An Instrumentalist Take on the Models of the Free-Energy Principle

Niccolò Aimone Pisano

Introduction

Predictive processing is one of the most popular axes along which research on cognition is developed (Hohwy, 2013; Clark, 2016). The core insight driving this research programme is that cognitive systems operate in probabilistic terms, formulating predictions about their environment and then adjusting them based on whether their expectations are met or not. Among the various elaborations of this conceptual pillar is the Free-Energy Principle (see Friston, Kilner and Harrison, 2006; Friston, 2009, 2010, 2012, 2013), which, in a nutshell, states that adaptive systems strive to keep their free-energy (a proxy for surprise, which is an information-theoretical notion) at a minimum, by making the case that they remain within a certain range of (unsurprising) states enabling their survival. In particular, this is done by behaving in a way that approaches optimal Bayesian inference. Based on a prior probabilistic distribution linking environmental states of affairs to the sensory states the system may enter in because of them, as well as on the basis of the actual sensory states the system enters in as a result of environmental influences, adaptive systems can try to act on their environment so as minimise the likelihood that they will enter in unsuitable states for their own survival.

It is then with the Free-Energy Principle (FEP), a conceptual framework establishing an important continuity between life and cognition, that I will be concerned in this paper. In particular, I will examine the question of whether we should adopt a realist or an instrumentalist approach to the models that are crucially involved in the study of adaptive systems. In doing so, I will make a novel use of some insights coming from the literature on scientific modelling in order to show that we should indeed embrace instrumentalism. However, this will have interesting consequences for attempts at exactly characterising the nature of cognition taking the FEP as a starting point.

I will proceed as follows. I will begin (section 1) by offering a general characterisation of the FEP. The purpose is to present as informally as possible the main ideas associated with it, as well as the theoretical tools it employs. Then (section 2), I will argue that the models involved in FEP-theorising should be plausibly understood as being isomorphic to their targets (although I will remain non-committal with respect to the structuralist view of scientific representation in general, especially when based on isomorphism). This will allow me (section 3) to turn the criticisms moved against isomorphism-based accounts of representation towards the modelling practice involved in the FEP. That is, maintaining that FEP-models represent their targets as they are, in a realist sense, is unwarranted. This is because the failure to establish an isomorphism between a model and its target leads to a failure on part of the former to represent the latter, and because it is highly unlikely that FEP-models are ever isomorphic to their targets due to unavoidable design choices (driven by explanatory interests) involved in modelling practice. Consequently, while FEP-models can be empirically adequate, we should refrain from interpreting them in a realist way and go instrumentalist instead.

Finally (section 4), I will consider what implications my argument in favour of an instrumentalist reading of FEP-models has for attempts at making use of the FEP to elaborate an account of what cognition exactly is, something which has raised considerable interest through the years, especially in the light of the consolidation of the 4E approach to cognition (embodied, embedded, enactive, extended views). My conclusion is that we should not dismiss accounts of cognition based on the FEP, as they may still be informative and further our understanding of the nature of cognition. Nonetheless, the prospects of settling the philosophical debates that sparked the interest in having a "mark of the cognitive" are not good.

1. The Free-Energy Principle

I will get things started by presenting the framework I will be concerned with in this paper. This will be an informal introduction, whose purpose is to outline the core concepts and ideas constituting the Free-Energy Principle (FEP), with a special focus on the way models are made use of in this framework. While the details of the FEP are quite technical, my presentation will be completely informal, thus unavoidably imprecise at times. I hope that this will be beneficial for those unfamiliar with the view, and not too irksome for those acquainted with it. But before starting, a couple of notes. The labels "active inference" and "Free-Energy Principle" are interchangeable, and are equally frequently employed in the literature. Here, I will tend to use the label "active inference" to refer to the sort of processes implementing the FEP, while I will use the label “Free-Energy Principle” for the conceptual framework as such.

Furthermore, the exact understanding of the epistemic status of the FEP is object of debate (see, for instance, Andrews, 2021, and Hohwy, 2021). The consensus is that the FEP adopts a “principle first” approach (Van Es and Kirchhoff, 2021, p.6623) in offering an array of mathematical instruments to conceptualise and describe self-organising systems and their behaviours. However, the FEP does not appear to be a proper scientific theory; rather, it seems to be better conceived of as a principle, or as a mathematical framework. In what follows, I will refer to the FEP as a theory exclusively when I will be discussing the FEP *qua* coupled with some (contextually unspecified) process theory concerning its implementation. Otherwise, I will refer to the FEP more neutrally as a “conceptual framework”.

