

# Living models or life modelled? On the use of models in the free energy principle

Thomas van Es 

Adaptive Behavior

1–15

© The Author(s) 2020

Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/1059712320918678

[journals.sagepub.com/home/adb](https://journals.sagepub.com/home/adb)



## Abstract

The free energy principle (FEP) is an information-theoretic approach to living systems. FEP characterizes life by living systems' resistance to the second law of thermodynamics: living systems do not randomly visit the possible states, but actively work to remain within a set of viable states. In FEP, this is modelled mathematically. Yet, the status of these models is typically unclear: are these models employed by organisms or strictly scientific tools of understanding? In this article, I argue for an instrumentalist take on models in FEP. I shall argue that models used as instruments for knowledge by scientists and models as implemented by organisms to navigate the world are being conflated, which leads to erroneous conclusions. I further argue that a realist position is unwarranted. First, it overgenerates models and thus trivializes the notion of modelling. Second, even when the mathematical mechanisms described by FEP are implemented in an organism, they do not constitute a model. They are covariational, not representational in nature, and precede the social practices that have shaped our scientific modelling practice. I finally argue that the above arguments do not affect the instrumentalist position. An instrumentalist approach can further add to conceptual clarity in the FEP literature.

## Keywords

Free energy, models, realism, instrumentalism, covariation, anti-representationalism

Handling Editor: Michael Kirchhoff, University of Wollongong, Australia

## 1. Introduction

In recent years, an increasingly popular view on life and cognition has taken shape, which is based on the free energy principle (FEP). FEP, in this context, started as a mathematical tool to understand the workings of the brain (Friston, 2002, 2003, 2005, 2011). Yet, since then, it has been applied to a wide variety of biological, psychological and social phenomena (Friston, 2013; Hesp et al., 2019; Ramstead et al., 2016, 2018). FEP is proposed to be a 'research heuristic' by which systems under investigation are modelled as free energy minimization systems (Ramstead, Badcock, & Friston, 2019). In some formulations, the system is also thought to implement this model either internally or in an embodied (or extended) sense so as to make statistical inferences about the external world (Bruineberg et al., 2016; Clark, 2016; Friston, 2012; Hohwy, 2013; Kirchhoff & Kiverstein, 2019). These different uses of FEP can be related to a difference between models *of* a target system made by scientists, and models *used by or embodied*

*by* a target system, the latter of which is further divided into neurocentric and embodied/extended approaches.<sup>1</sup>

In this article, I argue that we ought to disambiguate these two uses of modelling. Furthermore, I argue that only the former, instrumentalist understanding of free energy minimization in biology is justified, whereas the realist view that models are biologically implemented in organisms is not. This does not change the scope of what we can investigate with FEP, but it does constrain the conclusions that can be drawn from free energy models.

I will first give a brief overview of the FEP, explain its core theoretical posits and briefly discuss some of its applications. The distinction between models as mathematical constructs from models as they are ascribed to

---

Centre for Philosophical Psychology, Universiteit Antwerpen, Antwerpen, Belgium

### Corresponding author:

Thomas van Es, Centre for Philosophical Psychology, Universiteit Antwerpen, 2000 Antwerpen, Belgium.  
Email: [thomas.vanes@uantwerpen.be](mailto:thomas.vanes@uantwerpen.be)

organisms and brains will stand central in my explanation. This deviates from the usual manner in which the FEP is introduced (in which, as I will argue here, the two are typically conflated), but it is required by the goals of this article. Once FEP is on the stage, I shall disentangle the two uses of model and show that they are conceptually distinct. After this, I will argue that we should eschew a realist interpretation of these models. The first argument is the charge of overgeneration and trivialization of models. That is, if we accept a realist position on models in FEP, many non-living systems will be cast as modellers, entailing a trivialization of the notion of modelling. The second argument relies on Kirchhoff and Robertson's (2018) analysis of FEP as essentially covariational, not representationally contentful. Because models, both in representationalist and non-representationalist accounts, require more than mere covariation relations, a realist position on models in FEP is off the table. Finally, within scientific practice, covariational relations can be exploited to create models in the proper sense of the word, vindicating an instrumentalist approach to models in FEP. Before concluding, I will discuss some possible objections to my proposal.

## 2. The FEP explained: brain and life

### 2.1. A model of life

FEP is a principle-first approach to living systems (Friston, 2012). According to the second law of thermodynamics, any closed system's entropy will increase over time. Consider, for example, a bottle of perfume in a closed room. As soon as you open the bottle, the perfume will slowly spread throughout the entire room: its entropy will increase. Living systems, however, put in a lot of effort to not disperse randomly and, relative to all states they possibly could visit, remain within the subset of viable states (Friston & Stephan, 2007). In part, this is because they are semi-open systems: they regulate their exchanges with the environment to maintain homeostasis. Variational free energy is formally an upper bound on entropy, and thus, living systems minimize their free energy to stay alive (Friston, 2012, 2013). A research goal of FEP is to unearth the minimal characteristics that are necessary for any system to maintain its boundary with the environment. This is done with the idea that free energy minimization is an imperative of life (Friston, 2012). Via mathematical modelling, FEP can bring to light the relations between living systems at multiple spatiotemporal scales (Clark, 2017; Kirchhoff et al., 2018).

If we suppose that every organism has a clearly defined set of states, *characteristic* states, in which the organism remains alive, there is a boundary of states within which the organism is alive, outside of which the organism is dead (Friston, 2012). A state, here,

corresponds to a particular value for a particular variable in a model. The characteristic states of a living system, then, are a collection of values for particular variables in which the organism is alive. The characteristic states differ greatly per phenotype: a fish stays alive in rather different circumstances than we humans (Bruineberg et al., 2016; Friston, 2012; Kirchhoff et al., 2018). For any system at any particular point in time, we could make a probability distribution in which probabilities are assigned for each possible collection of states the system could possibly be in. If we assume such a distribution to be applicable to an organism (which is debatable, see Colombo & Wright, 2018; Colombo et al., 2018; see also Longo et al., 2012), then there will be high probabilities for those states in which it is alive, low probabilities for fatal states. Conversely, viable states will have a low uncertainty, and fatal states will have a high uncertainty: they will be rather surprising. As such, the organism would do best to avoid those surprising states. Moreover, 'surprise' in the long run, as it is used here, is considered formally equivalent to entropy, which in turn means that variational free energy is an upper bound on surprise (Friston & Stephan, 2007). This means that, by minimizing free energy, the organism will indirectly also minimize surprise in the long run.

Uncertainty plays a central role in FEP. An organism's primary imperative is to stay alive: it is supposed to steer clear from surprising states, or in other words, it is supposed to minimize the uncertainty or surprise as it is calculated in the model. In the long run, when the organism remains within surprising states, the imperative is thus to resolve this situation. That is, to make it less surprising. To do so, the organism needs to act. FEP theorists understand action to occur in terms of *active inference*. Active inference assumes we have a probability distribution over possible action policies of the organism, which in a particular context translate to particular actions to be taken (Friston, 2012; Kirchhoff & Kiverstein, 2019). It is an *inference* because it is an inferred prediction about a future state of the organism, and it is *active* because this inference is *conditional* on endogenous, organism-created action. That is, the prediction can be brought about by action of the organism (Friston, 2012). Put differently, the movement that the prediction is conditional on, is the action the organism ought to undertake in order to make the predicted state reality, to remain in unsurprising (viable, high probability) states. This means that, given the state of the organism and a probability distribution of action policies, we can predict a new probability distribution of action policies that will show which actions best minimize surprise in the current situation. That is, what the organism is best off doing to stay alive.

An alternative to using active inference to minimize free energy is to simply *update the probability distribution* so that previously surprising states are not so

anymore. This is called *perceptual inference* (Friston, 2005; Friston & Kiebel, 2009; Hohwy, 2013). This will not help if the surprising states are directly fatal, but it may be helpful, say, if a human for the first time learns to swim. Being in the water is a highly surprising state, which, if applying solely active inference, would have to be steered clear of. Yet, if the human learns to swim, this particular state is viable, in which case the probability distribution of states of the organism needs to be updated.

