and that these entities have different *similarity* relations to target systems. Specifically, biological systems that are able to minimise surprise will *appear* to minimise their variational free energy. Here is a case of similarity between the model and the target system. Godfrey-Smith (2009) puts the general view as follows: “model systems are fictional things which have various similarity relations to real-world systems. We learn about the former, and use that knowledge to illuminate and adapt us to the latter.” (2009, p. 108) In the very same way, investigations of how variational free energy is minimised in a model allows for insight into how surprise is minimised in target systems. Abstract entities can be seen as laying a path towards understanding the dynamics of real-world, target systems. It is this feature that allows fictional entities in model construction to tell you something about the world of target systems. This is all one need for the realist interpretation of FEP models.

5. Argument Two: Markov Blankets

We now consider work by Bruineberg et al. (2021) on the Markov blanket formalism underwriting the FEP. They argue that the use of the Markov blanket formalism in the literature on the FEP often takes the formalism (the map) to literally be a property of the territory.

The Markov blanket concept originates with Pearl (1988) in his work on probabilistic reasoning and Bayesian networks. In probabilistic networks, Markov blankets are used to model probabilistic relations between nodes or variables: e.g., nodes A and B can be modelled as conditionally independent from one another in virtue of a third node, C. In this case, C can be said to ‘shield’ or ‘separate’ A from B, and vice-versa. According to Beal (2003), the “Markov blanket for the node (or set of nodes) A is defined as the smallest set of nodes C, such that A is conditionally independent of all other variables not in C, given C.” (2003, p. 18) The key point here is that once a Markov blanket has been identified for any given node, e.g., A, this captures all the relevant information needed to *infer* the state of A. Markov blankets can be used in order to model (in)dependencies between different variables, which allows for an approach to probabilistic reasoning under uncertain circumstances. Bruineberg et al. (2021) call Markov blankets of this kind *Pearl blankets*.

Under the FEP, this formal model of identifying conditional independence between nodes in a network is *interpreted* in a specific way. That is, once a Markov blanket has been identified for a living system, the blanket states are modelled as active and sensory states, separating internal (organismic) states from external (environmental) states. Bruineberg et al. (2021) call Markov blankets of this kind *Friston blankets*. Unlike Pearl blankets, Friston blankets work to describe a real-world boundary demarcating an organism from its environment, and vice-versa. According to Bruineberg et al. (2021), the use of Friston blankets to draw a distinction between organism and environment is not licensed by appeal to the mathematical structure of Pearl blanket. Hence, Bruineberg et al. (2021) finish their article by providing FEP researchers with a choice:

> “the considerations presented in this paper leaves the FEP theorist with a choice. One can accept a rather technical and innocent conception of Markov blankets as an auxiliary formal concept that define[s] what nodes are relevant for variational inference [= Pearl blankets]. This conception is admittedly

> scientifically useful but has not yet led to any philosophically interesting conclusions about the nature of life or cognition. Alternatively, one can import a number of stronger metaphysical assumptions about the mathematical structure of reality to support a realist reading, where the blanket becomes a literal boundary between agents and their environment [= Friston blankets]. Such a strong realist reading cannot be justified by just ‘doing the maths’, but rather needs to be independently argued for, and no such argument has yet been offered.” (2021, p. 53)

The problem with this kind of argument should now be very clear: it fails to identify that the theoretical landscape contains more than two options. FEP researchers working with the Markov blanket formalism have a third choice; namely, they can legitimately employ an *abstract entity* (the Markov blanket formalism) to *indirectly represent* target systems in order to describe boundaries between states. Note that accepting some important, or even fundamental, link between Markov blankets and organism-world boundaries does not entail accepting the further literalist claim: that the formal and mathematical structure of a Markov blanket (the map) literally is the boundary (the territory). Those working with *Friston blankets* need not accept the literalist view that Bruineberg et al. (2021) attribute to them. This was the point we made above that scientific models are indirect representations of target systems.

