Free-Energy Principle, Computationalism and Realism: a Tragedy

Authors

Thomas van Es 1

Inês Hipólito 2, 3

1 Centre for Philosophical Psychology, Department of Philosophy, University of Antwerp (Belgium)

2 Berlin School of Mind and Brain, Humboldt University (Germany)

3 Wellcome Centre for Human Neuroimaging, University College London (United Kingdom)

Abstract

The free energy principle provides an increasingly popular framework to biology and cognitive science. However, it remains disputed whether its 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). We analyse the options critically, with particular attention to the question of representationalism. We argue that realism is unwarranted and conceptually incoherent. Conversely, instrumentalism is safer whilst remaining explanatorily powerful. Moreover, we show that the representationalism debate loses relevance in an instrumentalist reading. Finally, these findings could be generalized for our interpretation of models in cognitive science more generally.

Keywords: representationalism, realism, Free-Energy Principle (FEP), scientific models.

Acknowledgements

We are grateful to Manuel Baltieri, and Jo Bervoets for helpful comments and discussions that contributed to our work on this paper. This work has been funded by the International Postgraduate Scholarship by the University of Wollongong and by the Postdoctoral Fellowship by Humboldt University (IH); and the Research Fund Flanders (FWO), grant number 1124818N.

1 Introduction

The free-energy principle (FEP) provides a theoretical framework that primarily aims to unify biological and cognitive science approaches to life and mind (Friston, 2013). Yet it also has ambitions to underwrite classical and quantum mechanics so as to become a theory of every 'thing' (Friston, 2019).[1] In essence, it serves as a mathematical description of life, and states that any living, non-equilibrium steady-state system can be associated with the minimization of free energy.[2] Moreover, a state space description of an organism can be associated with what is called a *generative model* to the extent that a generative model is a joint distribution over (hidden) states and observation (statistical observation). We will explain the specifics of the FEP in more detail below in Section 2.

For now, it is important to note that the generative model has played a key role in certain process theories associated with the FEP, such as predictive coding and processing (Hohwy, 2013; Clark, 2016) and active inference (Ramstead, Friston, Hipólito 2020; Friston et al. 2020; Tschantz et al. 2020; Parr, 2020), yet its status remains unclear.[3] According to predictive processing theories, the generative model is literally implemented by a human brain to calculate the potential states of the environment (termed a *realist* approach), whereas other approaches take it to be an insightful statistical description that a non-scientifically trained organism has no access to (termed an *instrumentalist* approach). There is another debate as to whether this model is representational in nature or not (Gładziejewski, 2016; Gładziejewski and Miłkowski, 2017; Kiefer and Hohwy, 2018; Bruineberg et al., 2016; Kirchhoff and Robertson, 2018). The representations debate is associated with the general debate in the cognitive sciences regarding the concept of representation as a useful posit (Ramsey, 2007; Hutto and Myin, 2013, 2017). The idea is that going non-representationalist may save the generative model's causal efficacy, albeit technically *via* the generative process (Ramstead et al., 2019). In this paper, we shall engage with both debates, and argue that the representationalism debate is not relevant to the FEP. Realism is doomed to fail regardless of whether it is representationalist or not, and, conversely, instrumentalism can thrive either way, or so we shall argue.

Neuroimaging techniques offer important insights into the nervous system, such that we can develop explanations from patterns of activity and/or neuronal structures. However, patterned

[1] Every 'thing' as a system that can be modelled at non-equilibrium steady-state (NESS), such as typhoons, electrical circuits, stars, galaxies, and so on. NESS is a physical term that denotes any system that is far from equilibrium, and in a steady state with its environment. We elaborate on this in the next section.

[2] Our paper is neutral on the unifying ambitions of the FEP, and this discussion outside of the scope of this paper. Furthermore, there is an opposite proposal that focuses on entropy maximization instead (Vitas and Dobovišek 2019; Matyushnev and Seleznev 2006; Ziegler 1963).Yet these discussions are outside the scope of this paper.

[3] We discuss the relation between the FEP and its associated process theories in Section 2.

activity will not answer the question of whether or not it is representational. Indeed, experimental, neuroimaging data per se cannot answer the question of whether in its activity and interactions, the brain represents anything. This would be analogous to conducting experimental research to know whether or not objects represent the law by which they fall. The answer for ontological questions is not in empirical experiments. Thinking that the nervous system *represents* by the same properties as those we use to explain thus means taking a philosophical standpoint. To do that, we need to offer a sound philosophical argument. Thinking that it does, or does not, is a philosophical standpoint.

Inheriting from debates in philosophy of science around instrumentalism *v*s. realism, analytic philosophy of mind debates whether or not mental activity should be conceived of as representational. Scientific realism would prescribe that the technical terms used in modelling a target system also exist in the target system. Realism about Bayesian inference would thus dictate that the activity in the nervous system entails or is an intellectual representation that results from calculus between *posteriors*, *likelihoods* and *priors* (Rescorla, 2016). Instrumentalist thinking would be sceptical to accept the metaphysical assumption that the nervous system employs any of the tools used by scientists to model its activity. For instrumentalists, our capability to model, say, the auditory system, with prediction formalisms such as Bayesian inference, does not imply that the auditory system itself operates by applying Bayesian inference.

The aim of this paper is to show that, philosophically, instrumentalist thinking is less controversial, yet remains explanatorily powerful and can yield important insights in organism-environment dynamics. An instrumentalist attitude about the FEP is a safer bet without losing the potentially high returns. After briefly describing the FEP in Section 2, we assess two proposals made in the realist logical space, that of Representationalist Realism (RR), and Non-Representationalist Realism (NRR) in Section 3. We reject both of them and in Section 4 we proceed to offer positive reasons to embrace instrumentalism about the FEP. Given the activity-dependence feature of neuronal activity, Dynamic Causal Modeling (DCM), under the FEP, seems to be the most suitable and promising set of instruments to preserve the character of neuronal activation as we empirically know it to be – activity in coupled systems. From this angle, realist arguments look like forcing the world to conform with the anthropomorphic instrumental lens we use to make sense of it.

2 Free Energy Principle: essentials

The FEP is a mathematical formulation that states that a self-organising system is a dynamical system that minimises its free-energy. The FEP is based on three aspects. First, the observation of *self-organisation*, which refers to our observation of patterns, in time and space from interacting components, plays a crucial role in life sciences (Wedlich-Söldner and Betz, 2018; Hipólito 2019; Levin 2020; Fields and Levin 2020). A self-organised system can be described in terms of the structured dynamics of its behaviour. These patterns can be thought of by the light of density dynamics. That is, the evolution of probability density distributions over ensembled states (known as *variational Bayes*). A self-organising system is a system that, far from equilibrium, is in a steady state with its environment, or in non-equilibrium steady-state (NESS). To be in a *steady state* is to be in one specific state, typically averaged out over time. 'NESS' thus implies environmental exchange to maintain steady states. As such, living systems are considered to be at NESS, because their exchanges with the environment allow them to maintain their physical and structural integrity (considered their 'steady' state). Of course, living systems are in constant flux and thus are only *by approximation* in NESS. This brings us to the important feature doing the explanatory labour: entropy. Entropy, as measure of how things are, where low entropy indicates maintenance of integrity (states concentrated in small regions of the state space), and high entropy, its dissipation (states dissipated in the state space). So, the FEP focuses on entropy reduction.

This brings us to the second aspect: living organisms can be described as (stochastic) dynamical systems possessing attractors. A phase space is a space in which all possible states of a system are represented, where each possible state corresponds to one unique point in the phase space. The gain of energy translates to the expansion of the phase state. Conversely, the loss of energy formally parallels the contraction of the phase space, meaning an increase of certainty and minimisation of entropy or the maximisation of dissipation of energy in the system.