Finally, one important caveat is in order. As mentioned in the introduction, the FEP is tightly related to predictive processing, and, in a way, it can be thought of as a particular development of it. Accordingly, just like there exist numerous versions of predictive processing (predictive coding, prediction error minimisation…) which differ from one another under a number of respects, various readings of the FEP are available. Some are neurocentric, some are not; some are representationalist, some are not. For instance, Hohwy (2015) endorses a neurocentric, representationalist reading of the FEP, while Kirchhoff’s and Kiverstein’s (2019) is a non-neurocentric, non-representationalist reading. Here, my presentation of the FEP will be largely based on the latter approach, which is enactivism-flavoured[1]. With that being said, for the purposes of the argument I will make, it does not really make a difference what process theories are coupled with the FEP, nor whether one has (non-)representationalist or (non-)neurocentric inclinations; whenever the need to be explicit about such commitments should arise, this will be made clear.

[1] It is important to note, however, that I am not committed to the *actual* compatibility of the FEP and enactivism. There is an ongoing debate in the literature over the possibility to combine the FEP and enactivism (Allen and Friston, 2018; van Es and Kirchhoff, 2021; Raja et al., 2021; Di Paolo et al., 2022). This, however, will not concern us for present purposes, as the characterisation of the FEP presented in this paper will not make use of any technical notion coming from the enactive literature, nor from that on autopoiesis. Thus, its being enactivism-flavoured ought to be understood as pointing at features such as embodiedness, non-neurocentricity and non-representationalism, which are taken to be part of what motivates attempts at reading the FEP in properly enactivist terms, but which are not sufficient for establishing any robust relationship between the FEP and enactivism or autopoiesis.

1.1 The Free-Energy Principle: an overview

The FEP can be seen as a specific development of the currently popular view that cognitive systems are predictive systems. That is to say, cognitive system can be understood as approximating optimal Bayesian machines. As such, they formulate hypotheses in accordance with probability theory about their environment, which also includes the cognizer's internal states that are not involved in cognition. These hypotheses are what informs the cognizing organism's perceptions and, consequently, actions.

As it happens, there are various ways to interpret the Bayesian inferences cognitive systems perform, depending on how literally one takes them to occur (on this point, see Kirchhoff, Kiverstein and Robertson, 2022). One may adopt a fully literal reading, and claim that cognitive systems engage in explicit, personal-level inferences. However, this reading is likely to lend itself to all sorts of criticisms, among which some analogous to the well-known homunculus fallacy. As far as I am aware, this reading is not endorsed by many, if any, scholars working in the field, and I will accordingly leave it aside.

Alternatively, one may take a weaker, realist stance, and maintain that cognitive systems do engage in Bayesian inference, but not in a personal-level, explicit sense. In Hohwy's (2015, p.17) words: "The brain itself does not, of course, know the complex differential equations that implement variational Bayes", but nonetheless "the brain is literally Bayesian in much the same sense as the heart is literally a pump" (*ibid*.). This is the reading I have in mind in the current presentation of the FEP, and to which the instrumentalist stance I will argue for will be recommended as an alternative.

What the FEP adds to the predictive processing picture is a general principle that guides the Bayesian inferences corresponding to the various hypotheses a brain (or an organism, if one opts for an embodied reading) formulates: activities based on predictive processing tend to minimise (variational) free-energy[2]. Free-energy is an information-theoretical quantity which poses an upper bound on surprise[3]. That is, given the *actual*

[2] It can be observed that this formulation is ambiguous between two readings. According to the first, variational free-energy minimization is construed as the objective function that guides the drawing of approximate Bayesian inferences. According to the second reading, free-energy minimization is a sort of imperative that living and cognitive systems need to abide to in order to persist. I believe that both readings are viable, but the latter is more appropriate in this context.

[3] The surprise associated to some observation, or, more precisely, to the sensory states a system enters in as a result of being influenced by its environment, is the negative log probability of that observation.

states a cognitive system enters in because of the environmental data, and given the *predictions* concerning the states a cognitive system would enter in because of the expected environmental data, the measure of the mismatch between the actual and the predicted internal states, i.e. the free-energy, is always greater than surprise. Surprise, in turn, is a quantity closely related to Shannon entropy, as "on average, entropy is the long-term average of surprise" (Parr, Pezzulo, Friston, 2022, p.48)). Therefore, the FEP maintains that cognitive activities purport to minimise surprise, hence entropy, not directly, but by minimising its maximum value, which is set by free-energy.

The motivation behind the FEP is, in brief, the following. Shannon entropy (an information-theoretical quantity) is formally similar to the thermodynamic entropy (Colombo and Wright, 2021, p.S3472). Thermodynamic entropy, in turn, can be generally understood as a measure of disorder, and it naturally tends to increase, as per the second law of thermodynamics[4]. But living organisms, in order to remain alive, need to be organised in certain specific ways depending on the sort of organisms they are. Therefore, organisms need to "resist" this tendency towards disorder. Based on the FEP, engaging in cognitive activities is one way to do so.