Important technical terms surrounding FEP are the *Markov blanket* and the *generative model*. A *Markov blanket* is a statistical tool with which a system is divided between internal and external states with blanket states between them. Blanket states are the points of contact between internal and external states, and are comprised of sensory states through which external states influence internal states, and active states through which internal states influence external states. The Markov blanket itself, thus strictly seen, comprises only those blanket states between internal and external states. Furthermore, in this formalism, the internal and external states are statistically *conditionally independent* (Friston, 2013). That is to say that, given (conditional on) the blanket states, knowing more about internal states does not offer further insights into external states and vice versa. The sensory and active states are *conditionally dependent* on the external and internal states, respectively. This means that the sensory states are dependent on the external states, and the active states are dependent on the internal states. In the context of the organism, the internal states map onto the biophysical organism, the external states map onto its environment, and the active and sensory states map onto action and perception, respectively.

Intuitively, the environment, the external states, have a much wider range of variability than the internal states. This is because an organism can be in the ocean, on top of a volcano, deep underground in a cave or in the middle of an urban city, yet the organism itself, the internal states, remain semi-stable. This means that inferring the external states directly is computationally intractable (Friston, 2010). Yet, due to the conditional dependencies, one only needs to know the internal states and sensory states to be able to calculate an *approximate* prediction concerning the external states. For active inference, such an approximate inference conditioned on action will suggest a pathway that minimizes free energy or surprise. Note that when the organism moves in accordance with active inference, this constitutes an influence on the external states via active states, by the internal states. For perceptual inference, the same knowledge is needed, yet here the direction of influence is the other way around: the external states impose a particular influence on the internal states, *via* the sensory states.

*Generative model* is the term used for the probability distribution of the internal states (Friston, 2012). In the context of an organism, these are the states of the organism itself. It is these states that are essential to have access to in applying both active and perceptual inference. As we apply both active and perceptual inference in order to keep the modelled organism in viable states, minimizing its free energy, the generative model will recapitulate, *adapt to*, or covary with the structure of the external states, also called the *generative process* (Bruineberg et al., 2018; Kirchhoff & Kiverstein, 2019; Ramstead, Kirchhoff, & Friston, 2019). This relation with the generative process unfolds on ontogenetic and phylogenetic timescales. Consider how a fish's phenotype (the gills, a scaled body) shows covariation with the water it lives in, and how organisms during their lifetimes familiarize themselves with and adapt to new environments. This means that, from an outsider's perspective, the internal states hold *predictive value* over the external states, and in this sense, we can say that, from an outsider's perspective, the internal states can represent the external states (Hesp et al., 2019, pp. 10, 26; Friston, 2012).

## 2.2. Living models

So far, I have approached the FEP strictly as invoked in practices of scientific modelling. Standardly, however, these models identified in FEP theories are thought to be employed by, instantiated or encoded in living systems one way or another, or even in multiple ways on multiple hierarchical scales (Clark, 2017; Hesp et al., 2019; Kirchhoff & Kiverstein, 2019; Ramstead, Badcock, & Friston, 2018, 2019). Roughly, there is a neurocentric and an embodied variant. The former corresponds to *predictive processing* and is usually explained in terms of the brain *having* a model (Clark, 2016; Hohwy, 2013). The latter is influenced by enactive proposals, typically under the banner of FEP (though not always, see Kirchhoff & Kiverstein, 2019; Kirchhoff & Robertson, 2018) and is usually explained in terms of the system *embodying* a model, or *simply being* a model.

In neurocentric predictive processing, or predictive processing, the brain is thought to encode the generative model – the heart of the predictive machinery – explicitly. In vision, for example, the brain employs the model to resolve the underdetermination problem in perception (Hohwy, 2013).<sup>2</sup> Also known as the poverty of the stimulus problem, this derives from the consideration that any particular retinal image is coherent with an infinitude of possible actual external worlds. Yet, on a daily basis, we perceive a constant world: how does the brain achieve this? Neurocentric predictive processing's answer is that the brain is a Bayesian prediction machine that attempts to anticipate the incoming signals (Clark, 2013, 2016; Hohwy, 2013). The generative

model is thought to be a representational recapitulation of the causal-probabilistic structure of the external world, which will guide the predictions (Gładziejewski, 2016). As explained in Section 2.2, an exact prediction of the external states is computationally intractable. This is why, according to predictive processing, the brain engages in *approximate inference*. The inference using internal and sensory states, which the brain has access to, is thought to be computationally tractable. The probabilistic inference the brain engages in, thus *approximates* the direct inference of the external states. Whenever the brain doesn't get it right, prediction errors occur. The brain's imperative is to minimize such errors by engaging in either perceptual or active inference (Clark, 2013, 2016; Hohwy, 2013). A familiar example elucidates this. If I, say, walk into my office and predict the presence of a cup of tea that is not actually present in the stimulus entering my system, there are two ways to respond. I can apply perceptual inference and update the generative model of the world to not predict a cup of tea there anymore. I can also apply active inference, a prediction conditional on movement, and move about to see if the cup was occluded by other objects or even go to the kitchen to get myself a cup of tea. Interestingly, the *free energy* mentioned in Section 2.1 is formally equivalent to prediction error in the long run in predictive processing (Friston & Stephan, 2007; Hohwy, 2016). Under the FEP, free energy needs to be minimized, and under predictive processing, the formally equivalent value of prediction error needs to be minimized. This is, roughly, why predictive processing is considered a neural implementation of the FEP.

According to the neurocentric approach to predictive processing, 'prediction error minimization is the only principle for the activity of the brain' (Hohwy, 2016, p. 260). Any neural activity, the thought is, conforms to the principle of prediction error (or free energy) minimization. Yet, we figure in many more situations than looking for missing cups of tea, and often these situations have multiple layers of complexity. Consider a simple activity such as walking to university, which entails the long-term aspect of maintaining a job, but also getting to work today, manoeuvring the current traffic situation, dodging potholes or uneven roads and even lifting and putting down legs one by one to walk. All of this is thought to be controlled and computed by the brain. The brain's prediction machinery is thought to be hierarchically layered, mapping onto neurophysiological cortical layers (Kiebel et al., 2008, 2010). These layers deal with different levels of complexity, ranging from colours, surfaces and edges, objects, to objects and, further, objects-in-context (de Bruin & Michael, 2017; Hohwy, 2013).

Furthermore, the blanket states are considered to appear at the ends of the nervous centre, so that the extra-neural body also figures as part of the external states (Hohwy, 2016). This means that the brain not

only attempts to attenuate the external world best as possible but also the body it is in. Inherited through evolution, there are certain predictions that remain constant: those that pertain to the very essential states of the system to maintain homeostasis (Clark, 2013; Friston, 2010; Friston et al., 2009). Cases like hunger, then, appear as prediction errors to be alleviated by applying active inference, a prediction conditional on the action of, say, preparing a meal. In this way, predictive processing intends prediction error minimization to cover all bases.

In what I have called the embodied approach, the organism is thought at least to embody the model in its dynamics, and the model is (typically) considered to be non-representational (Bruineberg et al., 2016; Kirchhoff & Robertson, 2018). According to this view, the organism does not *have* a model, it *is* a model (Friston, 2013; Kirchhoff et al., 2018, p. 4). Ramstead, Kirchhoff, and Friston (2019) phrase it as follows:

generative models are *not explicitly encoded* by physical states. That is, they are *not encoded by states of the brain*. Rather, it is the adaptive behaviour of the system that implements or instantiates a generative model ... adaptive behaviour brings forth the conditional dependences [sic] captured by the generative model, that is, keeping the organism within its phenotypic, characteristic states. (p. 7, emphases in original)

This means that, as opposed to being encoded in neural activity, the model is implicit in the adaptive behaviour of the agent. Initially, this makes the generative model sound epiphenomenal. It seems that what actually does the work is the adaptive behaviour and the conditional dependencies, and that these relations can merely be *captured* in a generative model. This seems to imply that the model is merely a scientific construct that *captures* real statistical relations in the world. Yet, the organism is also thought to 'leverage' the generative model together with the *recognition density* (which, roughly, is a measure of the divergence between the prediction and the actual *sensory state* encountered by the system) (Friston, 2012; Ramstead, Kirchhoff, & Friston, 2019, p. 7). This means that the organism uses a calculated result (the recognition density) to minimize its free energy. In some sense, these probabilistic densities must thus be accessible to the organism. This seems to imply that the generative model is thus *not* epiphenomenal or a mere scientific construct. This reading is further consistent with the claim that it has a 'causal bite' by playing a vital role in action policy selection (Ramstead, Kirchhoff, & Friston, 2019, p. 9).