Thirdly, the states in which the self-organising system is at a point in time can be identified by the interactive role they play within the (multilevel) self-organisation scheme. States within the state space can be statistically differentiated by the application of a Markov blanket (Friston, 2020; Hipólito, Baltieri et al. 2020; Hipólito, Ramstead et al. 2020). By this formalism, we can partition the system into internal, external, active, and sensory states. Although internal and external states do not statistically influence one another (as they are conditionally independent), active and sensory states do statistically influence one another to the extent blanket states (internal, sensory, and

active) describe the patterned activity of an organism. By these lights, a system that is in NESS possesses a Markov blanket, though as a technical construct.

How can we formally account for the three aspects? The FEP prescribes the patterned activity of organisms in terms of minimisation of the free-energy as per Figure 1.

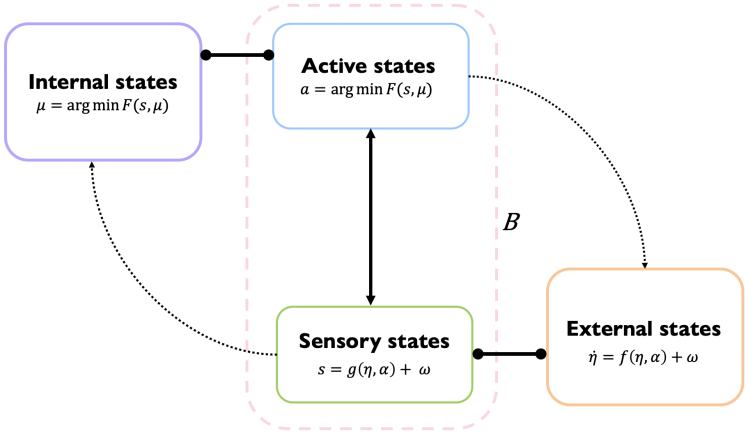

Figure 1. Minimisation of the free-energy by internal states and blanket states (*B*), comprehending sensory states and active state; and external states, which can be described by the equations of motion, as the function of (hidden states) of the world (η), active states (α), and noise or random fluctuations (ω).

Free-energy is a formal way of measuring the surprisal on sampling some data, given a generative model. Surprise refers to the "unlikeliness" of an outcome, as a measure of unlikeliness, within or in respect to a certain generative model. Mathematically, it qualifies how likely an outcome is, by measuring differences between posterior and prior beliefs of the observer. Technically, surprisal is the difference between accuracy (expected log likelihood) and complexity (i.e. Bayesian surprise or salience, as the informational divergence between the posterior probability and prior probability). Surprisal thus refers to the extent to which new data is 'surprising' to the prior model (surprisal should not be confused with psychological surprise in a day-to-day life setting or in information theory).

It is important to clarify that although the FEP is *related* to theories such as the Bayesian brain hypothesis (e.g. Knill and Richards 1996) and predictive coding (e.g. Hohwy 2013, Clark 2016), it does not entail them (at times a confusion in philosophy of mind). The FEP differs from predictive coding or the Bayesian brain hypothesis in a crucial aspect. The Bayesian brain hypothesis is the view that the brain performs inference according to Bayes's theorem, integrating new information in the light of existing models of the world. To do so, prior probability and likelihood are computed simultaneously to obtain the posterior probability. Predictive coding

(Hohwy, 2013) differs from the Bayesian brain hypothesis since it implies that prediction comes first, and is then corrected or updated by data. In this setting, two representations, bottom-down prediction, and bottom-up error signal, either match or mismatch (but see Orlandi and Lee 2018). The FEP differs from both the Bayesian brain hypothesis and predictive coding, by having at its target the reduction of entropy, rather than the maximisation of hypothesis likelihood given sensory data. The FEP does not join the discussion about the nature of computational processes (whether synchronous or sequential), because the FEP is a framework of states, not processes. The Bayesian brain hypothesis and predictive coding are process theories about how the principle is realised (Hohwy, 2020). The FEP, on the contrary, is a principle that things may or may not conform to. In this regard, the FEP, thus, stands in clear contrast with process theories such as the Bayesian brain hypothesis, predictive coding, or active inference.

The FEP is thus best seen as a research heuristic, a particular lense through which we can view and carve up the world. Associated process theories, then, are concerned with how the FEP is realized in real-world systems.[4] This crucial distinction sets us to realise that the FEP the FEP does not imply process aspects or features, such as representations, pertaining to the theoretical processes that aim to explain how the principle is realised. Yet the FEP does not in itself imply the representational tools employed by these process theories. Prior probabilities and likelihoods are tools used to explain the process by which variational free-energy is minimised. The FEP thus does not answer questions about the implementation of computational processes. Instead, the FEP targets the formal understanding of self-organising behaviour, not computational processes. It aims at explaining and understanding a system's behaviour from observing the self-organising system's patterns and making sense of them in terms of minimisation of variational free energy and entropy reduction.

3 Getting real about representations and models

The FEP provides powerful mathematical tools for the description and analysis of dynamic, self-organizing systems. However, the implications of these analyses are disputed. It is unclear what exactly they mean, what they say of the world or what we can do with them. Here we discuss the FEP along two axes, each with two possible values: 1) instrumentalism or realism, and 2)

[4] One could of course defend a predictive coding view of neural processing without subscribing to the FEP's grand ambitions, see Rao & Ballard (1999).

representationalism or non-representationalism, so that there are four possible lines of interpretation, i.e. combinations of philosophical takes on models in the FEP (see table 1).

FEP options	Realism	Instrumentalism
Representationalism	REP-REA	REP-INS
Non-representationalism	NRP-REA	NRP-INS

Table 1: Philosophical combinations under the models of FEP. Representationalist realism (REP-REA), non-representationalist realism (NRP-REA), representationalist instrumentalism (REP-INS), and non-representationalist instrumentalism (NRP-INS).

Realism and instrumentalism, here, concern the models and statistical manipulations that make up the FEP, and whether they are thought to be used and manipulated by the systems under scrutiny, independent of scientific inquiry (REA), or, conversely, whether they are thought to be scientific tools, wrought by humans in specific sociocultural environments to study particular systems (INS). Representation is a famously contested term in (philosophy of) cognitive science. Here, we shall use it to refer to at least something with representational content. That is, anything that represents some target system, does so in a way that the target system may not be so (Travis, 2004). This implies that representational content minimally has two aspects: 1) directedness, and 2) truth, accuracy or correctness conditions. First, we shall discuss the realist types: REP-REA and NRP-REA, before turning to the instrumentalist approach in Section 4.

3.1 Representationalist realism doesn't work

REP-REA is the view that the models and statistical calculations we use in the FEP formalism are literally employed by either a brain or an organism in its navigation of the world.