Now, how do cognitive systems minimise their free-energy? They engage in active inference, which consists in two complementary sorts of processes that do not need to take place sequentially; on the contrary, they can and typically do occur in parallel. On the one hand, systems update the probabilities upon which their predictions are based. In other words, while the prior probability distribution of the supposed causes of their observations is represented by the generative model, if confronted with surprising sensory inputs, cognitive systems modify their recognition model, which represents the observationally informed posterior probabilities of the causes of their observations (see Ramstead, Kirchhoff and Friston, 2020). On the other hand, cognitive systems also minimise free-energy by actively modifying their environment through action, consequently changing the inputs received so as to more closely approach the expected ones[5]. In this sense, cognitive systems are engaged in self-fulfilling predictive processing

[4] It is rightfully customary in presentations of the FEP to point out here that what is properly involved here is not the second law of thermodynamics, but the fluctuation theorem.

[5] It is worth mentioning that while up to this point I have been talking of free-energy having *variational* free-energy in mind, i.e. "actual" free-energy, when it comes to the active part of active inference the relevant sort of free-energy is *expected* free-energy, that is, the free-energy that is expected to be associated

which consists in selectively sampling their environment (see Hohwy, 2016). A circular dynamics is then in place: cognitive systems formulate predictions about their environment based on their previous information; if their expectations are not matched by the incoming data, they modify themselves and the environment from which the surprising data come, so that their subsequent predictions are less likely to clash with later data.

Some crucial remarks are in order. First, the reason why self-organising, adaptive systems (and, consequently, cognitive systems) are construed of under the FEP as attempting to minimise free-energy rather than surprise directly is that those systems cannot assess how surprising the states they enter in as a result of environmental inputs are. They just are in those states. A system does not represent its predictions to confront them with the environmental data in order to measure its corresponding surprise, as that would be a computationally intractable task. Rather, systems *embody* their recognition model: their internal states are interpreted within the FEP conceptual framework as being the predictions themselves, rather than representing them. In this sense, organisms are engaged in a process of self-evidencing. This means that, being themselves predictive models of their own environment, by gathering confirming evidence in favour of their predictions they correspondingly gather evidence in favour of themselves being good models of their environment (in accordance with Conant's and Ashby's (1970) good regulator theorem)[6].

Relatedly, the way systems update their recognition model (i.e. the way they change so as to embody different expectations) and act upon their external environment to modify it does not follow any higher-order rules. Their generative model, i.e. the patterns followed in reacting to surprise and consequently engaging in active behaviour is not the sort of thing that cognitive systems consult to obtain guidance over their behaviours. The generative model can only be inferentially abstracted away from the actual behaviours adopted by cognitive systems without it being at the cognitive systems' immediate disposal.

to the future states the system will enter in as a result of different behavioural policies. With that being said, I will keep using the generic term "free-energy" in the rest of this paper.

[6] It is worth emphasizing that this is a decidedly embodied, non-representationalist way of putting this point. Non-embodied, representational alternatives are available in the literature (e.g. Hohwy, 2015, 2016; Gładziejewski, 2016). The main difference is that, according to them, organisms (and cognitive systems) do not embody, or are not themselves, their own models; rather, they *have*, or *make use of* those models.

In short, cognitive systems are interpreted, within the theoretical framework of the FEP, as embodying a recognition model, and they are said to entail a generative model (Ramstead, Kirchhoff and Friston, 2020). The minimisation of surprise, the ultimate intended effect of cognitive activities, is not what cognitive activities tend to do *per se*, because what determines surprise is not available to cognitive systems. However, what free-energy depends upon, namely the recognition model and the sensory states a system enters in as a result of certain environmental data, *is* accessible to cognitive systems, which can then try to minimise it. And, since free-energy represents an upper bound on surprise, i.e. it constrains the maximum value of surprise, minimising free-energy has the consequence of indirectly minimising surprise.

1.2 Markov blankets and generative models: models in active inference

So far, I took the distinction between the states of a cognitive system and the environment (the external states) for granted. However, it should be clear that this separation is too important for the FEP to leave it unaddressed. This is because quantities like free-energy, surprise, and the probabilities involved in the generative model all need to be quantified on the basis of the internal states of the cognitive system and of what is part of the environmentally sourced data. The separation between internal and external states is mathematically handled with the help of Markov blankets, which are derived from Pearl's (1988) notion of a Markov boundary (a Markov blanket which does not contain other Markov blankets as its subsets).