There are thus two readings open to the embodied approach, both of which are consistent with at least some of the writing: a model-instrumentalist and a

model-realist reading (Bruineberg et al., 2016; Kirchhoff et al., 2018; Ramstead, Kirchhoff, & Friston, 2019). I will discuss this more elaborately in Section 3.1, but roughly it is as follows. The two readings map onto two different interpretations for what it means to *embody* or *simply be* a model, as well as for *approximate inference*. In the *model-instrumentalist* reading, the embodied model is a scientific construct. In the adaptive behaviour of the organism, particular statistical or covariational relations appear between the internal and the external states (e.g. a fish's gills on an evolutionary scale, a hand's covariation with the shape of the door knob on the scale of organismic activity; Ramstead, Kirchhoff, & Friston, 2019, p. 7). These statistical relations are *real* as the relations between the rings of a tree and the years it has been alive are real. The model that these statistical relations 'embody' and captures them is a scientific construct. Approximate inference here means that the organism behaves so that the statistical relations that are brought forth can be *cast* as conforming to the norms of probabilistic inference. In this sense, the organism does not engage in *any* form of inference itself, but it behaves so that the probabilistic relations *approximate* probabilistic, computational inferences. This reading will be defended in this article.

In the *model-realist* reading, the model embodied by the organism exists independent of our scientific modelling practices. The organism *literally is* a model in an objective sense. Under this reading, the organism can 'leverage' particular computational results from the model that it embodies in navigating the environment, without conscious access (Hesp et al., 2019; Ramstead, Kirchhoff, & Friston, 2019, p. 7). Approximate inference here is much the same as it is in predictive processing: directly inferring the external states is intractable, so the organism *approximates* this inference by way of computing over the internal and sensory states instead. In this reading, thus, the organism does engage in inference itself. However, the model that is leveraged is embodied by the organism, not encoded by the brain.<sup>3</sup>

There is a third route that takes both options, in which

the brain *does not just contain* a hierarchical generative model of the world, its dynamics *also instantiate* one – its form and function reflect a physical transcription of causal regularities in the environment that has been optimised by evolution within and across nested spatiotemporal scales. (Badcock et al., 2019, p. 6, emphases added)

In addition, 'different organisms instantiate unique "embodied models" of their specific biological needs and eco-niches' (Badcock et al., 2019, p. 7; referencing Allen & Friston, 2018; Friston, 2011; Gallagher & Allen, 2018; Ramstead et al., 2018). In short, this is a representational interpretation that takes a generative

model to be *encoded* in the brain like neurocentric predictive processing as we have seen above, and *instantiated* both in the brain and in the body conform the embodied approach (Badcock et al., 2019, p. 7).

An interesting result that has come from this modelling practice is that, by applying free energy minimization to a simulated primordial soup, self-organizing patterns appear naturally (Friston, 2012). This can hint at the broad dynamics that must've been in place for the actual ontogeny of life, and is an interesting model of what a minimal life form could require. The mathematical model is also applicable to single cells, organs, organisms and even extends into social situations, cultures, niche construction and evolution. FEP is typically said to unify theories of brain activity (Friston, 2010), but is now also argued to unify studies of life at multiple spatiotemporal scales (Bruineberg et al., 2018; Hesp et al., 2019; Ramstead, Badcock, & Friston, 2018, 2019). Under the FEP, a single modelling practice seems capable of capturing the dynamics of how the brain makes sense of the world as it does in predictive processing, how the organism copes with its environment as seen in the embodied approach, and so on. Indeed, Friston considers it to be 'a theory of every "thing"' (Friston, 2019). This supposed unifying ability is a central attractor for the FEP.

### 3. A model of life and life's model

The wide range of applicability of the Markov blanket formalism is a double-edged sword. On one hand, for many theorists a unified 'theory of everything' is appealing. On the other hand, it requires extra care. We may be able to model any two distinct phenomena using the same tools, yet this does not mean we can extrapolate findings in one application to the other. Nonetheless, I suggest that this may be happening in the FEP literature carelessly. In this section, I shall first show that there are two distinct applications of the FEP model. One use is consistent with an instrumentalist position on the model in which the model is a scientific tool used to study a system. The other use is strictly compatible with a realist position in which the model is implemented so that the system under scrutiny actively uses or literally embodies the model to make statistical inferences. Second, I will show that these two are often conflated, leading to certain invalid claims that overstate the FEP's accomplishments.

#### 3.1. Model entanglement in FEP

In Section 2, I separated scientific modelling practices from ascriptions to organisms and/or brains. In the FEP literature, these two notions typically seem to co-exist peacefully, and intertwine naturally as though one is a seamless continuation of the other. I suggest that

these two are to be distinguished by the agent involved, the modeller. In one take on modelling, the scientist models a self-organizing system (and its environment) to study it by way of a surrogate. This is the scientist's model, and is consistent with instrumentalism. In another take on modelling, the self-organizing system models its environment or 'simply is' a model of its environment, and exploits this model to navigate the world, and maintain its dynamics and physical integrity. This is compatible only with a realist position on models. Another way to pick them apart is that one is a model *of* life: a scientist's construal of relevant relations of the target system, and the other is a model *used by* life: purportedly a statistical model the organism uses unconsciously to navigate the world, life as a modeller. In this section, I will argue that, in the FEP literature, there is a conflation of models as they appear in science and models as ascribed to organisms.

The scientist's model often plays an important role in FEP literature. One may read that the models in FEP are '*representations* of dynamical systems', and 'may provide a *metaphor* for behaviour with different time-scales and biological substrates' (Friston, 2013, p. 1, emphases added); that is, rather than an objective part of nature, the model is human-made: 'an information-theoretic *construct*' (Constant, Ramstead, et al., 2018, p. 5, emphasis added). Furthermore, what makes it interesting is that 'it connects probabilistic *descriptions* of the states occupied by biological systems to probabilistic modelling or inference as described by Bayesian probability and information theory' (Friston, 2012, p. 2101, emphasis added; Korbak, 2019, p. 3).<sup>4</sup>

Representations or metaphors of a system are typically not to be taken literally: a portrait of your colleague is distinct from your actual colleague and the paper the portrait is made out of does not constrain the material your actual colleague is made out of (see also Di Paolo & Thompson, 2014; Tonneau, 2012). When Friston states that probabilistic descriptions connect to 'Bayesian probability and information theory', we do not seem to depart from the level of description. Moreover, both Bayesian probability and information theory are human-made theoretical constructs that are deeply embedded in multiple interrelated practices such as scientific, mathematical, probability-theoretic, and modelling practices. Each of these practices have evolved in intersubjective engagement and require teaching and practice for participation: they were only formed in very specific social contexts. Such formulations seem to indicate that FEP models are *descriptive* of actual biological dynamics, just like meteorological models of actual weather dynamics are used to make statistical inferences about future states of the weather.

Yet, as I have alluded to in Section 2, there are several ways that the mathematical machinery of FEP is ascribed to the organism under scrutiny. The representationalist, neurocentric approach lays their cards out

on the table most clearly. Friston's earlier work is focused on representational learning, and explicitly is aimed at unearthing the model that the brain encodes for this (Friston, 2002, 2003, 2005). Later too, Friston (2010) writes that 'an agent *must have* an implicit generative model of how causes conspire to produce sensory data' (p. 129). More specifically, he discusses 'the form of the generative model and how it *manifests* in the brain' (Friston, 2010). This indicates a realist take on models as encoded, exploited and manipulated in the neural architecture. This is also the approach taken up by Hohwy (2013, 2016), Gładziejewski (2016) and Clark (2013, 2016) most notably, and has seen widespread further influence. Their position is, roughly, that the brain encodes a generative model of the extra-neural world that, in some sense, recapitulates the causal-probabilistic structure of the world so as to maintain homeostasis (via active inference) and infer the causes of sensory inputs (via perceptual inference).