Bruineberg et al. (2021) make a good deal out of the fact that the shift from Pearl blankets to Friston blankets requires *an interpretative step*. It requires interpreting Markov blanket states in terms of sensory and active states. But, any kind of interpretation is done by a theorist, rendering the identification of Markov blanket states arbitrary. This is a fair point. However, science is unlikely to progress in the absence of interpretation. First, any application of a model to a target system requires interpretation of the model for it to be a model of the target system. This is the case for any kind of model, whether it be concrete, mathematical, computational or some other kind of model. Recall our example of the scale model of a bridge at the start of the paper: even a scale model of a bridge must be interpreted as a model of a bridge for it to be meaningful for those using it. As a matter of fact, interpretation is also part of working with Pearl blankets. Importantly, interpretation does not imply arbitrariness. Interpretation does not rule out scientific realism about the use of the Markov blanket formalism in the FEP.

A useful reminder of the necessity of interpretation in science, can be found in a classical discussion on the role of *hypotheses* in science by the 18th Century scholar, Emilie du Châtelet. In her *Foundations of Physics*, she says the following of hypotheses:

> “There must be a beginning in all research, and this beginning must almost always be a very imperfect, often unsuccessful attempt. There are unknown truths just as there are unknown countries to which one can only find the good

> route after having tried all the others. Thus, some must run the risk of losing their way in order to mark the good path for others; so it would be doing the sciences great injury, infinitely delaying their progress, to banish hypotheses as some modern philosophers have." (1740, p. 147)

For the scientist to construct a hypothesis they must ask questions, led by their explanatory interests. Model building involves interpreting structures in the world in a meaningful way. If this is on the right track, then science without any kind of interpretation, without any kind of explanatory or interest on behalf of researchers is a false idealisation of the scientific method. Moreover, there is no need to think, or so we submit, that these features of the scientific method create any problem for scientific realism.

Finally, Bruineberg et al. 2021 argue against a scientific realist interpretation of the Markov blanket formalism on the grounds that the FEP applies to *any* open dynamical system. Therefore, they contend, the claim that the FEP unifies biology and cognition is problematic. This is true (see e.g., Kirchhoff 2018; Kirchhoff et al. 2018). Yet, explanatory scope issues are mute with respect to scientific realism. Or, differently put, one cannot derive a claim about antirealism by appealing to the explanatory reach of the FEP. One can only establish that the FEP applies to more systems than biological systems.

We end this tour into the world of blankets with a further reflection and qualification on *literalism*. A scientific realist might admit that in certain circumstances, there will be local reasons to posit abstract entities and give them a literal interpretation. But, in such cases the scientific realist will not be providing a literal interpretation of these abstract entities across the map-territory relation. Consider what Psillos (2011) says of the Carnot engine:[4] "The model of a (fully reversible) Carnot engine is such that it cannot represent exactly and accurately any worldly engine. This was known to Sadi Carnot himself—as well as to anybody else—and this knowledge might be enough to justify taking the Carnot engine as a fiction." (2011, p. 6)

It is obvious that the Carnot model is a model with *mixed claims*. We addressed this above, and will say no more here. Yet, it is worth highlighting that the scientific realist might take literally the theoretical description of the (fictional) Carnot engine. In discussions about the FEP, the scientific realist might take literally the theoretical description of the Markov blanket formalism, without endorsing the additional claim that the formalism itself literally is the boundary being modelled.

This brings up an important issue to which we shall now turn: *approximation*. As we understand it, the Markov blanket formalism itself is not literally true of target

[4] A Carnot engine is a theoretical thermodynamic cycle proposed by Leonard Carnot that estimates the maximum possible efficacy that a heat engine can turn into work.

systems. It is an approximate way of representing the boundaries of target systems. The scientific realist may endorse the claim that the theoretical description is literally true, but only approximately true of the target system. This leaves it open to providing further details about the model and its relation to target systems. We now consider this issue with respect to Bayesian inference and generative models in the FEP.

6. Argument Three: Approximate Bayesian Inference

It is common to read that instrumentalism about the FEP is motivated by appeal to free energy minimisation being an approximation of Bayesian inference (Ramstead et al. 2020; van Es & Hipolito 2020; van Es 2020).