The notion of a Markov blanket is a graph-theoretic one, and it was "introduced as a way of separating a node in a Bayesian network from other nodes in the network" (Menary and Gillett, 2022, p.41). Hence, *per se*, Markov blankets are a purely formal tool used in the study of artificial Bayesian networks, and they do not straightforwardly correspond to any real-world state of affairs. This has led many[7] to forcefully contest the use that is made of Markov blankets in the literature on the FEP. For, a clear move is made from the original, epistemic use of the Markov blanket formalism in a non-empirical context, in the direction of a metaphysically committed use in an empirical context. In

[7] Bruineberg et al. (2021); Menary and Gillett (2022); but also Facchin (2021), although in this case limited to the extent in which Markov blankets can be used to settle disputes over vehicle externalism.

other words, what is contested is the legitimacy of taking this formal device to be applicable to the real world, in the sense of being able to fully account for the demarcation of the boundaries of self-organising systems purely in virtue of formal characteristics. If Markov blankets are to be used in this way, some interpretive assumptions are required, and such assumptions cannot be extrapolated from the Markov blanket formalism as appearing in its proper graph-theoretic domain.

I am sympathetic with these criticisms. However, for present purposes, I will assume that it is conceptually legitimate to make use of the Markov blankets formalism in the way the literature on the FEP does. To be clear, this is not to say that I am assuming a realist interpretation of the Markov blankets. Rather, I am conceding for the sake of the argument that it is not a category mistake to maintain that Markov blankets are what delimits the boundaries of self-organising systems.

The core idea underlying the Markov blanket formalism, as employed in the context of the FEP, is that something is part of a living system in so far as it plays a statistical role in shaping the later developments of the system. Consider a set of objects, {1, 2, 3, 4, 5}[8]. Let us suppose that some of its elements are conditionally dependent upon some other elements. That is, depending on the states that the latter are in, the former have a varying chance to enter in certain other states at a later time. In particular, suppose that: 3 is probabilistically relevant for 4 (so that depending on the state 3 is at a certain time, 4 will have a certain probability of entering in some state rather than another at a later time); 1 is also relevant in an analogous way for 4; 2 is relevant for 3; and 4 is relevant for 5. We have the following situation:

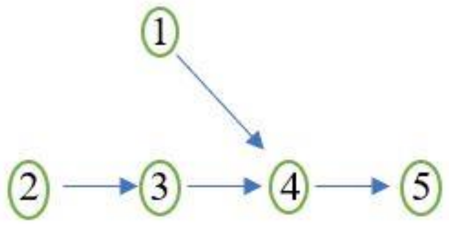

Fig.1

[8] This presentation of the core idea behind the Markov blanket formalism is largely based on Clark's (2017) and Hohwy's (2017) illustrations.

Let us now define “the parents of a variable X [as] the variables whose directed connections lead immediately to X; [and] the children of a variable X [as] the other variables to which the X leads immediately through its directed connections” (Facchin, 2021, p.5). In our scenario, the parenthood relationships are as follows: 2 is a parent of 3; 4 is a parent of 5; 1 and 3 are parents of 4.

With the needed terminology in place, it is possible to define a Markov blanket for each one of these elements as a set of nodes that makes a given node of the model in question conditionally independent from all the other nodes in the model. This set includes the parent(s), the children, and the parents of the children of the target element, and it is such that all the other elements of the model are probabilistically uninformative with respect to the task of determining the states of the element in question. For example, the Markov blanket associated to 4 in the illustration above would be constituted by its parent nodes 1 and 3, and by its child 5, while the blanket associated with 3 would comprise its parent 2, its child 4, and its child’s parent 1.

Now, how is all this employed within the FEP framework? As anticipated, this formalism is used to separate the internal states of a cognitive system from its external, environmental states. Accordingly, the node upon which a given Markov blanket is centred represents the internal states of the system itself, while the nodes constituting the relevant blanket constitute the blanket states of the system. These states are still part of the system in question, but they are not statistically separated from the environment. In particular, two sorts of blanket states can be individuated: the sensory states, which are statistically dependent on the environment and on which the internal states of the system are, in turn, dependent; and the active states, namely those which statistically depend on the internal states of the system, and upon which the environment is statistically dependent. To illustrate this with one example from above: if the internal states of the system are represented by node 4, the sensory states would be represented by nodes 1 and 3, while node 5 would represent the active states of the system.