Prime examples of the REP-REA view come in the form of process-theoretic offshoots of the FEP, such as predictive coding, predictive processing, or, more generally, PEM theories of cognition (see for accessible introductory texts Hohwy, 2013; Clark, 2016). By employing Bayesian epistemology, scientists refer to the model of the nervous system by using technical terms pertaining to the Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Such that, the *posterior*, as the probability of "A" being true, given that "B" is true is equal to the *likelihood*, as the probability of "B" being true given that "A" is true, times the *prior*, as the probability of "A" being true, divided by the probability of "B" being true.[5]

A realist description of the model in Bayesian technical terms involves the assumption that the activity of the nervous system itself entails the representational properties of the model. That is, the activity of the nervous system aims at an intellectual representation that results from combining *posteriors*, *likelihoods*, and *priors*. The use of this technical wording is so customary, that some scientists apply it interchangeably for the brain and the model of the brain. This is the case of Pouget and colleagues (2013) stating that "there is strong behavioural and physiological evidence that the brain both *represents probability distributions* and *performs probabilistic inference*" (p. 1170, emphasis added). This causes many philosophers to take for granted that the technical terminology used in models of the brain is applicable to the brain itself, helping to paint a picture of the brain as an "inference machine" or the "Bayesian brain hypothesis". (Helmholtz 1860/1962, Gregory 1980; Dayan et al. 1995; Friston 2012; Hohwy 2013; Clark 2016). According to these views, the agent (sometimes considered to be the brain, sometimes the organism, see Hohwy, 2016; Corcoran et al., 2020) is essentially a prediction machine. Subpersonally — that is, unbeknownst to the acting individual — the system predicts, in accordance with Bayes' theorem, what is most likely to occur next, given the current state of affairs and its knowledge of the world. According to PEM's best bet, this knowledge is thought to be of the causal-probabilistic structure of the world, and stored in a 'structural-representational' format (Gładziejewski, 2016; Gładziejewski and Miłkowski, 2017; Kiefer and Hohwy, 2018). As we will see, to get explanatory bite out of the brain as (literally) an inference machine, philosophers on the realist bench need to use the full force of the technical terms employed in the model. Terms such as *posteriors*, *likelihoods*, and *priors*, then directly refer to the nervous system, both in representational (Kiefer and Hohwy 2019; Clark, 2016; Hohwy 2018), and non-representational, or seemingly enactivism inspired proposals (Kirchhoff, 2018; Hohwy, 2018; Ramstead et al. 2019; Hohwy 2020).

At issue in the debate is whether the intellectual process that we apply by using scientific tools in the investigation of a target phenomenon needs necessarily to be supposed as an ontological feature of the target phenomenon. Indeed, Linson and colleagues (2018) attentively

[5] See how Baltieri & Buckley (2017) address a similar issue.

note that as convenient as it may be, expressions such as "the brain "is" Bayesian or "implements" Bayesian models can lend itself to misunderstanding cognition's ontological commitments" (p. 14). Although representational language is often used by scientists, it remains to be seen whether this is explanatorily additive or a mere gloss (Cheremo, 2009). The proposed structural-representational format is intended to address these worries, yet, as van Es and Myin (2020) argue, does not seem to solve the well-known problems of invoking representation in cognitive science.

Consider Ramsey's (2007) job description challenge: a representation must minimally fulfil its explanatory role qua its representational status (Ramsey, 2007). A representation as used in cognitive science, it is thought, must fit the job description of what representations *do*. That is, the representation is to be explanatorily powerful *in virtue of being* a representation (and not, say, a covariation relation with an inoperative representational gloss). Further, a representational relation is a three-part relation: 1) a target system, which is represented by 2) a representational vehicle, which in turn is used as such by 3) a representation-user with access to (1) and (2) and the representationally exploitable relations between the two (Nunn, 1909-1910, as cited in Tonneau, 2012). This is closely related to the mereological fallacy (Bennett and Hacker, 2003). In slogan-form, this says that 'brains don't use models or representations, agents do.'

With regards to the two latter points, most of the REP-REA work relies on the philosophical assumption of the organism or the brain as the representation-user with access to the target system and the representational vehicle. Indeed, this assumption permeates much of the technical work and philosophical thinking of the nervous system, made popular as an analogy between the brain and scientist (Helmholtz 1860/1962, Gregory 1980; Hohwy 2013; Clark 2016; Yon, Lange and Press 2019). The longstanding Helmholtzian view (1860/1962), that 'unconscious inferences' are much like the inferences scientists draw, is also supported by Clark (2016):

> the experimenter is here in roughly the position of the biological brain itself. Her task – made possible by the powerful mathematical and statistical tools – is to take patterns of neural activation and, on that basis alone, infer properties of the stimulus (p. 95).

One must be on guard in this respect. Organisms or brains, unlike scientists[6], do not possess a perspective of an external observer. Thinking that it does, requires a sound argument that is not

[6] See Bruineberg, Kiverstein and Rietveld (2018) for a similar approach.

yet offered in the literature. In fact, the hypothesis for the brain as an ideal observer has been often rejected in literature, most recently (Brette 2019; Hipólito et al. 2020) as a bad metaphor (Mirski and Bickhard 2019; Reeke 2019). In agreement with this, Mirski et al. (2020) call for encultured minds in replacement of error reduction minds. Finally, there is what Hutto and Myin (2013, 2017) have termed the hard problem of content. This, very essentially, is the problem with grounding representational content in a naturalistic manner.

It may be fruitful to briefly rehearse here the attempt to meet the job description challenge, and why it fails (Gładziejewski, 2016; van Es and Myin, 2020). Essentially, the positive account is that structural representations as used in, for example, predictive coding, can meet the job description challenge, and solve the hard problem of content (Gładziejewski, 2016; Gładziejewski and Miłkowski, 2017; Kiefer and Hohwy, 2018). Gładziejewski uses the *compare-to-prototype* strategy, which he borrows from Ramsey (2007). He, first, analyzes a prototypical representation — in this case, a cartographic map — and distinguishes 4 features that make a cartographic map the representational tool that it is, and, second, argues step by step how each feature is present in the predictive coding account of structural representation. Cartographic maps, he argues, are structural representations of the terrains they represent, because they (1) exhibit structural similarities to the target, "(2) guide the actions of their users, (3) do so in a detachable way, and (4) allow their users to detect representational errors" (Gładziejewski, 2016, p. 566).

A counterexample shows why this analysis fails at capturing what makes a cartographic map a representation in the first place. Specifically, van Es and Myin (2020) show that a cardboard box and a table top could meet the four conditions, without engaging in any sort of representation whatsoever. Say that you're walking, holding a cardboard box that you hope to place on the table top when home. The cardboard box is structurally similar to the table top at least in terms of its relatively flat surfaces. Indeed, structural similarity comes cheap, so condition (1) is met. This structural similarity is *exploitable*, as Gładziejewski stresses (see also Shea, 2007), and can be used to guide actions of the user. In this case, the structural similarities between the cardboard box and the tabletop can be exploited so that the box can be successfully placed on the tabletop. Moreover, the structural similarities are a *fuel to success* so that, counterfactually, had they not been in place, the actions would be unsuccessful (Gładziejewski and Miłkowski, 2017). Had the cardboard box's surfaces not been similarly flat, but instead convex, the box would not afford to be placed on the tabletop stably, and the box may have ended up falling off instead. As such, condition (2) is met. *Detachability* requires that the exploitable structural similarities are exploitable in some way in the absence of the target system. The map can be used at home to plan a trip, what turns to take where, in the absence of the terrain it represents, for example. Similarly, the exploitable structural

similarities between the cardboard box and the tabletop can be used to plan where to place the box, whilst walking home and thus in absence of the tabletop. This means that condition (3) is met. Finally, *error detection* requires that the agent can detect when the supposed representation has erred in representing the target system. A cartographic map may be dated and not properly represent the current state of a city's roads. It is then the manner in which the structural similarities between the map and the terrain *do not* hold that *fuels* the failure of the navigational activity, which can be detected by way of feedback from the real world: you may not be able to take a turn that, following the map, you need to take to arrive at your destination. Returning to our cardboard box, a misalignment in the relations between the surface shape of the respective items, say, if the table top is convex or tilted and slippery, will result in a falling box. Surely, we can detect the fall by way of feedback from the real world. We may see it, it may fall on our feet, it may make a noise, etc. As such, the job description challenge was set up mistakenly, so van Es and Myin (2020) argue. It is not these four features that (conjointly) make a cartographic map a representation — lest a cardboard box represents a tabletop for the same reason. As such, REP-REA's best attempt at standing up to the job description challenge cannot get off the ground.

This also places pressure on the extent to which the other issues can be deemed resolved. The aforementioned brain as a representation-user problems remains unsolved. Further, the hard problem of content is about correctness or veridicality conditions, yet the 'error' detection minimally required here is met by a misaligned surface relation between a cardboard box and a tabletop, which falls short of *representational* error detection. After all, we may *fail* at doing something, without this being *representational* in nature.