In section four, we considered the status and role of variational free energy in the FEP. The relevant issue about literalism versus approximation arises once questions are asked about how organisms minimise free energy. The usual proposal is that free energy minimisation can be shown to be equivalent (under certain mathematical assumptions) to Bayesian inference given a generative model. A generative model is a *probabilistic* specification of how some kind of data might have been generated. According to Parr & Friston, a generative model “expresses prior beliefs about unobserved hidden states [i.e., causes of sensory input], the probabilistic dependencies between these states, and a likelihood function that maps hidden states … to sensory data …” (2018, p. 2) A generative model can be used to predict new data. It can be used to infer the hidden states that may have caused the observed data (Beal 2003). This allows a Bayesian rendition of the FEP; namely, that internal, sensory and active states of an organism can be understood as engaging in optimising their posterior beliefs over a (generative) model as new evidence (sensory input) is being generated.

Here the instrumentalist can make their case. Very few think that systems such as brains *literally* compute Bayes’ rule - the set of computational steps used to update beliefs given new evidence under that particular framework. Insofar as brains *do* such things as Bayesian inference, they do Bayesian inference *approximately*. To suggest otherwise would be a clear case of conflating the mathematics with the territory (cf. Andrews 2021). Hence, the FEP, even when given a Bayesian articulation, is not the claim that organisms perform optimal or literal Bayesian inference to minimise surprise. The FEP is the claim that the minimisation of variational free energy conforms to an *approximation of Bayesian inference*. It is the nervous system that can be said to engage in approximate Bayesian inference, given that neuronal dynamics seek to anticipate sensory input (or, seek to maximise the likelihood of sensory observations). The top-down anticipatory cascades are a good approximation of the true posterior so long as surprise is kept to a minimum.

Norton (2012) defines ‘approximation’ as inexact descriptions of a target system (2012, p. 207). Since the FEP as an approximation is inexact, it is not literally true of its target systems. The status of the FEP as an approximation may therefore be taken to lend support to instrumentalism about the FEP. Such an inference however again relies upon the literalist fallacy, since nothing about approximation shows that scientific realism is false.

A small proviso to the above paragraph. One must be cautious to distinguish the notion that FEP models feature approximations of various kinds from the distinct notion that the respective target systems themselves approximate a form of (variational) Bayesian inference. One might hold that FEP models really describe the organism as approximating Bayesian inference, and that the organism does indeed literally approximate Bayesian inference but not in the sense of how a computer-based model would do this (the organism or its central nervous system presumably does not allocate numbers across a probability space). This is very much how Hohwy approaches the issue here:

> “The key point is that approximate inference is not just like exact inference except with approximate values, rather the ‘approximation’ is the minimization of the KL-divergence [the difference between posterior and recognition densities that is minimised during the minimisation of variational free energy]: an approximation of the states of the system to the states that it would have if it were indeed computing exact inference. This means that it is correct to say that such a system is literally doing ‘approximate inference’ even if it does not literally do ‘inference’ in the sense of ‘computations over probability distributions’.” (2020, p. 17)

Hohwy’s point is that it can be true to claim that brains approximate Bayesian inference even if brains do not literally compute belief updating schemes such as Bayes’ rule. If the claim that brains engage in approximate Bayesian inference is true, then it is possible to say that there is a *similarity* between models in cognitive neuroscience and their target systems. Work in cognitive neuroscience is increasingly done with approximate models of Bayesian inference. The similarity between Bayesian models and the brain allow cognitive neuroscientists to gain further insights into the functional and structural organisation of the brain. This is consistent with scientific realism. As Giere (1988) puts it (in a different discussion):

> “One way science advances is by discovering new aspects of the world, that is, new respects in which models might resemble the world. Science also advances by discovering some respects in which similarities between model and world are not as commonly thought. Neither sort of advance, however, is inconsistent with … realism.” (1988, p. 107)

Of course, logically speaking, anything is similar to something else in some way, or another. This means “the claim of similarity must be limited (as least implicitly) to a specified set of respects and degrees.” (Giere 1988, p. 93) Under the FEP, the similarity between a scientist's generative model and the dynamics of the brain turns on the idea that the neuronal dynamics *conforms* to the scientists model. What this means is that if both neuronal dynamics and the scientific model succeed in reducing surprise, they both converge on the following: the ability to keep a low conditional entropy over their constituent states. More specifically, in a recent review of the field, Isomura (2021) concludes that progress in theoretical neurobiology demonstrates that standard neural networks perform variational Bayesian inference under the form of a generative model. This is another way of conceptualising the issue of similarity between models and target systems, since it “demonstrate[s] that standard neural networks - comprising biologically plausible neural activity and plasticity models - can perform … inference.” (Isomura 2021, p. 1) Note that this connects nicely with Hohwy’s (2020) claim that “it is correct to say that such a system is literally doing ‘approximate inference’ even if it does not literally do ‘inference’ in the sense of ‘computations over probability distributions’.” (2020, p. 17)