There is also an embodied approach to the FEP. Where the neurocentric approach clearly commits to a realist position, we have seen in Section 2.2 that the embodied approach remains ambiguous, mirroring Friston's pioneering writing (Friston, 2012, 2013). This is conspicuously expressed in the descriptive language employed by FEP theorists. The embodied model finds a basis in the claim that 'biological systems can distil structural regularities from environmental fluctuations [...] and embody them in their form and internal dynamics' (Friston, 2012, p. 2101). This is contrasted with the notion that an organism *has* a model of its environment. Kirchhoff (2018) writes,

by 'model' it does not follow that an organism has an internal, representational model of its niche and that it is this model that does all the cognitive work (if you like). Instead, an organism *is* a model, viz., the causal and statistical regularities reflected in some environment are reflected in some phenotype, i.e., model. (p. 761; see also Kirchhoff et al., 2018, p. 4)

An example of this is 'the physiological make-up of a fish, say, as a model of the fluid dynamics and other elements that constitute its aquatic environment – its internal dynamics depend on the dynamics of the niche' (Bruineberg et al., 2016; Kirchhoff et al., 2018, p. 4).

Taken in this sense, in the embodied approach, the FEP seems like a *definition* of life, or more specifically, 'a mathematical *formulation* of how adaptive systems (that is, biological agents, like animals or brains) resist a natural tendency to disorder' (Friston, 2010, p. 1, emphasis added), so that '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' (Ramstead, Kirchhoff, &

Friston, 2019, p. 4). Implied, but not explicit, is that the system *does not actually* predict, infer, track or minimize a quantity called variational free energy, but *merely looks as if*. The probabilistic model merely tracks certain real statistical relations in the organism-environment system. In this sense the organism is thought to *approximate* inference. It seems again that, with regard to the status of the model, we take the scientist's perspective. When we take a fish's phenotype as an example, and we look at the shape of the fish, the scales and the gills, they can, for an external observer, represent or hold predictive value with regard to the water it lives in. We, as external observers, can then make inferences on the basis of what are here marked as the internal states (the phenotype) about the external states (the environmental niche), and in this sense, we can use the fish's physiology as a model for the external states.

The instrumentalist reading seems encouraged as more realist-leaning terminology is often accompanied by 'scare quotes' or termed *implicit*, implying it should be read in a non-standard way. In Bruineberg et al.'s (2018) model of niche construction under the FEP, we read that the 'effect of the agent on the environment can be understood as the environment "learning" about the agent', and in this sense, 'the agent and the environment "get to know each other"' (Bruineberg et al., 2018, p. 162). In Badcock et al. (2019) attempt at a 'free-energy formulation of the human psyche', we read that 'brain dynamics (i.e. the general "behaviour" or "ensemble dynamics" of neural mechanisms) can be described as realising an implicit hierarchical generative model: a Bayesian hierarchy of "hypotheses" or "best guesses" about the hidden causes of our sensory states' (p. 5).

Recall that, according to Ramstead, Kirchhoff, and Friston (2019), 'it is the adaptive behaviour of the system that implements or instantiates a generative model ... adaptive behaviour brings forth the conditional dependences [sic] captured by the generative model' (p. 7). The model is thus 'realised' through a specific sort of action of the organism (p. 9). This could be taken to mean that what is *real* is adaptive behaviour of the organism, and the particular statistical relations that exist between the organism and its environment that can be *described* as a generative model. In this sense, when the organism 'leverages' the recognition density (a measure of the match between the prediction of the generative model and the actually encountered input), perhaps what the organism exploits there is *not* the recognition density itself, but merely the conditional dependencies between the organism and its environment. These conditional dependencies could be thought of as covariational in nature (Bruineberg & Rietveld, 2014). In practice, this happens when I exploit the covariation of the shape of my hand and the door knob in opening the door, or the structural similarities between

a cardboard box and a table top when I place the former on the latter. As such, the embodied approach seems to afford an instrumentalist reading.

Hesp et al. (2019) also emphasize the *implicit* status of the model in FEP, and they argue 'it means that these Bayesian concepts do not require the system itself to be "conscious" of inferences in any way or that these inferences need to be "explicit" and couched in propositional or linguistic terms' (p. 231). 'Implicit', thus minimally means that no conscious access is required, and that the inferences made need not be propositional or linguistic in nature. This purely negative description of what implicit inference means rules out the caricature of an organism consciously engaging in advanced statistical computations whenever it moves about. This is consistent with an instrumentalist reading in which the organism does not engage in inference at all, it merely enacts certain statistical relations that can be captured in a model. Yet, it is also consistent with a predictive processing reading in which the organism, subpersonally and unconsciously, infers the world in strictly probabilistic, non-propositional and non-linguistic terms (Wiese, 2017).

Indeed, it is also said that the organism directly uses (or 'leverages') the generative model that it embodies, or conversely, that the generative model actively controls the organism's coupling to the environment. For example, Bruineberg et al. (2018) assume 'the agent *uses its generative model*' (p. 164, emphasis added), and the 'generative model functions to regulate and control the agent's coupling to the environment' (Kirchhoff & Kiverstein, 2019, p. 59). More explicitly, Kirchhoff and Kiverstein (2019) state that '*generative models* coupled in active inference to generative processes are uniquely equipped to *make use of* the properties of an organism's embodiment and associated species-typical environment' (p. 59). Here, the generative model *makes use of* the properties of an organism's embodiment and associated environment. As such, it cannot simply be that the generative model is merely a 'mathematical formulation' of a complex interplay between an organism and its environmental niche. It is seen as separate from the organism's embodiment, insofar it can *make use of* properties thereof. Indeed, in each of these citations, it seems the generative model is ascribed active force in the causal web of an organism: it has 'causal bite' (Ramstead, Kirchhoff, & Friston, 2019). Under an instrumentalist reading, a model cannot have a direct *causal bite*: a description of a particular behaviour does not feature in causing the target behaviour (barring self-referential descriptions). If it were merely a description, it would describe features of and (statistical) relations between the system and its environment (or the agent-environment system) that have causal bite. The model, on itself, remains on a descriptive level.

Despite their theoretical differences, as far as the status of models in FEP is concerned, the embodied

approach in which the organism simply is a model is thus equivalent to neurocentric predictive processing. In both approaches, the model exists independent of human practice, computational results are either encoded and manipulated, or embodied and leveraged. This is also seen in the manner in which inference can be said to be approximated. Contrary to a model-instrumentalist reading, approximate inference in both neurocentric predictive processing and a model-realist embodied approach is cashed out in terms of an inference over internal and sensory states, as opposed to a direct inference of the external states (which is intractable).

For both neurocentric and embodied approaches to the FEP framework, it thus seems that there is a conflation of models that are created in a scientific endeavour and a model that is used, leveraged or employed by, or instantiated in (the dynamics of) an organism. With no sense of direction, FEP theorists seem to veer from scientific models of life to living models.

### 3.2. The error in conflation

‘So what?’ one may think. There are two possible modes of interpretation, and one may, for various reasons, prefer one over the other. The issue is that these two interpretations are often muddled and few take care to distinguish the two (Colombo & Wright, 2018). Consider how Friston (2013) sums up exactly why free energy minimization and inferentialism are central to life:

[1] Under ergodic assumptions, the long-term average of surprise is entropy. [2] This means that minimizing free energy – through selectively sampling sensory input – places an upper bound on the entropy or dispersion of sensory states. [3] This enables biological systems to resist the second law of thermodynamics – or more exactly the fluctuation theorem that applies to open systems far from equilibrium. [4] However, because negative surprise is also Bayesian model evidence, systems that minimize free energy also maximize a lower bound on the evidence for an implicit model of how their sensory samples were generated. [5] In statistics and machine learning, this is known as approximate Bayesian inference and provides a normative theory for the Bayesian brain hypothesis. [6] In short, biological systems act on the world to place an upper bound on the dispersion of their sensed states, while using those sensations to infer external states of the world. (p. 2, numbering added)

Let us dissect this quote sentence per sentence. In [1], Friston points out an equivalence in a formal model of long-term average of surprise and entropy under specific assumptions. In [2], we see a formal implication of [1]. In [3], the terms in the model are linked to their target system: biological systems. In [4], the terms in the

model (*average of surprise* and *entropy*) are also linked to Bayesian model evidence. Here, he thus introduces a further formal equivalence. So far, Friston stated a mathematical formulation of organisms’ resistance to the second law of thermodynamics and offered formal equivalences of terms in that formulation. Yet, the line gets blurred when Friston introduces ‘the implicit model of how the organism’s sensory samples were generated’ in [4]. The aforementioned model is now ‘implicitly’ implemented in the system. In [5-6], this quickly gets expanded with further equivalences to entail that ‘biological systems ... us[e] those sensations to *infer* external states of the world’ (emphasis added). The organism here uses the model to make statistical inferences about the external states of the world. Simplified a little, it seems that the argument in the above quote can be put as follows:

1. Biological systems with  $X$ .
2. Formally,  $X$  is represented by  $x$ .
3. Formally,  $x$  is equivalent to  $y$ .
4. Formally,  $y$  is equivalent to  $z$ .
5. Thus, biological systems with  $z$ .