3.2 Non-representationalist realism doesn't work

Acknowledging the deeply rooted issues with representations, there is a strand of FEP that advocates a non-representationalist approach. Though it is not always clear whether specific accounts are to be placed in instrumentalist or realist camps (see van Es, 2020 for discussion), we shall discuss here a realist interpretation of the relevant literature, and explain exactly why it cannot work. We associate the NRP-REA literature with the slogan that the *brain* does not *have* a model, the *organism is* a model (Friston, 2013; Ramstead et al., 2019; Hesp et al., 2019). Here we will take 'to be a model' to mean that, essentially, the organism is, embodies or instantiates a model relative to its phenotype, the type of organism that it is, independently of our human, sociocultural modelling practices. This is to say that the model *really* exists, and is actively being used, manipulated or 'leveraged' by any and all self-organizing systems to minimize their free energy.

Kirchhoff and Robertson (2018) argue that the FEP falls short of ascribing representational models to acting organisms. Their target is Kiefer and Hohwy's (2018) notion that the agent has a measure of misrepresentation in terms of the KL-divergence between the prior probability distribution and the posterior probability distribution. The prior probability distribution here refers to the state *before* encountering new evidence, and the posterior probability distribution here refers to the state *after* encountering new evidence. Upon encountering new evidence, the model's probability distribution will be updated to reflect the newfound evidence and how it affects the different aspects of the model. In a sense, then, the difference between the prior and the posterior will be a measure of the extent to which the model has been changed in the updating process. This is then, for the system, a measure for the extent to which its initial model was misaligned. Further, *if* we take the generative model to be representational in nature, the KL-divergence becomes a measure of the extent to which the system *misrepresented*. Yet, Kirchhoff and Robertson point out that the model comparison in the KL-divergence only measures Shannon covariance, not representation (2018). Barring a representational assumption, this means, they suggest, that this falls short of providing a measure of misrepresentation, and only succeeds in providing a measure of covariational misalignment (Bruineberg and Rietveld, 2014; Bruineberg et al., 2019; Kirchhoff and Robertson, 2018). As such, what actually does explanatory work in FEP is the minimization of negative covariance, *not* the minimization of (representational) prediction error.

If we cannot invoke representations in our realist account of the FEP machinery, what does this leave us with? Key terms in the FEP conceptual toolkit are the *generative model*, a probability distribution over sensory states parameterized by the internal states, the *generative process* by which it is placed into contact with external states via active states, and Bayesian updating of the model (Corcoran et al., 2020; Ramstead et al., 2019; Friston, 2013). Without representations, one may wonder, can we still have a generative model? For this, we need to briefly explore what it takes for anything to be called a model. In the current discourse in philosophy of science, there is a wide variety of accounts with regard to what makes a model, and how it is that they can tell us anything about their target systems.

There are many varieties of models in use in a scientific context, such as scale models, analogous models, idealized models and more. Standardly conceived, each of these is a model of its target in virtue of *representing* that target (Frigg and Hartman, 2020). Bayes's theorem itself is a placeholder, that when furnished with relevant information becomes a model that affords predicting the activity or behavior of the target system given certain conditions. Minimally, a model such as Bayes' theorem, is required to have three features: (1) access (furnishing information or

data); (2) a target (neuronal activity); (3) structural similarity (similar causal relations). If we would create a Bayesian model of, say, neuronal activity, then access is accounted for by the furnishing information or data, the target is neuronal activity and structural similarity needs to hold by way of a dynamical covariational relation so that if *X* would wiggle in the registered neuronal activity, something needs to wiggle in the model as well.

If we take a model to be essentially representational, it should be clear, a mere covariation relation is insufficient to warrant model status. Moreover, Bayesian inference can be seen as the implementation of Bayes' theorem on a specific Bayesian model in light of new evidence. In light of this, it seems that without representation, the *generative model*'s status as a model needs to be revoked. This presents a serious problem to those defending NRP-REA (such as Kirchhoff et al., 2018; Ramstead et al., 2019; Bruineberg et al., 2016).

Let us consider this more closely. A generative model is a probabilistic mapping of potential external states relative to the internal states. If we have a multi-dimensional state space that describes a particular system's internal states, with an axis for each variable associated with the system, then the generative model is what tells us, given this state space description, the probability of the possible values for each variable of the external states. This can be extremely useful because each of those variables represents one or some behavioural features of the target system it is a description of. A description, of course, is a form of representation. If we are to take away the representational characteristics of the generative model, the variables over which it is a probability distribution do not actually represent anything at all[7]. It would be a probability distribution over variables that in no way stand in or are to be seen as surrogates for real-world characteristics or features.[8] It thus seems that, without representation, the generative model is no model at all, and thereby unfit to aid an agent in navigating its environment, as NRP-REA would have it.

NRP-REA seems like a no-starter. Unless, there would be a way of making sense of surrogacy, of something *standing in* for something else *without* reference to representation or representational content. Luck has it that Guilherme Sanchez de Oliveira challenges the representational capacity and motivation even for *scientific models*. *Prima facie*, this seems like the exact way out NRP-REA's generative models require: modeling without representation (2018, 2016). Yet de Oliveira's work may not be the hero they need. He argues that scientific modelling isolates a model from its context, and in doing so, "constrains our ability to see how the nature of

[7] We discuss one strand of definitions of 'model': the representational one. Nonetheless, this makes up the bulk of the philosophical literature on the efficacy and ontological status of models. Below we discuss the only outlier position.

[8] This point can be understood in Wittgenstein's understanding of 'nonsensical' propositions, where variables would be radically devoid of meaning, that is to say, transcend the bounds of sense. If we remove the representational characteristics of the generative model, the variables over which it is a probability distribution do not have any referent, i.e. are 'nonsensical' propositions.

the phenomenon is shaped by what brings it about (the individual scientists, the research context, disciplinary traditions, and technological possibilities in addition to properties of the target)" (de Oliveira 2016, p. 96). The challenge to representational properties is that "we get caught up on ethereal metaphysical concerns that have nothing to do with the phenomenon in the real world of scientific practice" (de Oliveira 2016, p. 96). Moreover, elsewhere de Oliveira argues that to judge models' epistemic virtue and their ontological status in terms of their representational relation to a target system is contradictory (2018). In a nutshell, if a model is inherently representational, and if modelling is thus a representational activity, this means that:

> scientists can use models to learn about target phenomena because models represent their targets, *and* that models represent their targets because scientists use them as representations of those targets—in short, this would mean that the reason scientists *can* use models to study real-world phenomena is that they *do* use them to study real-world phenomena. (de Oliveira, 2018, p. 14, emphasis in original)[9]

Vicious circularity ensues. Essentially, the *use* of models is justified in terms of their representational status, yet the representational status itself is grounded in our use thereof. Whether de Oliveira is correct in this analysis of modelling practices in science is irrelevant to our current debate, yet his proposed alternative *is* relevant.

A non-representational approach to models as de Oliveira (2018) suggests, keeps only what is essential to our modelling practices. There are at least two features we can distinguish: a model is 1) mediative or surrogative in that it mediates between the modeller and the target system or *stands in* (or surrogates) for the target system to the modeller, and 2) requires training in specific, socioculturally embedded modelling practices (de Oliveira, 2018). He further notes that mediation nor surrogacy are necessarily representational in nature: consider our use of toy guitars or miniature-sized footballs as surrogates for their professional counterparts. These surrogates, further, aid in the 'skill-development and learning transfer' practices. As it is for the toy guitar, so it is for the model, de Oliveira argues (2018). Indeed, scientific models result mostly from procedures and processes of negotiations materially extended across laboratories in the world, and, thereby, across cultures, and from experts to students. We use models to learn about complex systems, and use this knowledge in our manipulation of the target systems indirectly by, for example, informing policy makers. As such, in our scientific endeavours, models can be naturalistic and useful, whilst only counting as representational when embedded, manipulated, and viewed within the appropriate sociocultural practice (Hutto and Myin, 2013, 2017). Here too, it is important to note that these models are devised, employed and explored by agents, not by their

[9] This argument is directed at 'mind-dependent' views of the representational relation of models, according to which, our *use* of models *as* representations is crucial to their representational status. See de Oliveira (2018, p. 9-12) for a discussion of mind-independent view that has long gone out of fashion.

brains. Conversely, a scientific model outside of social practices is simply a device with contingent relations to another system that in itself is senseless (in the Fregean sense) and, thereby, holds no explanatory capacity.