So, a cognitive system is represented by means of the Markov blanket formalism as constituted by its internal states and by its blanket (sensory and active) states. It is at this point that we can see how the predictive processes involved in the FEP enter the scene. The bulk of a cognitive system is not directly statistically related to its environment, as it is not directly acted upon by the environmental states of affairs, nor does it directly act upon them. Furthermore, given that the boundaries of cognitive systems are determined by a Markov blanket, whatever lies outside those boundaries is not immediately available to the cognitive system as a whole, as it is something “other”, external. For this reason, cognitive systems can only make educated guesses about what actual states of affairs cause them to enter certain states (about the inaccessible generative *process*), as well as about what sort of impact on their environment certain courses of action (*policies*) will have. Such predictions are based on: prior probability distributions concerning the likelihood of various external states of affairs; the statistical correlations between those states of affairs and the resulting sensory states caused by them; the expected sensory states which will be later on caused by the environment depending on the policies undertaken. Having all this in mind, one can say that the actual behaviours adopted by cognitive systems *entail* a generative model, that is, a model which represents the aforementioned factors that ultimately lead to the observable actual behaviours.

Consider the following example. I am sitting next to a pond, on a summer evening. At some point, I slap my arm, then reach for some mosquito repellent. Based on my observable behaviour (my active states), in the light of the FEP, one can infer that I have entered some surprising sensory state such as one caused by an insect bite, and I have undertaken action so as not to enter in a similar undesirable state later on. In this sense, I, as a cognitive system, with my observable behaviours, can be said to entail a generative model. What would the generative model include in this case? Well, first of all a probability distribution concerning my sensory states. Some of them are not harmful, hence they are not associated with much increment in entropy, and, therefore, carry little surprise (recall that entropy is the long-term average of surprise). Some others, such as a stinging feeling, are instead surprising. Second, a likelihood distribution concerning the potential causes of my sensory states; in our example, it is unlikely that a tiny dart shot with a blowpipe hit me, while it is much more likely that a mosquito bit me.

Hence, from the previous scenario one can infer a model of the mechanisms that unfold while I engage in my observed behaviour. Certain sensory states are surprising, they are more likely to be caused by mosquito bites than by tiny darts, and I believe that by using mosquito repellent I will not enter such states anymore in future. That is, applying mosquito repellent is a good (expected) free-energy minimizing strategy that will allow me to reduce future surprise. This is the sort of information that a generative model[9] contains: a specification of how the cognitive system takes the world to be, of what counts as a surprising state, and of the action policies apt for future surprise minimisation.

A question that has been in recent years discussed (e.g. Colombo, Elkin and Hartmann, 2021; Van Es, 2021) concerns the status of the models involved in active inference. Are generative models just instrumentally conceivable theoretical constructs which FEP-scientists make use of, which do not correspond to anything cognitive systems really avail themselves of? Are they instead to be interpreted in a realist way, so that, as Parr, Pezzulo, and Friston (2022, p.172) suggest, the scientist's task is that of "recover[ing] the parameters of the generative model that a subject's brain uses to produce behavior – the *subjective* model […by using their…] own generative model (of how the subjective model produces behavior) – the *objective* model"? I will address this question in section 3. But before doing so, I will need to explain what notion of model seems to be at work within the FEP.

2. Models and isomorphisms

In the previous section I have presented the central ideas of the FEP, as well as the sorts of models of self-organising systems that are constructed based on the theory. Namely, living (and cognitive) systems are said to entail a generative model, i.e. the system's statistical representation of how the sensory inputs are generated as a result of the interaction with the external environment. Such systems are statistically separated from their environment via a Markov blanket, comprising the sensory states the system enters in because of the environment's influence, as well as the active states the system enters in to manipulate the environment itself. Here, I will lay the grounds for the subsequent examination of the way the FEP-informed models are to be understood. Specifically, I will introduce the structuralist view of scientific theories, dating back to Van Fraassen (1980) (and notably discussed in Van Fraassen, 2008) and I will focus on the most

[9] Notice that I am here talking only about generative models rather than recognition models. The difference between the two is, again, that the former have to do with prior probabilities, while the latter with posterior probabilities. For present purposes, focusing on generative models and leaving recognition models aside will be of no consequence.

relevant aspect of the theory for my purposes, namely the idea that scientific models relate to their targets via some morphism.

2.1 The structuralist view

One of the broadest topics from the general philosophy of science is the issue of scientific representation: how does it work? Is it different from other forms of representation, such as artistic representation (see Callender and Cohen, 2006)? Of particular interest for our purposes is the structuralist approach, which is also sometimes referred to as the "mapping account" (Pincock, 2004; Nguyen and Frigg, 2021). The core idea behind this conception of scientific representation is that scientific models represent their target phenomena by being similar to them in a specific way, namely by being related to them via some morphism. That is to say, insofar as models can be conceived of as structures (set-theoretical entities composed by a domain of elements and a set of relations defined over that domain), they can represent their targets if some morphism, i.e. some structure-preserving mapping, exists between the model and the target.