The issue is, I argue, that a formal representational relation between two systems is conflated with one of natural identity. Put differently, the equivalence of term  $x$  that represents target feature  $X$  in the world, with a different term  $y$  does *not* warrant the interchangeability of target feature  $X$  with  $y$ .  $X$  concerns a real-world feature, whereas  $x$  and  $y$  are both *mathematical descriptions* of such a real-life feature. What it *does* warrant, however, is the following. If  $X$  can be mathematically described as  $x$ , and  $x$  is equivalent with  $y$  and  $z$ , then  $X$  can be mathematically described as either  $x$ ,  $y$ , or  $z$ . *Within* the realm of mathematical description, these terms are equivalent. In the above quote by Friston (2013, p. 2), there is a hidden argument in which a mathematical formulation of behaviour is turned into an underlying mechanism for that behaviour.

This is unjustified, yet gives off the pretence that the battle on how to interpret modelling under FEP is already won, hidden behind complex mathematical and statistical jargon. It suggests that for a biological system to model and infer the external states of the world is a tautology (Friston, 2013; Friston & Buzsáki, 2016). We see this being picked up in the literature too. Constant, Bervoets, et al. (2018) for example, say, ‘To be a living system then *means* dynamically modeling oneself in relation to one’s body, and one’s environment’ (p. 3). To a lesser extent, Hohwy (2016) makes a similar claim. In this, the representational relation between internal and external states within the Markov blanket formalism is reified so that the internal states in themselves represent the external states, irrespective of there being an external observer. This is taken to hold for the target



system as a matter of course once one accepts the basic description of life under the FEP (Hohwy, 2016).

The above discussion of model-realist versus model-instrumentalist readings of the embodied approach in FEP also seems indicative of the issue. In one reading, there is no conflation: there is a realist attitude with regard to the statistical relations enacted by an organism's adaptive behaviour, and an instrumentalist attitude with regard to the scientific model that captures these relations in a mathematical formalism. Approximate inference is understood in terms of the relations brought forth by the behaviour approximating inference on the basis of the model as we could do as modellers. Yet, in another reading of the same literature, there is a conflation. Simply being alive, then, means that the organism will model its environment because of the statistical relations that it is accompanied by. The existence of statistical relations, as we have seen, does not warrant the (science-independent) existence of a model in itself: the rings in a tree trunk covary with the years the tree's been alive, but this does not entail that the rings model the environment in any sense independent of human practice. This conflation seems to permeate the very basis of the FEP approach to cognition. As such, it is imperative to distinguish models *of* a target system from models *used by* a target system.

#### 4. Anti-realism: against the life's model interpretation

I have argued that in the FEP literature, two approaches to models, instrumentalist and realist, are present without being properly distinguished. Blurring the lines between these two takes on models can cause theoretical mishaps. Here, I will argue that, because of these errors, only the instrumentalist approach to FEP models is warranted.

##### 4.1. Overgeneration of models

The applicability of Markov blankets extends far beyond only biological systems. Friston (2013) even says that '*any system that exists will [...] engage in active inference*' (p. 2). Although this formulation seems to draw no boundary at all, even weaker ways of understanding FEP related notions such as Markov blankets have very broad applicability. At least every single biological system has a Markov blanket, ranging from single cells to macroscopic organisms (Kirchhoff et al., 2018). Certain non-living systems such as Huygens pendulums also have a Markov blanket, and engage in active inference (Friston, 2013; Kirchhoff et al., 2018). '[C]arefully chosen nodes of the World Wide Web surrounding a particular province' may constitute a Markov blanket (Ramstead et al., 2018). The Markov

blanket formalism is also applicable to a process called 'niche construction', according to which organisms and their environments co-evolve and mutually influence each other. Here, the organism models the environment, and the environment models the organism, because, from the environment's perspective, the organism is external (Constant, Ramstead, et al., 2018). Ramstead et al. (2016) further suggest it is applicable to human cultures and may be able to elucidate how cultural practices take shape due to social free energy minimization.

This wide range of applicability of the formalism constrains the range of applicability of a realist approach. Although we may be able to imagine a statistical model implemented in a neural architecture, it is harder to imagine a bacterium subpersonally engaging in advanced statistics (Hohwy, 2013, 2016). Stranger still is when 'the environment' is thought to model the organism. It is unclear what the physiological substrate of this model could be, as the 'environment' is composed of a wide variety of physiologically diverse systems like different and different sorts of trees, plants, animals and so on. This would require a sort of hive-mind of systems that may have a hard time communicating. The 'environment' is also defined relative to a particular agent, and, as the agent moves, the constituents of the 'environment' are also in constant flux. The World Wide Web, or human cultural evolution all do not easily afford the autonomy of *implementing* models and *making statistical inferences*.

The overgeneration of Markov blankets is not new in the literature. A similar issue is dealt with in Kirchhoff et al. (2018), in which an extra property is suggested to be added to distinguish non-autonomous active inference systems from autonomous active inference systems. Huygens pendulums, for example, engage in '*mere* active inference', whereas organisms engage in '*adaptive* active inference'. The distinction between the two is marked by whether one is 'entirely "enslaved" by its here-and-now – and, in particular, its precedents' or not, the latter entails that one is capable of 'modulation of [one's] sensorimotor coupling to [one's] environment' (Kirchhoff et al., 2018, p. 5). This is thought to constrain the ascription of autonomy according to FEP to systems that we would actually consider autonomous. Certainly, this clears Huygens pendulums from the ascription of autonomy, but one may wonder whether an organism's environment, or human cultural evolution are only 'enslaved' by their here-and-now and their precedents any more than single living organisms. It seems they may allow a stronger sense of novelty or adaptability to their perspectively determined external states than a Huygens pendulum. There is also the question to what extent an autonomous system is *not* 'enslaved' to its here-and-now and its precedents. We may be able to modulate our sensorimotor coupling to the environment, but one may wonder to what extent this is done in a way that is *not* enslaved by

our environment and our interactional history (our precedents) with that environment. There is surely a sense in which our environment, our phylo- and ontogenetic interactional histories can be said to simply *determine* our sensorimotor coupling with that environment. Consider the purported open-endedness in minimizing the prediction error in me expecting a cup of tea on my desk. Whether I will either ‘update my model’ and accept the cup is not there (perceptual inference) or get myself a cup of tea (active inference) will depend on my current state and my interactional history. Whether I am thirsty or not, but also whether I, through my previous interactions, have learned that a cup of tea can be conducive to a productive afternoon and how to make a cup of tea, are essential in the determination of my actions. Indeed, when considering the whole organism–environment system, the issue may be less open-ended.

Moreover, though this novel distinction between *mere* and *adaptive* active inference touches on a *similar* issue, it does not affect overgeneration of modelling in itself. Indeed, the Markov blanket formalism is still applicable to Huygens pendulums, environments and cultural practices, even if we do not grant them autonomy, and they are still said to be modellers of their external states. This is an issue because for none of these things it is clear how they could implement a model, nor make statistical inferences. Proposing a solution by suggesting we take an instrumentalist approach to non-autonomous systems and a realist approach to autonomous systems, is *ad hoc*. It would require additional argumentation or a principled reason to show that of all systems that can be modelled in a particular way, only those that are also autonomous should be described as implementing and using that model, whereas the others merely *act as if*.