A related question to ask is: what does working with generative models enable scientists to do? Note that we are here shifting perspective from the target system itself to the theoretical model used by the scientist. According to Turner & Zandt (2018), new Bayesian techniques using approximate methods are making it “possible to fit these … models to data. These techniques have even allowed simulation-based models to transition into neuroscience, where tests of cognitive theories can be biologically substantiated.” (2018, p. 1) In the FEP literature, this sort of work is starting to emerge. Generative modelling under the FEP is being used to explore how biological systems such as the brain are able to anticipate sensory observations. For example, Parr et al. (2019) investigate generative models under active vision. Of the significance of this work, they say:

> “In brief, we used magnetoencephalography in combination with eye-tracking to assess the neural correlates of a form of short-term memory during a dot cancellation task. Using dynamic causal modelling to quantify changes in effective connectivity, we found evidence that the coupling between the dorsal and ventral attention networks changed during the saccadic interrogation of a simple visual scene. Intuitively, this is consistent with the idea that these neuronal connections may encode beliefs about “what I would see if I looked there”, and that this mapping is optimized as new data are obtained with each fixation.” (2019, p. 1)

We can see from this example how cognitive neuroscientists are using the FEP to construct generative models that they intend to be at least indirect representations of systems in the real world. The *indirectness* of models is a result of the *complexity* of target systems. Generative-model-based simulations allow for reduction of the extreme complexity in systems being modelled.

We have focused on the relation here between generative models as they are constructed in computational neuroscience and neural systems. Scientific realism is sometimes taken to imply that organisms encode or implement such a model, enabling them to navigate a dynamic environment. van Es (2020) argues against such a realist claim first by noting correctly that it is the “adaptive behaviour of the system that implements or instantiates a generative model”. It is the organism’s adaptive behaviour that “brings forth the conditional dependencies captured by the generative model” (van Es 2020, p. 7; both quotes). van Es (2020) goes on to argue that it is adaptive behaviour that does the work of minimising free energy, and the generative model is “merely a scientific construct that captures real statistical relations in the world.” (2020, p. 318) We agree. However, we take these points to illustrate the similarity we have argued for above between the properties of the generative model and adaptive behaviour, such that one can capture *real statistical relations in the world* by means of generative modelling. It is this crucial point about the relation of relevant similarity between the generative model of the scientist and the dynamics of the adaptive behaviour of organisms that is missed by van Es and others defending instrumentalism about the FEP.

7. Argument Four: Ergodicity

We now turn to consider the final argument for instrumentalism about the FEP. This argument problematises a central commitment of the FEP: that organisms can be modelled as random dynamical systems with *bounded attracting states*, meaning that one can model systems that seek to maintain non-equilibrium steady states as having an *ergodic density*.

Random dynamical systems are systems whose states are subject to random fluctuations. This means that if one models a system as an *ergodic* system, the states of such a system will, after some amount of time, converge to what is called a random global attractor (i.e., a set of invariant states). Ergodicity refers to the “time average of any measurable function of the system converges (almost surely) over a sufficient amount of time. This means that one can interpret the average amount of time a state is occupied as the probability of the system being in that state when observed at random.” (Friston 2013, p. 2) Given this assumption, one can model the proportion of time a system spends in any region of its phase space as equivalent to the probability of such a system being in this region of its phase space. An uncontroversial example is tossing a fair coin. With enough time spent tossing a fair coin, the probability of it landing on heads is equivalent to the time spent flipping the coin. Ergodicity concerning non-biological cases, such as coin tosses, is one thing. It is quite a different matter to assume ergodicity about biological systems. Colombo &

Palacios (2021) argue that “while in physics one can ‘pre-state the phase space for target systems based on stable invariances and symmetries, historical processes studied principally in evolutionary and population biology…involve symmetry breaking, which makes phase spaces structurally unstable, ever-changing and unpredictable” (2021, p. 13). They argue on these grounds that the systems studied in biology are non-ergodic, calling into question the possibility of modelling biological systems in terms of random dynamical attractors. The FEP is taken to apply both to physical systems that can be modelled in these terms, and to biological systems that Colombo & Palacios (2021) argue do not admit of such modelling. Thus Colombo & Palacios (2021) argue that the FEP achieves its generality at the expense of biological realism about its models.