Now that we have sketched de Oliveira's motivation for an account of a non-representational approach to models, we need to see whether this helps NRP-REA's predicament. It essentially comes down to whether FEP's generative models display both (1) the mediative or surrogative, and (2) the skill development and learning transfer features of models, if they are ascribed to (or used, implemented, instantiated, or leveraged by) any free energy minimizing system. Feature (1) is easily shown, as the generative model (but more specifically the generative process by which the model is brought into contact with the external world) is considered crucial in determining action policies (Ramstead et al., 2019). Feature 2, however, is, as emphasized above, clearly sociocultural in nature. It is by becoming enculturated in a scientific ecosystem, being trained by experts in the practice, that we attain the relevant sensibilities with regards to construing and manipulating a model, as well as how to leverage it to further our understanding of the target system. The generative model, in NRP-REA, is to be used in some way or another by *any* free energy minimizing system, unconsciously. That is to say that the way the generative model is envisioned to be leveraged by an organism does not take into consideration the practice that a trainee would need to undergo to become skilled at using complex, statistical models.

In sum, regardless of whether a model is to be seen as a representational device or not, the generative model, if it is to be given a realist reading as is done in NRP-REA, *cannot* reasonably be said to be a model taking into consideration the current state of the literature on the ontological and epistemic status of models. In this section we have first discussed the KL-divergence option if we take models to be essentially representational. We have argued that for a model to actually be about something, refer to something, i.e. it needs to be representational (under this notion), yet the KL-divergence approach resists this. Without this, the probability distribution we apply Bayes' rule to can't actually get off the ground. We have also discussed de Oliveira's option that models are not representational. Yet here too Bayesian inference doesn't hold without learning transfer, professional training, and so on. This means that Bayesian inference will not get off the ground. After all, Bayesian inference is a particular mode of manipulation of a Bayesian model of a target system. These manipulations are performed by agents embedded and trained in sociocultural modelling practices unavailable to the NRP-REA theorist's notion of a generative model as leveraged by an organism. Though we have provided a potential way out for NRP-REA by considering an account of modeling that does not rely on representation, it turned out to be a dead end. This means that NRP-REA, despite carefully avoiding the well-known representationalist's

pitfalls, is incapable of balancing themselves on the Bayesian-enactivist tightrope. If neither representationalism nor non-representationalism can make a realist interpretation take off, we may need to consider whether instrumentalism fares any better, and if so, what good it actually does to go instrumentalist. Why not give up on the FEP project in its entirety if the models it describes are not literally employed by the organisms we study?

4 Why instrumentalism works

Instrumentalism is, broadly, the idea that the models we use to describe important and interesting statistical relations between, among others, organisms and their environments do just that: describe. It resists the temptation to conceive of organisms as having access to our human sociocultural heritage of making and exploiting models.[10] As such, instrumentalism in itself is characterized by ontological agnosticism with regards to what *actually* makes a system tick. Instead, it is concerned with accurately describing organism-environment dynamics and the interesting relations that may surface.

In this section, we want to first explicate why instrumentalism does not run into any of the issues that realism does. Of interest here is the way in which the question of representationalism transforms from being of vital interest to the FEP project to an interesting related question that helps conceptualize the framework. Second, we want to delve into how instrumentalism can work for us. This is important to emphasize, because otherwise it may seem like we are only losing explanatory ambitions, without gaining anything in return.

4.1 The representational collapse and the safer bet

In Section 3, we raised a few concerns with the realist perspective on the FEP. The realist perspective we considered as the model of the FEP, and thus Bayesian inference, is in one way or another literally used, employed, instantiated, embodied, or 'leveraged' by any free energy minimizing system, or at least organisms. We argued that the position is untenable, regardless of whether we take a representationalist or a non-representationalist stance. Bayesian inference using a statistical model of a target system is commonly seen as a representational activity, yet there is no naturalistically viable answer as to how this works outside of our own socioculturally developed representational practices as scientists or philosophers, as we discussed in Section 3.1.

[10] Humans, of course, do have access to our sociocultural heritage. *Prima facie*, this one might consider a 'humans-only' approach. However, the use of models does not, by way of sociocultural heritage, become *innate* (Satne and Hutto, 2015). We have been exposed to imagery all around us, exponentially so the younger you are, which influences our skillset. Though enculturated in a wide variety of representational practices, the particular skill of employing Bayesian inference remains rather *niche*, making it a tough sell for universality. The distinction between activity being *conform* a computational principle and actually *computing* according to this principle is relevant, but is outside the scope of the current paper.

Subsequently, in Section 3.2 we find that the non-representational approach and its covariational escape has its explanatory concepts fall one by one. When it concerns essential organismic behavior, we show that without representational content, there is no model; without a model, there is no Bayesian inference; without Bayesian inference, there is no realism. As such, realism is untenable across the board. Yet instrumentalism is not *evidently* free of worries.

Here we shall briefly discuss the issues encountered by realism, why they don't concern the instrumentalist approach, and also why, in general, instrumentalism is a much *safer* bet. We use the same methods, the same conceptual tools, but how they are employed differs wildly. In an instrumentalist perspective, Bayesian inference, as well as any potentially associated representational activity, is not said to be performed, embodied or leveraged by any system other than those humans that have been trained to do so. The same applies to models, and modeling activities, but of course also to any other sociocultural activity such as writing, whether that is formally, calligraphic or graffiti. In the instrumentalist take, organisms do not model anything in and of themselves, but they could potentially be trained by others to engage in certain modeling practices that aim to explain and predict scientific phenomena. FEP models, as well as the inferences we make with them about their target systems, are specific to our human scientific practices of studying the world by way of using idealized surrogates. Where these models originate in, and how they can serve as tools that help us understand the world, then, becomes a question for the history and philosophy of science, *not* for the cognitive sciences.

We see a similar transformation of the issue of representationalism. In the introduction of Section 3, we sketched the possibilities along the two axes of interest: realism *vs* instrumentalism, and representationalism *vs* non-representationalism, leading to four positions: REP-REA, NRP-REA, REP-INS and NRP-INS respectively. For the realist position, REP-REA and NRP-REA are extremely different accounts of how living and cognitive systems navigate their environment. Either the system forms a rich, representational model of (the causal probabilistic structure of) the external world (Hohwy, 2013; Gładziejewski, 2016), or the system covaries adaptively with its environment by 'leveraging' a stipulated generative model (Ramstead et al., 2019; Kirchhoff and Robertson, 2018). Notice, however, that *qua* the FEP, cognitive science, and biology, the representationalism question enters the domain of philosophy of science. Indeed, if we go instrumentalist, as far as our scientific endeavour is concerned, *it doesn't actually matter* whether the models we use are representational or not, it just matters *that they work* (de Oliveira, 2018, pp. 18-20). As such, instrumentalism does not solve the issues of realism, rather, the issues do not even apply to instrumentalism. In fact, they are *dissolved*.