#### 4.2. Covariation, no content, no model?

Let us entertain the possibility that the mathematical construct is implemented biologically, either neurally by having a model, or organismically by simply being a model, independent of modelling practices. It is still open whether the organism actually uses, or embodies and leverages, a *model*. In this section, I will argue that a realist position on the *mathematical mechanisms* described by the FEP still does not warrant a realist position on the *models* as used by FEP theorists. This relates to the FEP debate on representations. The primary argument will be roughly as follows. In Section 4.2.1, I discuss that models in FEP as thought to be implemented in organisms are not representational, but covariational in nature (Bruineberg et al., 2016; Bruineberg & Rietveld, 2014; Kirchhoff & Robertson, 2018; Ramstead, Kirchhoff, & Friston, 2019). Continuing, in Section 4.2.2 I consider what it takes to be a model. The literature on the epistemic and

ontological grounds of models in science may comprise a wide variety of positions, but most agree that they are representational in some way, shape or form (Frigg & Hartmann, 2018). This means that, in virtue of being non-representational when used by the organism, the mathematical machinery as thought to be instantiated in the organism’s dynamics cannot be a model. Recently, a non-representational position on scientific models has been proposed that could, *prima facie*, be considered a solution for the model-realist FEP position (de Oliveira, 2018). Yet, this is to no avail, because the pragmatist appeal to human practices (de Oliveira, 2018) relies on is unavailable to FEP model-realists. As such, even if we entertain the possibility that the mathematical constructs are implemented biologically, these constructs are not models.

**4.2.1. Covariation, no content.** It will be fruitful to first describe Kirchhoff and Robertson’s argument for a non-representationalist interpretation of FEP. Recall that according to FEP, an organism’s primary imperative is to minimize its free energy. In predictive processing, the neural implementation of the FEP, this means that the brain continuously attempts to match its internally generated predictions with the signals that enter the system. To call this representational, it is crucial to show that *misrepresentation* is possible. Kiefer and Hohwy (2018) have attempted to do this in reference to the Kullback–Leibler divergence (see also Friston, 2013). Roughly, they argue that this divergence is a measure of the difference between the brain’s estimate and the incoming signal.<sup>6</sup> This, they argue, means that it measures misrepresentation.

However, as Kirchhoff and Robertson (2018) argue, the Kullback–Leibler divergence only measures a *Shannon-informational* divergence. Shannon information is covariational information, in itself not representational (Godfrey-Smith & Sterelny, 2016). This means, roughly, that two particular systems covary: when one system changes, it is likely that the other changes in a similar fashion, broadly construed. In this sense, if a scientist knows that system A and system B covary, then knowledge of one system’s state increases the information they have about the other system’s state. A typical example of covariance is the rings on the trunk of the tree that reliably covary with the amount of years it has lived. Covariance is not necessarily a one-to-one relation, and systems typically covary more or less strongly. In this particular sense, Kirchhoff and Robertson (2018) argue that the internal states of the organism covary with the external states: reliably, but not per se one-to-one. After all, we do make mistakes regularly and this betrays misalignment. This misalignment, however, is a form of negative covariance, not one of misrepresentation. As such, for system A to ‘model’ or ‘infer’ system B, the dynamics of system A

must covary reliably with the dynamics of system B (Bruineberg & Rietveld, 2014, p. 7; Kirchhoff, 2018, p. 762). This relation is thus not representational. Both modelling and inference are captured by dynamical covariation (Friston, 2013).<sup>7</sup> Kirchhoff and Robertson (2018), unlike Bruineberg and Rietveld (2014) for example, do not argue to alter FEP by expanding it with a theory of affordances or fit. They merely show that the mechanisms proposed by FEP are, contrary to popular opinion, in themselves *not representational*.

Although the generative model is not representational, Ramstead, Kirchhoff, and Friston (2019) emphasize it does use ‘exploitable structural similarities’ (p. 11). Such exploitable structural similarities lie at the foundation of the notion of structural representation popular in predictive processing (Gładziejewski, 2016). However, proponents concede mere exploitable structural similarity is insufficient to ground a notion of representation. Indeed, when I open a door, I exploit the structural similarities between the shape of my hand and the door knob. Yet, my hand in no sense *represents* the knob. As such, the exploitable structural similarities do not allow for representations to be snuck back into the FEP.

**4.2.2. No model?** What makes a model a model? Where do models get their epistemic import from? How can it be that from studying a model of a target system, we can learn more about the target system? These questions have been discussed elaborately (Frigg & Hartmann, 2018). Despite a multiplicity of views, there is a broad consensus that models are inherently representational. The debate is, largely, about *how* models are representational. A few options that constitute this representational relation are isomorphism (van Fraassen, 1980) or similarity (Giere, 1988; see Frigg & Hartmann, 2018 for an overview). The representational relation is then thought to be constituted solely by the extent of similarity the model displays to the target system. These views have been heavily criticized, and there is now a broadly shared consensus to include reference to *use*: a model is representational because it is used as such, which may include further reference to objective similarities between the model and the target system (de Oliveira, 2018, p. 12; see Giere, 2010; Suarez, 2003; van Fraassen, 2008 for example). A more recent non-representationalist account suggests that a model is a surrogate used for a target system within a particular scientific social practice of creating models ‘in terms of skill development and learning transfer’ (de Oliveira, 2018, p. 23).

There are thus essentially two options: models are representational in some way, shape or form, or they are not, in which case their epistemic import is explained pragmatically by surrogacy and human practices of skill development and learning transfer. Above

I have there is good reason to think that models in FEP are non-representational. That which is purported to measure misrepresentation, an essential feature of representation, instead measures negative Shannon-informational covariance. This means that, if we follow the broad consensus on models in science, ‘models’ as they are instantiated in organisms according to the FEP simply *do not count* as models. The non-representationalist view on models in science is of no help either. This has, roughly, two conditions: the purported model should (1) be used as a surrogate of a target system and (2) be embedded in the scientific social practice of modelling (de Oliveira, 2018). Although the surrogacy condition could potentially be met, models in FEP cannot rely on human practices that are thought to be *products* of the social interaction that the models are thought to *precede*. As such, a realist position on the mathematical relations as described in FEP still does not warrant a realist position on these models: covariance is simply too weak a notion.

## 5. Model-instrumentalism, covariation-realism

Above I have argued that there are two issues with model-realism in FEP. The first is that the model is widely applicable, leading to over-generation of models and trivialization of the notion. The second is that the notion of modelling in FEP is not representational, but covariational in nature. Models as they are commonly understood in science, however, *are* representational. A non-representational approach exists, but to no avail for the FEP realist. De Oliveira’s approach relies on a model’s embeddedness in scientific social practices. FEP models are thought to figure at the very core of life itself: they should *precede* the sorts of social behaviour that form our practices. The mechanisms described by the FEP thus do not seem to fit the bill for a model anymore. One may now wonder whether this also affects the instrumentalist position on models in FEP. Perhaps, instead, we should not speak of models in FEP *tout court*.

I argue that this is not necessary. Overgeneration in itself is not an issue. As a modeller, one chooses to model a target system in a particular way. There being a multiplicity of potential target systems that could be modelled using the same tools does not detract from the legitimacy of the model. Furthermore, models as implemented in organisms in FEP cannot rely on a sociocultural practice of using models in science. An instrumentalist approach to modelling in FEP does not encounter the same obstacle. After all, FEP models are created by experts in mathematical modelling that have been trained in a scientific context. Indeed, it seems that models in the FEP as formed, used and exploited by people in a scientific sociocultural context are prime

exemplars of models, whether we take a representationalist view or not. As such, the instrumentalist view quite clearly is free of aforementioned worries.

In Section 2.2, I discussed the embodied approach to the FEP, and in Section 3.1, I discussed more elaborately the extent to which the theory as proposed in the literature affords an instrumentalist or a realist reading. I argued that there is plenty of wiggle room, allowing for both a model-instrumentalist and a model-realist reading of the same literature. Relying on this wiggle room, we can see the broad contours of an instrumentalist approach. In this view, the *models* nor the *mathematical machinery* are taken to be *instantiated* in the organism. The adaptive behaviour, nor the relations between the organism and its environmental niche instantiate, encode or embody a generative model in any realist sense. There is *approximate* inference so that the organism's adaptive behaviour corresponds to probabilistic inference in a scientific model. Yet, the organism does not engage in statistical inference in any realist sense. It is not for the organism, nor the brain, as it is for the scientist (contrary to Hohwy, 2013, 2016).