Two important remarks are in order. First, morphisms are functions, and as such they can only be defined over mathematical objects. Properly speaking, then, models are not morphic to their target phenomena *qua* natural entities, but to the mathematical structures encoding the data concerning such phenomena (see Nguyen, 2016). Second, depending on their exact properties, morphisms can be of different kinds: there can be isomorphisms, partial isomorphisms, homomorphisms, partial homomorphisms... For reasons that will be made explicit in a moment, I will here focus only on isomorphisms.

Let A = $< D; P^n_j >$ and B = $< E; Q^n_j >$ be two structures, respectively, a model and its target system, where D and E represent the domains of, respectively, A and B, and P^n_j and Q^n_j represent the n-places relations on, respectively, D and E. A function $f: D \rightarrow E$ is an isomorphism just in case two conditions are met. First, f is a bijection, so that no two elements of D are mapped on the same element of E, and for each element of E there is an element in D which is mapped on it. Second, f is relation-preserving, so that, for any n-tuple $(x_1, \ldots, x_n) \in D$, $P^n_j [x_1, \ldots, x_n]$ iff $Q^n_j [f(x_1), \ldots f(x_n)]$, and for any n-tuple $(y_1, \ldots, y_n) \in E$, $Q^n_j [y_1, \ldots y_n]$ iff $P^n_j [f^{-1} (y_1), \ldots f^{-1} (y_n)]$. Informally, this means that the model A univocally represents all and only the elements of the target system B, and all and only the relations existing among elements of the model are associated to relations among the elements of the target system. In other words, the model A is an accurate and complete representation of the target system B.

Now, I maintain that there is reason to take the structuralist view in its isomorphism-based guise to be the sort of account that best describes the modelling practice FEP-theorists engage in. When constructing an *objective* generative model for some cognitive system, FEP-theorists gather data about their target system's behaviour. Then, they construct a mathematical model which purports to describe the *subjective* generative model[10] used by the cognitive system in engaging in the active inference processes leading to the observed behaviour. Such objective generative model can be understood as a structure whose domain's elements stand for the parts of the system whose states are taken to be relevant for the processes in question. Moreover, the relations defined over such domain can be understood as corresponding to the statistical correlations the system takes to exist among the elements of the domain. Finally, given that both the objective and the subjective generative models are mathematical structures, there is no threat of a category mistake arising from attempting to establish an isomorphism between them.

Before moving on, a couple of points need to be addressed. The first is a brief caveat concerning the scope of my claim. I am not claiming that the structuralist account of scientific representation is, in general, correct. That is to say, nothing in my argument relies on structuralism correctly identifying the necessary and/or sufficient conditions for the occurrence of scientific representation. The claim I am instead making is the substantially weaker one that, even if structuralism is not the ultimately correct account of scientific representation, it does seem to reflect at least the modelling practice involved in the FEP.

The second point is lengthier. According to the structuralist view, a model represents its target only if there exists a morphism among the structures corresponding to the two of them. But what I have sketched above is a way to interpret objective and subjective generative models as structures, without saying anything specific about the morphism allegedly mapping one onto the other. Furthermore, I have anticipated that I would have taken isomorphism to be especially relevant for our purposes, as opposed to

[10] Notice that this is not the same as the generative *process*, as the generative process is what the subjective generative model purports to reconstruct, but it is not what the objective generative model is constructed to capture.

other, weaker, morphisms. Why is that? The answer can be extrapolated from works such as Hohwy’s (2016) and Kirchhoff’s and Kiverstein’s (2019, 2021) which discuss the issue of how to draw the boundaries of cognitive systems via the Markov blankets formalism.

2.2 Isomorphism and FEP-models

Isomorphism is a strong kind of morphism, as it requires that there is a relation-preserving bijection among two structures. Because of this, it has been noted that isomorphism-based theories of representation struggle to account for misrepresentation (Suárez, 2003). By misrepresentation, I mean a representation which is inaccurate (a representation possessing features which are not possessed by the target) or incomplete (it fails to represent some features of the target), but which nonetheless counts as a representation of its target. This inability to accommodate misrepresentations is problematic if isomorphism is meant to be the mapping upon which scientific representational processes are founded. In fact, it is not unusual that scientific models are incorrect in one way or another without losing their representational power as a result, as they instead should, were their powers based on isomorphism.