This is particularly interesting when we consider the non-representationalist view presented in the literature (Kirchhoff and Robertson, 2018; Ramstead et al., 2019). In Section 3.1, we placed this view in the NRP-REA camp for argumentative purposes, conceding that the literature itself can technically be read in multiple ways (van Es, 2020). Yet we can see now that arguing for a non-representational take on the models as employed in the FEP, only makes sense under realist assumptions. Only if we *assume* the entire statistical machinery at work in the FEP is literally employed by free energy minimizing systems, does it really matter whether these models imply representationalism (and the problems this is accompanied by) or not. Consequently, this puts the enactivism-inspired 'no representation, just covariation' project in the FEP literature in a bind. It is either doomed to fail (under realist assumptions) or irrelevant (under instrumentalist assumptions).

At this juncture, one may either deem instrumentalism the god-given gift without philosophical problems, *or* suspect that there is something deeply worrying about it. Or a bit of both, we don't judge. Yet it's exactly this *lack of judgment* that may seem suspect. The realist took a plunge, and, or so we argue, failed. They took a risk and came up empty. Yet it may seem the instrumentalist just waited by the sideline, and only remained safely untouched because they never moved in the first place. That is, it may seem the instrumentalist is only safe from issues because *it doesn't actually make any claims* about the world. It may seem empirically vacuous, without even the promise of helping us understand the world and its distinguishable systems any better. In the remainder, we shall argue that despite giving up the realist claims on the world, instrumentalism in the FEP has much explanatory capacity to offer with respect to new insights in making sense of systems' interactions in terms of patterned activity.

4.2 The stakes of instrumentalism or models in neuroscience

In neuroscience, we use different imaging techniques and formal languages to understand the activity of the nervous system. Formal, or mathematical languages are developed and applied to make sense of the overwhelming amount of data collected from imaging the brain, where different formalisms correspond to different models. If the model shows similar patterns of activation to those directly collected from functional neuroimaging, we can obtain not only insights into the neuronal activity itself, but also draw and test new hypotheses related to and within that model. The model, in scientific practice, is a representation of the nervous system to the extent it holds explanatory capacity. This is, as we know, the goal *par excellence* of computational neuroscience.

Indeed, computational neuroscience simulates the neuronal processes to infer models that explain and predict the phenomena. There are two major ways to model neuronal processes. One is Structural Causal Models (SCM), which typically applies machine learning or information theory to model the system in conformity with the presence or absence of 'information'. The goal in SCM is precisely to display topological maps of brain structures as per the presence or absence of 'information' amongst highly connected (neural) modules or nodes.[11] This is the set of techniques *par excellence* of brain mapping.[12] The other way of modelling neuronal processes is by Dynamical causal models (DCM), employed to explain the activity-dependent patterns found in the nervous system. Applied to the FEP is the modelling of activity-dependence in coupled systems by means of dynamical formalisms. As simulation models that aim to hold predictive capacity, both models – SCM and DCM – apply the statistical tools of Bayesian epistemology[13], viz. Bayesian inference.

4.3 How instrumentalism can work for us

The main question for the FEP is, not about processes, but self-organising behaviour. As we have explained in section 2, the FEP aims at explaining and understanding a system's behaviour from observing the self-organising system's patterns and making sense of them in terms of minimisation of variational free energy and entropy reduction.[14] As a principle, the FEP is expected to apply to different levels of self-organisation.

The behaviour of (self-organising) systems can be described as acting to minimise expected free energy, and to reduce expected surprisal. Living systems, such as cells in a tissue, neurons in a network, brains in organisms, organisms in environments and so on, enacting their environments, could be thought of as actions for epistemic affordance. By epistemic affordance we mean actions that avoid dissipation (resolve uncertainty and, thereby, expected free energy).[15] In order to avoid dissipation, opportunities for resolving uncertainty become attractive. Appealing to dynamical systems theory, this can be described as a random dynamical attractor: a dynamical system in which the equations of motion have an element of randomness or fluctuations to them. An example of a random dynamical system is a stochastic differential equation, describing and accounting for the

[11] Where the aim is to highlight the structure by determining (predicting the likelihood) of connections between modules in terms of information being exchanged between modules - thus by the presence or absence of information.
[12] See Pearl (2001); Spohn (2010); Bielczyk et al. (2019); Borsboom, Cramer Kalis (2019); Straathof et al. (2019).
[13] See (Talbot, 2016).
[14] We do not claim that FEP offers the ultimate answer to *all* behavior.Yet it may be key in making sense of certain biologically essential levels of cognition.
[15] This does not mean that propositional information is extracted from the environment.

important aspect of noise. Brown (1827), examining the forms of particles immersed in water, "observed many of them very evidently in motion". Albert Einstein (1905) noted they arose directly from the incessant random pushes, or perturbations, to the particle made by molecules in the surrounding fluid. Langevin (1908) formulated the first stochastic equation to describe Brownian motion emphasising the dynamical behaviours observed in the interplay between deterministic processes and noise.[16] Randomness or fluctuations (such as Brownian motion, or even cell or neuronal activity) are 'noisy' to the extent that their origin implies 'degrees of freedom'. Notably, noise can drastically modify the even deterministic dynamics.[17] Importantly, this means that stochastic dynamical systems, accounting for noise, are equipped, at least in principle, to capture how existing states contribute to adaptation. State-space models are among the most suitable sets of techniques (Razi and Friston 2016) to model the unfolding activity or behavior of a system subject to fluctuations and noise, described by an ordinary differential equation (ODE):

$$\dot{\mathbf{x}}(t) = f\big(\mathbf{x}(t), \boldsymbol{\theta}, \mathbf{u}(t)\big) + \mathbf{w}(t) \qquad (1$$

Where f denotes the coupled dynamical system where θcorresponds to the parameters of the influences; $\mathbf{x}(t) = [x_1(t), x_2(t),...,x_n(t)]^T$ represents the rate of change over the time in state variables $\mathbf{x}(t)$. And, finally $\mathbf{w}(t)$ represents the random influences that can only be modelled probabilistically. Although random influences play an important role in 'stochastic' systems, they are typically de-emphasised in most formal applications by being replaced or absorbed into prior distributions over parameters. Considering however the relevance of noise, e.g. to stabilize unstable equilibria and shift bifurcations[18]; motivate transitions between coexisting deterministic stable states or attractors; or even induce new stable states that have no deterministic counterpart, taking random fluctuations as priors, we think, blurs the line between dynamical and deterministic systems. Models, instead of aiming to represent something, should be able to capture the essential aspect of random influences (the patterns we alluded to above) and thus offering a more comprehensive understanding of behaviour. In a real-world scenario, systems, like cells, neurons, or organs, can be described as

[16] This is especially relevant under the observation that at the very least noise acts as a driving force exciting internal modes of oscillations in both linear and nonlinear systems (where the latter corresponds to the enhanced response of a nonlinear system to external signals, see Jung, 1993; Gammaitoni et al., 1998; Lindner et al. 2004).
[17] Even if it is possible to use deterministic equations of motion to study a system subjected to the action of a large number of variables, the deterministic equations need to be coupled to a "noise" that simply mimics the perpetual action of many variables.
[18] The parameter value at which the dynamics change qualitatively (Arnold 2003).

subject to fluctuations or noisy environments. However these systems 'act', can be understood as avoiding dissipation (or resolving uncertainty and, thereby, expected free energy).

There is no reason to think of this form of action as intellectual thinking (i.e. representing). The intellectual part of the story is our scientific or philosophical attempt to make sense and understand observed behaviour. So, we describe behavior or actions that, from an external observer standpoint, look as if the subjects of enactment were asking 'what would happen if I did that' (Schmidhuber 2010). We can develop and use tools of process theories (e.g. predictive coding, predictive processing) to explain how systems resolve uncertainty (and thereby minimise the free energy). We can use formal terminology, such as *intrinsic value*, *epistemic value of information*, *Bayesian surprisal*, and so on (Friston 2017), to develop models that explain the neuronal processes enabling and underlying things becoming salient to a system to resolve uncertainty. For example, to develop models that explain neuronal excitatory and inhibitory projections in terms of predictions and prediction errors, respectively. In this scientific route, an open question for process theories in relation to the FEP is which theory, predictive coding, Bayesian filtering, belief propagation, variational message passing, particle filtering, and so on, if any, conforms to the FEP. More precisely, which model, if any, conforms with the FEP.