Instead, what in this view does warrant a *realist* position are the different statistical and covariational relations between the organism and its environment. This view takes a realist position on the notion that adaptive behaviour brings forth, or displays the conditional dependencies, that can then, by scientific modelers, be *captured* in a generative model (Ramstead, Kirchhoff, & Friston, 2019, p. 7). This view also takes a realist approach to the notion, further, that we form increasingly tighter covariational relations with our environment both in a niche-construction, designer environment sense and in an active, ontogenetic adaptational sense, such as described in FEP terms in (Bruineberg et al., 2018; Clark, 2016, section 9.5; Constant, Ramstead, et al., 2018; Hesp et al., 2019). The mathematical model created by Bruineberg et al. (2018) in particular is a good way to investigate minimal conditions for niche construction, as well as investigate certain statistical relations between the agent and the environment. That is to say, an instrumentalist view need not impose constraints on the FEP research programme itself. It does, however, impose constraints on the conclusions that can be drawn from the findings.

There are a few possible objections to constraining FEP models to an instrumentalist take. One possible objection is that models in FEP simply are quite unlike any model in science. Indeed, though scientific models are thought to be representational, these models are covariational, and they are further thought to be implemented by organisms to minimize free energy or prediction error. System A can be said to model the dynamics of system B if the dynamics of A covary with the dynamics of B (Bruineberg & Rietveld, 2014; Kirchhoff & Robertson, 2018). This is simply a *special case* of modelling.

First, this does not match the consensus view on models as described in Section 4. Even a special case should be thought to be captured in a definition. If it is too special to be captured in the definition, this may be good reason to consider that, whatever it is, it may not be a model. Second, persisting it is a model results in trivialization of the notion, and it may not indicate to a reader what the author thinks it does. When we say that the organism *has* or *is* a model that the organism uses or leverages to make statistical inferences, it is difficult to imagine what this means if it is not contentful. According to Kirchhoff and Robertson (2018) and Bruineberg and Rietveld (2014), both to model and to infer are captured by dynamical covariation between two systems. A model is thus a covariation relation. A statistical inference is, too, a covariation relation. As both a model and an inference are now identical, it becomes difficult to see what it would mean for an organism to *use* its covariation relation with the dynamics of B (the model) to covary with the dynamics of B (to make a statistical inference). The errors shown in Section 3.2 are made here. By maintaining a realist position on models in FEP, even in this deflationary sense, we find that some theorists draw unjustified conclusions. This may be avoided if we use clearer terminology. If an FEP model unearths statistical (in)dependencies and covariation relations between internal and external states, we may as well call them by name: statistical relations and covariation.

Another objection is that the best explanation for why our predictive models work is that the organism actually uses these models itself.<sup>8</sup> It is true that, if an organism were to employ the same models we use to predict its behaviour, that would explain our predictive success, but this does not work *vice versa*. We model weather dynamics, animal population rates as well as biodiversity increases or decreases, yet we are not inclined to take a realist position on these models, despite their predictive successes. For biological organisms to be the exception to this rule, we would need additional argumentation. Note that this is a distinct point from scientific realism versus instrumentalism *tout court*. The model-instrumentalist view is perfectly compatible with certain forms of realism with regard to the particular features of the target system that are picked out by the model. Recall the model-instrumentalist reading of the embodied approach. This is a realist position with regard to the patterns of covariation and the statistical relations, but an instrumentalist position with regard to the model that captures those relations. Put differently, the *scientifically devised* models of organism-environment systems (for example) pick out *real* statistical covariation relations.

## 6. Conclusion

In this article, I have argued that we should take an instrumentalist approach to models in FEP. I have shown that in the literature instrumentalist models as created by

scientists and realist models as thought to be implemented in and used by organisms are often conflated, and this can lead to erroneous conclusions. Furthermore, a realist take on models in FEP is unwarranted. This is due to two reasons. First, a realist approach leads to overgeneration and trivialization of models. Second, a realist approach to the mathematical mechanisms described in FEP still does not warrant a realist take on models: the mechanisms described are covariational in nature, which falls short of meeting the conditions of both representational and non-representational takes on modelling in science. I have further shown that an instrumentalist approach to models in FEP is safe from these arguments, and is a justified use of modelling. This does not keep one from embracing the results of FEP modelling, and taking a realist approach to the particular dynamical relations between, say, the organism and its environment. I suggest that, in going forward, an instrumentalist position on models in FEP should be maintained. This signals clearly to the reader what is being meant and may allow us to avoid conflation. Indeed, if a model of a target system unearths a covariational relation between internal and external states, we should take it as the covariational relation that it is.


### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Thomas van Es is funded by an Aspirant grant from the FWO (Research Foundational Flanders), Grant number: 1124818N.

### ORCID iD

Thomas van Es  <https://orcid.org/0000-0002-4097-3566>

### Notes

1. In this article, I will refer to such embodied/extended approaches as ‘the embodied approach’. This captures multiple views, with plenty of differences among them. Whereas Bruineberg et al. (2016) focus on an embodied approach, Kirchhoff and Kiverstein (2019) argue for a more flexible, context-sensitive approach, ranging from extension into the world to merely the brain (see Clark, 2017; Hesp et al., 2019; Ramstead, Badcock, & Friston, 2018, 2019 for similar positions).
2. Note that the underdetermination problem depends on how the situation is set up. When we take a single snapshot of the world, there is strong underdetermination. In an ecological situation, however, we do not take snapshots of the world, but action and perception unfold simultaneously over time. If one approaches perception as a doing (thus with duration, extended through time), the poverty of the stimulus disappears (Gallagher & Allen, 2018; Myin, 2016). If one approaches perception not like a film as a sequence of snapshots, but instead as a diachronically constituted process with duration, our attunement to sensorimotor contingencies shaped by a phylo- and ontogenetic interactional history, much of the uncertainty of the external world dissipates (Di Paolo et al., 2017; Kirchhoff, 2015; O’Regan & Noë, 2001). Indeed, if we note that we do not passively take in the world, but actively explore it, the poverty of the stimulus ceases to exist as such. Exploration of this idea is outside the scope of this article, but the references in here offer accessible introductions.
3. It is not clear *how* the organism infers on the basis of the model that it embodies. In neurocentric predictive processing, one can imagine the brain, cast as an agent, manipulating and making inferences on the basis of a statistical model, analogous to the way we do in our modelling practices. There is a, fairly obvious, mereological fallacy here (see Hohwy, 2016), and a few other issues pertaining to invoking representations independent of human practices reappear (Hutto & Myin, 2013; Tonneau, 2012). Nonetheless, there is a clear *modus operandi* for the exploitation of the model. In the embodied sense, it is much less clear. What agent engages in inference? What does it mean for the organism to *implicitly* infer the external states over and above the organism’s behaviour to instantiate particular probabilistic relations that *can be captured* in a model?
4. There are plenty more of such phrasings in the FEP literature. Friston (2013) writes that ‘any system that exists will *appear* to minimize free energy’ (p. 11). FEP’s success makes it an ‘important *metaphor* for neuronal processing in the brain’ (Friston, 2012, p. 2101, emphasis added).
5. As an anonymous reviewer pointed out, it is important that this is not what models in the FEP are used for in science. Instead, they are descriptive models that allow for a mathematical formulation of the organism-environment dynamics, but they do not afford new predictions, *per se*. This is only meant to point out the manner in which talk of inference makes sense in the FEP framework.
6. Technically, the Kullback–Leibler divergence measures the difference between the brain’s recognition model and the actual posterior probability (Friston, 2013; Kiefer & Hohwy, 2018). This means that it measures the divergence of the brain’s approximation of the prior probability and the ‘true state of affairs’ (Kiefer & Hohwy, 2018, p. 23). See Friston (2013) and Kiefer and Hohwy (2018), and Kirchhoff and Robertson (2018) for a more technical discussion on the matter.
7. The equivalence of inference and covariation is also picked up in Badcock et al. (2019). We read that ‘our actions will tend to infer or reflect the statistical structure of the environment to which they are coupled’ (p. 6). Both ‘infer’ and ‘reflect’ are used as interchangeable, equivalent terms. Reflection here should be thought of in the way that the body of a fish reflects the environment it has evolved and lived in (Friston, 2013; Kirchhoff et al., 2018). This reflection, as we have seen here, is captured by a covariation relation between the organism and its environment, which in turn means that to infer is again seen as captured by a covariation relation.