To respond to this worry, some have attempted to frame structuralist accounts of scientific representation not in terms of isomorphism, but in terms of weaker morphisms, such as partial isomorphism (e.g., Bueno, 1997; Bueno and French, 2011) or homomorphisms (Bartels, 2006; for a critical discussion, see Pero and Suárez, 2016). The idea guiding these alternatives to isomorphism is to enable structuralist accounts of representation to successfully deal with misrepresentation, that is, to allow models to maintain their representational characteristics despite inaccurately or incompletely depicting their targets. Adopting a morphism weaker than isomorphism may afford more flexibility in accommodating misrepresentation, which is beneficial for accounts of scientific representation. However, there is evidence coming from discussions over FEP-models suggesting that isomorphism proper is what is relevant in that context. I have specifically in mind debates over the question of whether for each cognitive system there exists a unique Markov blanket enclosing it.

Recall that the task of a FEP-researcher, according to some of the most prominent upholders of the view, is to construct an objective generative model, which is meant to reflect the subjective generative model giving rise to the observable free-energy minimising behaviours a cognitive system displays. This suggests, as Clark (2017, p.12) pointed out[11] and as also Parr, Pezzulo and Friston (2022, pp.106-110) appear to think, that there is some degree of arbitrariness involved in this modelling practice, in the sense that the choice of a certain model is not univocally dictated by the intrinsic characteristics of the target. Rather, the determination of the boundaries of the cognitive system in question will plausibly be significantly influenced by our explanatory goals. Specifically, depending on the behaviours *a modeller* individuates, and on how these behaviours are individuated, different alternative mechanisms and generative models may be relevant in originating them. It seems then that it is in principle impossible for us to be sure that a given objective generative model will match the subjective generative model the cognitive system makes use of. Remember that the subjective generative model is, by definition, a probability distribution concerning the cognitive system's take on the way its sensory states are generated. But the subjective generative model need not, and typically does not, precisely capture the actual way in which such states originate from the environmental influence upon the system. In other words, the subjective generative model is *not* the same as the actual generative *process*. This is particularly problematic if one intends to recover the subjective generative model as opposed to reconstructing the generative process. The reason is that, as external observers, on the one hand we have access to the observable behaviours which are a function of the subjective generative model, and on the other we may have access to part of the generative *process*. Consequently, we may incorrectly parametrize an objective generative model based on the information coming from our access to the latter, which may not correspond to the actual parameters of the subjective generative model.

An illustrative analogy may be helpful. Consider the earlier example in which I reach for my mosquito repellent while sitting next to a pond on a summer night. An external observer watching the scene may notice a mosquito biting my arm. Thus, they would assume that it is this environmental influence that causes some surprising sensory state, which, in turn, ultimately leads me to reach for the mosquito repellent as an attempt to avoid future similar surprising states. However, it is possible that I did not notice the

[11] "Complex living beings are composed of layer upon layer of Markov blankets […]. Different explanatory purposes drive us to highlight some of these blankets (of blankets) at the expense of others. But no blanket or set of blankets is privileged in and of itself".

mosquito bite, and I intended to apply it on myself just because I am particularly fond of its smell. Indeed, the behavioural data the observer takes into consideration in reconstructing what is going on may be partial (by observing the behaviour a little longer, it may turn out that I wanted to check the expiry date of the repellent) or altogether incorrect (I was really trying to reach for something else). Be that as it may, it is not just the observable behaviour of the cognitive system in question that drives our construction of the objective model. A non-negligible role is played by our consideration of facts that we (correctly or incorrectly) take to be part of the generative process, which we assume to be relevant inputs for the subjective generative model's delivering a certain behavioural output.

In short, one can arguably maintain that our modelling practices in the context of the FEP are importantly influenced by our explanatory interests. Consequently, one may be inclined to pick a certain Markov blanket rather than another, because, with respect to certain behavioural data, conceiving of the relevant cognitive system in one way rather than another may seem more appropriate. Therefore, there seems to be reason to think that it is not possible to find a principled criterion to individuate cognitive systems: there is no way to draw these boundaries fully independently from any explanatory interest. At any given time, there is a multitude of potential Markov blankets one may pick to separate a cognitive system from its environment, and, consequently, to constrain the subsequent objective generative model one will construct.

Some have challenged this final conclusion, which is sometimes labelled (e.g. by Kirchhoff and Kiverstein (2021)) "*proliferation*". For instance, Hohwy (2016) has made a case for privileging the brain as the cognitive system of interest. But, even if one were to accept his proposal, the issue would just be pushed further back, rather than solved. For, even granting that the brain is by default the cognitive system of interest, the threat of an ensuing slippery slope shrinking down the dimension of that cognitive system would emerge (see Anderson, 2017). What constitutes the outer layer of the brain, and how should we model the different parts of the brain? Should we take just the outermost layer of neurons to constitute the Markov blanket? Why not the next inner one? This is referred to as the "*shrinkage*" problem by Kirchhoff and Kiverstein (2021), who also proposed their solution to both *proliferation* and *shrinkage*. According to them, there is not a single, persisting Markov blanket that demarcates the boundaries of a cognitive system throughout its existence. Nonetheless, at any given time it is possible to determine which among the various alternative Markov blankets one should choose to demarcate the boundaries of the cognitive system in question: the right Markov blanket is the one which, in terms of average free-energy minimisation, best accounts for the continued existence of the relevant system for a target period of time.