Yet from the fact that it is possible to model a process, it does not necessarily follow that the target phenomenon represents the intellectual tools we use to model it. Consider a moving object that can be explained by Newton's Law of Motion. That we can model the movement by that formalism, does not follow that the object represents the law by which it falls. Few people would claim that the object *represents (or embodies, instantiates, implements, employs, leverages)* the laws by which it moves. Because science does not back this up, those who wish to do so, are committed to a philosophical assumption that moving objects, like cells, or organs like the nervous system, represent laws, principles, or the intellectual tools we use to describe processes conforming to laws or principles (*posteriors*, *likelihoods*, and *priors*). Friston, Wiese and Hobson (2020) are on the guard on this matter, pointing that, from the fact that it is possible to map states, "does not mean that the resulting descriptions refer to entities *that actually exist*" (p. 17, emphasis added).

FEP is not in itself a commitment to the picture that an organism, and/or its nervous system literally is a hierarchical system that itself aims at representation (Baltieri and Buckley 2019; Gallagher 2020; Hipólito et al. 2020; Williams 2020). This is because the FEP targets understanding behavior, from the observation of its dynamical states, in terms of self-organisation towards the aim of avoiding dissipation. In the FEP, the notion of salience plays an essential role read according

to the reduction of entropy. What becomes salient is what reduces entropy, put simply. It is these enactive and cultural aspects that are lacking from process theories aiming at developing representational models. Explaining why things become salient is an *explanandum* of the FEP. In this setting, active inference, as a process theory, is an important FEP associate tool to explore the processes enabling and underlying things becoming salient, because it accounts for salience as an attribute of the action itself in relation to the lived world. In pushing in this direction, active inference seems formally equipped to a more accurate description/model of real-world sociocultural scenarios. But active inference is a process theory, i.e. it aims at explaining *the processes by which* things become salient to an agent, not *why* they become salient - that is a goal of the FEP.

The instrumentalist account we propose here understands the use of models without the need to assume that the target system also engages in a representational activity. From the fact that we can generate a high probability value that allows us to draw claims about behaviour, from within our model of an enactive system, we are not licenced to assume that the enactive system itself represents the laws by which it adapts. Such a claim would imply a further claim: that nature, essentially, represents. This does not seem metaphysically reasonable. We think that instrumentalism associated with FEP offers sufficient explanatory power without falling into problematic realist assumptions. In what follows we explain how the FEP, as a tool, holds explanatory capacity for the investigation and understanding of organisms enacting the environment.

In Section 3 on realism, we discussed the KL-divergence argument against representational aspects of FEP (Kirchhoff and Robertson 2018). As we attempted to show in Section 4.1, this argument against representationalism only holds under a realist assumption. The KL-divergence 'solution' to the problems with representations becomes irrelevant. Indeed, only if the model is actually thought to be used, manipulated (or 'leveraged') by the organism, does it actually make sense to try and resolve representationalist worries. Yet in our instrumentalist account this is a non-issue.

In conclusion, we do not think that there are convincing reasons to believe that organisms or systems engage in representation, nor to think that our scientific models are themselves necessarily representational. Situated in a sociocultural practice, models allow us to make culturally informed inferences about the likelihood of something being the case with the target system (i.e ontological claims). So, the instrumentalism we propose does not assume that generative models used in the modelling are models that are used by organisms or systems themselves, nor that

models are representational outside of the culture they are developed in. Neutral with regards to realist ascriptions, our instrumentalist account for the offers sufficient explanatory power to explain the behaviour of systems or organisms without falling into unnecessary philosophical problems.

5 Conclusion

In this paper we have defended the instrumentalist take on the FEP, arguing that the realist approach is a non-starter, regardless of whether it is representationalist or not. Crucially, the question as to whether systems do or do not model their environment will not be decided by neuro-imaging studies or the models we employ in interpreting the data. This is a philosophical matter that should be dealt with by way of philosophical argumentation. We have argued that the representationalist realist (REP-REA) position does not hold up because of the as of yet missing naturalistic grounding of representations independent of sociocultural practices, including structural representations (van Es and Myin, 2020; Hutto and Myin, 2013, 2017). The non-representationalist realist position (NRP-REA) purports to solve the issues of REP-REA by removing representational content from the story. Yet it does not hold up because without content, there is no model and no Bayesian inference. The instrumentalist does not face the same problems, as they do not ascribe the modeling activity to the organism under scrutiny. The question of representationalism then turns into a general philosophy of science debate on the ontology of models in science, on which the validity or usefulness of the FEP does not hinge (de Oliveira, 2018). The instrumentalist position, then, means that we take the statistical machinery to be a helpful description of real life systems, potentially offering deep insights into the relevant statistical relations between organism and environment. The instrumentalist does *not* take the models we make of the organisms to be employed by the organisms themselves *in virtue* of our capacity to model them.

The difference between realism and instrumentalism is thus primarily ontological in nature: in realism, there is an ontological claim with regards to the status of models in living systems, whereas in instrumentalism there is no such ontological claim. This may be seen as a weakness, as it looks as though the instrumentalist position only gives up explanatory ambitions relative to the FEP realist. This is true. However, the ambitions given up on, we argue, are never going to be met. If this is on the right track, the realist's ambition is a fata morgana, if you will. As such, instead of chasing ghosts, the instrumentalist position is more realistic in their ambitions. There is, within

this more modest framework, still plenty of insight to be gained into the workings of life and cognition by way of dynamic causal modeling (DCM). In sum, we argue that modesty and ambition go hand in hand when it comes to models and the FEP.

Bibliography

Baltieri, M., & Buckley, C. L. (2017, September). An active inference implementation of phototaxis. In *Artificial Life Conference Proceedings 14* (pp. 36-43). One Rogers Street, Cambridge, MA 02142-1209 USA journals-info@ mit. edu: MIT Press.

Baltieri, M., & Buckley, C. L. (2019). Generative models as parsimonious descriptions of sensorimotor loops. *arXiv preprint arXiv:1904.12937*.

Brown, R. (1828). XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *The Philosophical Magazine*, 4(21), 161-173.

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., 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.

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

Chemero, A. (2009). *Radical embodied cognitive science*. MIT press.

Corcoran, A. W., Pezzula, G., and Hohwy, J. (2020) From Allostatic Agents to Counterfactual Cognisers: Active Inference, Biological Regulation, and The Origins of Cognition. *Biology and Philosophy, 35*(3). https://doi.org/10.1007/s10539-020-09746-2

Crauel, H., Flandoli, F. (1994) Attractors for random dynamical systems. *Probab. Th. Rel. Fields 100***,** 365–393. https://doi.org/10.1007/BF01193705

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. Neural computation, 7(5), 889-904.

Einstein, A. (1905). On the movement of small particles suspended in stationary liquids required by the molecularkinetic theory of heat. *Ann. d. Phys, 17*(549-560), 1.

Fields, C., & Levin, M. (2020). Scale-Free Biology: Integrating Evolutionary and Developmental Thinking. *BioEssays*, 42(8), 1900228.

Frigg, R. and Hartmann, S. (2020) Models in Science in *The Stanford Encyclopedia of Philosophy (Spring 2020 Edition)*, Zalta, E. N. (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/models-science/>.

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, *62*(2), 1230-1233.

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

Friston, K. J., Fortier, M., & Friedman, D. A. (2018). *Of woodlice and men: A Bayesian account of cognition, life and consciousness. An interview with Karl Friston*. ALIUS Bulletin, 2, 17-43.