8. Rescorla (2016) and Korbak (2019, p. 16) appear to defend something akin to this position.

## References

- Allen, M., & Friston, K. (2018). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, 195, 2459–2482.
- Badcock, P. B., Friston, K. J., & Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews*, 31, 104–121.
- Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*, 195, 2417–2444.
- Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience*, 8, Article 599.
- Bruineberg, J., Rietveld, E., Parr, T., van Maanen, L., & Friston, K. J. (2018). Free-energy minimization in joint agent-environment systems: A niche construction perspective. *Journal of Theoretical Biology*, 455, 161–178. <https://doi.org/10.1016/j.jtbi.2018.07.002>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 181–253.
- Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.
- Clark, A. (2017). How to knit your own Markov blanket resisting the second law with metamorphic minds. In T. Metzinger, & W. Wiese (Eds.), *Philosophy and predictive processing*. MIND Group. [https://www.research.ed.ac.uk/portal/files/39856902/How\\_to\\_Knit\\_Your\\_Own\\_Markov\\_Blanket.pdf](https://www.research.ed.ac.uk/portal/files/39856902/How_to_Knit_Your_Own_Markov_Blanket.pdf)
- Colombo, M., & Wright, C. (2018). First principles in the life-sciences: The free-energy principle, organicism and mechanism. *Synthese*. Advance online publication. <https://doi.org/10.1007/s11229-018-01932-w>
- Colombo, M., Elkin, L., & Hartman, S. (2018). Being realist about Bayes, and predictive processing. *The British Journal for the Philosophy of Science*. Advance online publication. <https://doi.org/10.1093/bjps/axy059>
- Constant, A., Ramstead, M. J. D., Veissière, S. P. L., Campbell, J. O., & Friston, K. J. (2018). A variational approach to niche construction. *Journal of the Royal Society Interface*, 15, Article 20170685. <http://doi.org/10.1098/rsif.2017.0685>
- Constant, A., Bervoets, Jo., Hens, K. & Van de Cruys, S. (2018). Precise worlds for certain minds: An ecological perspective on the relational self in autism. *Topoi*. Advance online publication. <https://doi.org/10.1007/s11245-018-9546-4>
- de Bruin, L., & Michael, J. (2017). Prediction error minimization: Implications for embodied cognition and the extended mind hypothesis. *Brain and Cognition*, 112, 58–63.
- de Oliveira, G. S. (2018). Representationalism is a dead end. *Synthese*. Advance online publication. <https://doi.org/10.1007/s11229-018-01995-9>
- Di Paolo, E., Burmann, T., & Barandiaran, X. (2017). *Sensorimotor life: An enactive proposal*. Oxford University Press.
- Di Paolo, E., & Thompson, E. (2014). The enactive approach. In L. Shapiro (Ed.), *Routledge handbook of embodied cognition* (pp. 68–78). Routledge.
- Frigg, R., & Hartmann, S. (2018). Models in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2018 ed.). <https://plato.stanford.edu/archives/sum2018/entries/models-science/>
- Friston, K. (2002). Functional integration and inference in the brain. *Progress in Neurobiology*, 590, 113–143.
- Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, 16, 1325–1352.
- Friston, K. (2005). A theory of cortical response. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360, 815–836. <https://doi.org/10.1098/rstb.2005.1622>
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(11), 127–138.
- Friston, K. (2011). Embodied inference: Or ‘I think therefore I am, if I am what I think’. In W. Tschacher, & C. Bergomi (Eds.), *The implications of embodiment: Cognition and communication* (pp. 89–125). Imprint Academic.
- Friston, K. (2012). A free energy principle for biological systems. *Entropy*, 14, 2100–2121. <https://doi.org/10.3390/e14112100>
- Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10, Article 20130475.
- Friston, K. (2019). *A free energy principle for a particular physics*. Unpublished manuscript.
- Friston, K., & Buzsáki, G. (2016). The functional anatomy of time: What and when in the brain. *Trends in Cognitive Sciences*, 20(7), 500–511.
- Friston, K., Daunizeau, J., & Kiebel, S. J. (2009). Reinforcement learning or active inference? *PLOS ONE*, 4(7), Article e6421.
- Friston, K., & Kiebel, S. (2009). Cortical circuits for perceptual inference. *Neural Networks*, 22, 1093–1104.
- Friston, K., & Stephan, K. (2007). Free energy and the brain. *Synthese*, 159(3), 417–458.
- Gallagher, S., & Allen, M. (2018). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*, 195(6), 1–22.
- Giere, R. N. (1988). *Explaining science: A cognitive approach*. University of Chicago Press.
- Giere, R. N. (2010). An agent-based conception of models and scientific representation. *Synthese*, 172(2), 269–281.
- Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, 193(2), 559–582.
- Godfrey-Smith, P., & Sterelny, K. (2016). Biological information. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. <https://plato.stanford.edu/entries/information-biological/>
- Hesp, C., Ramstead, M., Constant, A., Badcock, P., Kirchhoff, M., & Friston, K. (2019). A multi-scale view of the emergent complexity of life: A free-energy proposal. In G. Georgiev, J. Smart, C. L. Flores Martinez, & M. Price (Eds.), *Evolution, development, and complexity: Multiscale models in complex adaptive systems* (pp. 195–227). Springer.
- Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
- Hohwy, J. (2016). The self-evidencing brain. *Noûs*, 50(2), 259–285.



- Hutto, D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. MIT Press.
- Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology*, 4(11), Article e1000209.
- Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2010). Perception and hierarchical dynamics. *Frontiers in Neuroinformatics*, 3, Article 20. <https://doi.org/10.3389/neuro.11.020.2009>
- Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, 195(6), 2387–2415. <https://doi.org/10.1007/s11229-017-1435-7>
- Kirchhoff, M. (2015). Extended cognition & the causal-constitutive fallacy: In search for a diachronic and dynamical conception of constitution. *Philosophy and Phenomenological Research*, 90(2), 320–360.
- Kirchhoff, M. (2018). Predictive processing, perceiving and imagining: Is to perceive to imagine, or something close to it? *Philosophical Studies*, 175(3), 751–767.
- Kirchhoff, M. D., & Kiverstein, J. (2019). *Extended consciousness and predictive processing: A third wave view*. Routledge.
- 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, Article 20170792. <http://doi.org/10.1098/rsif.2017.0792>
- Kirchhoff, M., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations*, 21, 264–281.
- Korbak, T. (2019). Computational enactivism under the free energy principle. *Synthese*. Advance online publication. <https://doi.org/10.1007/s11229-019-02243-4>
- Longo, G., Montévil, M., & Kauffman, S. (2012). No entailing laws, but enablement in the evolution of the biosphere. In *Proceedings of the 14th international conference on genetic and evolutionary computation conference companion* (pp. 1379–1392). <https://dl.acm.org/doi/pdf/10.1145/2330784.2330946>
- Myin, E. (2016). Perception as something we do. *Journal of Consciousness Studies*, 23(5–6), 80–104.
- O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences*, 24, 939–1031.
- Ramstead, M., Badcock, P., & Friston, K. (2018). Answering Schrödinger's question: A free energy formulation. *Physics of Life Reviews*, 24, 1–16.
- Ramstead, M., Badcock, P., & 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.
- Ramstead, M. J. D., Kirchhoff, M. D., & Friston, K. J. (2019). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*. Advance online publication. <https://doi.org/10.1177/1059712319862774>
- Ramstead, M., Veissière, S., & Kirmayer, L. (2016). Cultural affordances: Scaffolding local worlds through shared intentionality and regimes of attention. *Frontiers in Psychology*, 7, Article 1090. <https://doi.org/10.3389/fpsyg.2016.01090>
- Rescorla, M. (2016). Bayesian sensorimotor psychology. *Mind & Language*, 31(1), 3–36.
- Suarez, M. (2003). Scientific representation: Against similarity and isomorphism. *International Studies in the Philosophy of Science*, 17(3), 225–244.
- Tonneau, F. (2012). Metaphor and truth: A review of representation reconsidered by W. M. Ramsey. *Behavior and Philosophy*, 39(40), 331–343.
- van Fraassen, B. C. (1980). *The scientific image*. Clarendon Press.
- van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford University Press.
- Wiese, W. (2017). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 16(4), 715–736.

## About the Author



**Thomas van Es** is a doctoral researcher at the University of Antwerp under supervision of Erik Myin. He has published on non-representational approaches to the Embedded View of vision and predictive processing. Financed by an FWO grant, he studies Prediction Error Minimization theories, in particular the use of representational contents, and their intersection with Enactivism.