Colombo & Palacios (2021) frame their discussion of trade-offs in terms of Levins’ (1966) framework in biology. Levins (1966) uses the term ‘realism’ in a particular way - in a way that does not, or need not, converge with scientific realism. According to Weisberg (2006): “In some passages, he uses the term ‘realism’ as a synonym for accuracy. In others, it is related to diverse considerations including the number of factors included in the model …” (2006, p. 635) With this in mind, we think that even if the FEP trades off realism (in the sense of Levins) for generality and mathematical precision, there is nothing about this that makes it problematic to endorse scientific realism about the FEP. More accurate representations of target systems can be gained by working to show how causal properties of target systems are reflected in FEP models (for an overview of implementation details of the FEP, see Da Costa 2020).

We suggest that ergodicity is best understood as a way of *interpreting* the behaviour of systems whose dynamics are subject to random environmental fluctuations. If one models a biological system as an ergodic system, it is possible to show that long-term surprise is entropy. This implies that minimising free energy results in keeping a low entropic distribution over internal states in ways that are interestingly similar to biological systems whose dynamics are anticipatory. Furthermore, it is now fairly common to weaken the notion of ergodicity. We have seen above how the FEP requires that populations *on average* tend to frequent the same kinds of phenotypic states. It is reasonably intuitive to think that biological systems tend to return to similar phenotypic states over time and that their adaptive behaviours depend upon them doing so. Body temperature is a good example, fluctuating around an average of 36.5C in humans. Ergodicity can in this weak sense be said to represent biological systems as an approximation of their true adaptive behaviours. We have seen above how the role of approximation in scientific modelling is no threat to understanding scientific models in realist terms.

Conclusion

This paper has focused on the status of scientific realism concerning the FEP. In this context, many have found the view that FEP models are idealised and approximate models of target systems a compelling reason to defend instrumentalism about the FEP. We have argued that all the reasons provided (so far in the literature) for this instrumentalist rendition of the FEP are fallacious. They all commit, in one way or another, what we termed the *literalist fallacy*. This is the fallacy of inferring the truth of instrumentalism based on the claim that the properties of FEP models do not literally map onto real-world, target systems. In the end, we hope to have shown that a realist interpretation of the FEP is a live and tenable option. We take this to imply that the following assertions are live ones: (1) FEP models describe (partly, at least) a mind- independent structure (e.g., discoveries of how self-organisation in open systems is achieved); (2) the FEP highlights a semantic perspective on models, seeing models as truth-conditioned descriptions of target systems - they can be true or false; and (3) the FEP exhibits an epistemic perspective on models in terms of idealisation, approximation and abstract, indirect representation. In the end, what makes models based on the FEP realistic models of biological systems is that the dynamics of biological systems that behave adaptively, returning to a state of dynamic equilibrium when randomly perturbed, are interestingly similar to the models scientists make of such systems using the mathematics of the FEP.

References

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

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

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

Bruineberg, J., Dolega, K., Dewhurst, J., and Baltieri, M. (2021). The emperor's new Markov blankets. *Behavioral and Brain Sciences*. Forthcoming.

Châtelet, E. (1740). *Foundations of Physics*. https://historyofwomenphilosophers.org/project/du-chatelets-foundations-of-physics/

Colombo, M., and Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle. *Biology & Philosophy*, 36:41 https://doi.org/10.1007/s 10539-021-09818-x 13

Constant, A. (2021). The free energy principle: it's not about what it takes, it's about what took you there. *Biology & Philosophy*, 36, 1-17.