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

Friston, K., Parr, T., Yufik, Y., Sajid, N., Price, C. J., & Holmes, E. (2020). Generative models, language and active inference. *PsyArXiv*. DOI: 10.31234/osf.io/4j2k6

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

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

Friston, K.J.; Wiese, W.; Hobson, J.A. Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. Entropy 2020, 22, 516.

Gallagher, S. (2020). Action and interaction. Oxford University Press.

Gammaitoni, L., Hänggi, P., Jung, P., & Marchesoni, F. (1998). Stochastic resonance. *Reviews of modern physics*, *70*(1), 223.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese, 193*(2), 559–582.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors. *Biology & philosophy, 32*(3), 337–355. https://doi.org/10.1007/s10539-017-9562-6

Gregory, R. L. (1980). Perceptions as hypotheses. Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 290(1038), 181-197.

Gulli, R. A. (2019). Beyond metaphors and semantics: A framework for causal inference in neuroscience. Behavioral and Brain Sciences, 42.

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.

Hipólito, I. (2019). A simple theory of every 'thing'. *Physics of life reviews, 31*, 79-85.

Hipólito, I., Baltieri, M., Friston J, K., & Ramstead, M. J. (2020). Embodied Skillful Performance: Where the Action Is. *Synthese*.

Hipólito, I., Ramstead, M., Constant, A., & Friston, K. J. (2020a). Cognition coming about: Self-organisation and free-energy: Commentary on "The growth of cognition: Free energy minimization and the embryogenesis of cortical computation" by Wright and Bourke (2020). *Physics of Life Reviews*.

Hipólito, I., Ramstead, M., Convertino, L., Bhat, A., Friston, K., & Parr, T. (2020b). Markov blankets in the brain. *arXiv preprint arXiv:2006.02741*.

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

Hohwy, J. (2018). The predictive processing hypothesis. *The Oxford handbook of 4E cognition*, 129-146.

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

Hutto, D. D., & Myin, E. (2013). *Radicalizing enactivism: Basic minds without content*. Mit Press.

Hutto, D. D., & Myin, E. (2017). *Evolving enactivism: Basic minds meet content*. MIT press.

Jung, P. (1993). Periodically driven stochastic systems. *Physics Reports*, *234*(4-5), 175-295.

Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, *195*(6), 2387-2415.

Kiefer, A., & Hohwy, J. (2019). Representation in the prediction error minimization framework. Routledge handbook to the philosophy of psychology, 2nd ed. Oxford, UK: Routledge.

Kirchhoff, M. (2018). Predictive brains and embodied, enactive cognition: an introduction to the special issue. *Synthese*.

Kirchhoff, M., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations, 21*, 264–281.

Lemons, D. S., Gythiel, A., & Langevin's, P. (1908). "Sur la théorie du mouvement brownien [On the theory of Brownian motion]". *CR Acad. Sci.(Paris)*, 146, 530-533.

Levin, M. (2020, July). Robot Cancer: what the bioelectrics of embryogenesis and regeneration can teach us about unconventional computing, cognition, and the software of life. *In Artificial Life Conference Proceedings* (pp. 5-5). One Rogers Street, Cambridge, MA 02142-1209 USA

Linson A, Clark A, Ramamoorthy S and Friston K (2018) The Active Inference Approach to Ecological Perception: General Information Dynamics for Natural and Artificial Embodied Cognition. *Front. Robot. AI 5*:21. doi: 10.3389/frobt.2018.00021

Lindner, B., Garcıa-Ojalvo, J., Neiman, A., & Schimansky-Geier, L. (2004). Effects of noise in excitable systems. *Physics reports*, *392*(6), 321-424.

Martyushev, L. M., & Seleznev, V. D. (2006). Maximum entropy production principle in physics, chemistry and biology. *Physics reports*, *426*(1), 1-45.

Mirski, R., & Bickhard, M. H. (2019). Encodingism is not just a bad metaphor. *Behavioral and Brain Sciences*, 42.

Mirski, R., Bickhard, M. H., Eck, D., & Gut, A. (2020). Encultured minds, not error reduction minds. *Behavioral and Brain Sciences*, 43.

Nunn, T. P. (1909-1910). Are secondary qualities independent of perception? *Proceedings of the Aristotelian Society, 10,* 191-218.

Orlandi, N. & Lee, G. (2018). How Radical is Predictive Processing? in Eds., Colombo, Irvine, & Stapleton, *Andy Clark & Critics*. Oxford University Press

Parr, T. (2020). Inferring What to Do (And What Not to). *Entropy, 22*(5), 536.

Pearl, J. (2001). Bayesian networks, causal inference and knowledge discovery. *UCLA Cognitive Systems Laboratory, Technical Report*.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, *16*(9), 1170–1178. https://doi.org/10.1038/nn.3495

Rao and Ballard, (1999) Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-field Effects. *Nature Neuroscience, 2*(1):79-87

Rescorla, M. (2016). Bayesian sensorimotor psychology. *Mind & Language, 31*(1), 3–36.

Ramsey, W. M. (2007). *Representation reconsidered.* Cambridge University Press.

Ramstead, M. J. D., Kirchhoff, M. D., & Friston, K. J. (2019). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior, 28*(4), 225-239.

Ramstead, M. J., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy*, *22*(8), 889.

Ramstead, M. J., Kirchhoff, M. D., Constant, A., & Friston, K. J. (2019). Multiscale integration: beyond internalism and externalism. *Synthese*, 1-30.

Razi, A., & Friston, K. J. (2016). The connected brain: causality, models, and intrinsic dynamics. IEEE Signal Processing Magazine, 33(3), 14-35.

Reeke, G. N. (2019). Not just a bad metaphor, but a little piece of a big bad metaphor. *Behavioral and Brain Sciences*, *42*.

Satne, G., & Hutto, D. (2015). The Natural Origins of Content. *Philosophia*, *43*(3), 521–536. https://doi.org/10.1007/s11406-015-9644-0

Shea, N. (2007). Consumers need information: Supplementing teleosemantics with an input condition. *Philosophy and Phenomenological Research, 75*, 404–435.

Travis, C. 2004. The silence of the senses. *Mind 113* (449): 57–94.

Tonneau, F. (2012) Metaphor and truth: A review of *Representation Reconsidered* by W. M. Ramsey, *Behavior and Philosophy, 39/40*, 331-343.

Tschantz, A., Baltieri, M., Seth, A. K., & Buckley, C. L. (2020, July). Scaling active inference. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

van Es, T. (2020). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*. https://doi.org/10.1177/1059712320918678

van Es, T. and Myin, E. (2020) Predictive processing and representation: How less can be more. In Mendonça, D., Curado, M., and Gouveia, S. S. (eds) *The philosophy and science of predictive processing*. Bloomsbury.

Vitas, M., & Dobovišek, A. (2019). Towards a general definition of life. *Origins of Life and Evolution of Biospheres*, *49*(1-2), 77-88.

von Helmholtz, H. (1962). Handbuch der physiologischen optik. 1860/1962. & Trans by JPC Southall Dover English Edition.

Wedlich-Söldner, R., & Betz, T. (2018). Self-organization: the fundament of cell biology.

Williams, D. (2020). Predictive coding and thought. Synthese, 197(4), 1749-1775.

Yon, D., de Lange, F. P., & Press, C. (2019). The predictive brain as a stubborn scientist. Trends in cognitive sciences, 23(1), 6-8.

Zarghami, T. S., & Friston, K. J. (2020). Dynamic effective connectivity. *Neuroimage, 207*, 116453.

Ziegler, H. (1963) Some extremum principles in irreversible thermodynamics with application to continuum mechanics. In: Sneddon, I.N., Hill, R. (eds.) *Progress in Solid Mechanics,* North-Holland, Amsterdam, pp. 91–193