Corcoran, A., Pezzulo, G., and Hohwy, J. (2021). From allostatic agents to counterfactual cognisers: active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, 35, 1-45.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., and Friston, K. (2020). Active inference on discrete state spaces: a synthesis. arXiv preprint. arXiv:2001.07203

Friston, K. (2019). A free energy principle for a particular physics. Unpublished manuscript.

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

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

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

Giere, R.N. (1988). *Explaining Science: A Cognitive Approach*. Chicago: The University of Chicago Press.

Godfrey-Smith, P. (2009). Models and fiction in science. *Philosophical Studies*, 143, 101-116.

Godfrey-Smith, P. (2003). *Theory and Reality*: *An Introduction to the Philosophy of Science*. Chicago: The University of Chicago Press.

Hesp, C., Ramstead,M., Constant, A., Badcock,P., Kirchhoff, M.D., and Friston, K. (2019). A Multi-scale view of the emergent complexity of life: A free energy proposal. In M. Price et al. (eds), *Evolution, Development, and Complexity: Multiscale Models in Complex Adaptive Systems* (pp. 195-227). Springer

Hohwy, J. (2020). Self-supervision, normativity and the free energy principle. *Synthese*, 199, 29-53.

Isomura, T. (2021). Active inference leads to Bayesian neurophysiology. *Neuroscience Research*, https://doi.org/10.1016/j.neures.2021.12.003

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

Kirchhoff, M. D. (2018). Hierarchical Markov blankets and adaptive active inference. *Physics of Life Reviews*, 1-2, doi: https://doi.org/10.1016/j.plrev.2017.09.001.

Kirchhoff, M. D. (2015). Species of realization and the free energy principle. *Australasian Journal of Philosophy* 93(4), 706-723.

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

Laudan, L. (1984). *Science and Values*. Berkeley: University of California Press.

Levins, R. (1966). The strategy of model building in population biology. In: Sober E (ed) Conceptual issues in evolutionary biology, 1st ed. Cambridge MA, MIT Press, pp 18–27

Mann, S., Pain, R., and Kirchhoff, M.D. (2021). Free energy: A user's guide. Pre-print: http://philsci-archive.pitt.edu/19961/

McMullin, E. (1985). Galilean idealisation. *Studies in the History and Philosophy of Science*, 16(3), 247-273.

Norton, J. (2012). Approximation and idealisation: Why the difference matters. *Philosophy of Science*, 79(2), 207-232.

Palacios, E. R., Razi, A., Parr, T., Kirchhoff, M., and Friston, K. (2020). On Markov blankets and hierarchical self-organisation. *Journal of theoretical biology*, 486, 110089.

Parr, T., Mirza, B., Cagnan, H., and Friston, K. (2019). Dynamic causal modelling of active inference. *Journal of Neuroscience*, 39(32), 6265-6275

Parr, T., and Friston, K. (2017). The anatomy of inference: Generative models and brain structure. *Frontiers in Computational Neuroscience*, https://doi.org/10.3389/fncom.2018.00090

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Psillos, S. (2011). Living with the abstract: realism and models. *Synthese*, 180, 3-17.

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. London: Routledge.

Ramstead, M., Friston, K., and Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, 22, 889; doi:10.3390/e22080889

Ramstead, M., Badcock, P., and Friston, K. (2019). Variational neuroethology: Answering further questions: Reply to comments on 'Answering Schrödinger's question: A free-energy formulation'. *Physics of Life Reviews*, 24, 59–66.

Stanford, K. (2003). Pyrrhic victories for scientific realism. *The Journal of Philosophy*, 100(11), 553-572

Turner, B., and Zandt, T. (2018). Approximating Bayesian inference through model simulation. *Trends in Cognitive Sciences*, 22(9), 1-15.

Tschantz, A., Seth, A.K., and Buckley, C.L. 2020. Learning action-oriented models through active inference. *PLoS Computational Biology*, https://doi.org/10.1371/journal.pcbi.1007805

van Es, T. (2021). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*, 29(3) 315–329

van Es, T., and Hipólito, I. (2020). Free energy principle, computationalism and realism: a tragedy. http://philsci-archive.pitt.edu/18497/.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, 104(102), 639-659.

Weisberg, M. (2006). Forty Years of 'The Strategy': Levins on Model Building and Idealization. *Biology & Philosophy*, 21, 623–645