For now, it is not important to settle the dispute over the potential plurality of Markov blankets associated with a cognitive system. What matters for the present purposes is an implicit assumption that underlies all the views I have sketched, namely that there is *one correct way* to demarcate the boundaries of a cognitive system, and, consequently, to construct the objective generative model. It does not matter whether the subjective generative model separated from the environment by a Markov blanket is always the same (and identifiable with the brain, as per Hohwy). Nor does it matter whether it is diachronically negotiable (as per Kirchhoff and Kiverstein), or whether different subjective generative models are to be constructed relative to different observable behaviours (Parr, Pezzulo, Friston, 2022, p.56). Once a target phenomenon is pinpointed, to construct an objective generative model one has to assume that there is a unique subjective generative model associated with that phenomenon. This, at last, is the reason I believe that FEP-modelling, i.e. the construction of objective generative models, needs to be construed in terms of isomorphism, rather than in terms of some weaker morphism. Although there may not be a correct way to individuate the target phenomenon, namely the active inferences a cognitive system engages in, the underlying assumption is that for each putative target phenomenon there is a single subjective generative model, which needs to be reconstructed by modellers. If the morphism between objective and subjective generative models is weaker than an isomorphism, then the objective generative model would fail to be sufficiently similar to the subjective generative model so as to ensure that it corresponds to the unique subjective generative model associated with the phenomenon to be modelled, whichever that may be and however that may be individuated.

Before moving on, one potential worry needs to be addressed. So far, I have argued that any morphism short of being an isomorphism would be insufficient for the purpose of picking the right subjective generative model. But there is a sense in which isomorphisms themselves may not be fully adequate for the purpose. As it has been long well-known (for an early elaboration of this point, see McLendon, 1955), isomorphisms are not as strong as they may appear at first. Indeed there are two senses in which, even though an isomorphism can be established, an objective generative model may fail to be correctly related to the intended subjective generative model. First, there may be more than just one isomorphism holding among two structures. Second, a structure typically is isomorphic to more than just one other structure. I don't think that the former consideration is especially troublesome for present purposes. The latter, instead, casts some doubts over the real adequacy of isomorphism as the morphism grounding the representational link meant to exist between subjective and objective generative model. In fact, it appears that isomorphisms are vulnerable to criticisms akin to the ones moved against less stringent morphisms: they are not strong enough to guarantee that the objective generative models will map onto the right subjective generative models.

I acknowledge the legitimacy of this worry. However, I believe that the sense in which isomorphisms are weaker than it would be desirable for the present purposes is different from the sense in which other morphisms are. Let me illustrate what I intend by this with a brief thought experiment.

Imagine that you take a perfectly clear picture (call it *pic1*) of a woman named Alice. It turns out that, unbeknownst to you, the woman you took a picture of is not really Alice, but her identical twin sister, Beth. Despite your photograph not being *really* a picture of Alice, it would make no difference in any relevant sense that this is so: it possesses all the features you can possibly be interested in, were you to examine the picture in order to learn something about the physical appearance of Alice. In this sense, it checks all the *required* boxes for being accepted as a representation of Alice, although that may not be *enough* for really being such. Nonetheless, had the woman in the picture really been Alice, and not Beth, the picture would be a perfect representation of Alice: the reason why *pic1* is not perfectly adequate does not have to do with the properties of the picture itself, but on external circumstances. On the other hand, if your picture (*pic2*) also does not come out as perfectly as you hoped (say, her left arm is left out of the frame, or it is not clear whether she has freckles or not), the resulting picture may not be good enough to be used to learn everything you may be interested in about Alice's physical appearance. This would be the case even if Alice had been the subject of *pic2*: at least part of what makes *pic2* inadequate has to do with the properties of *pic2* itself.

The point I am driving at is that *pic1* and *pic2* are inadequate in different senses. Neither has *everything* it takes to be a perfectly useful representation of Alice (they are both insufficient). But in the case of *pic1*, this does not have to do with the features of the picture itself, and there is no practical difference as a result. On the contrary, in the case of *pic2*, because of some features of *pic2* itself, you may either be unable to learn certain things about Alice, or you may learn the wrong things (perhaps because, due to some light trick, her eyes appear of a different colour). This, I maintain, is analogous to the different ways in which isomorphisms and weaker morphisms are not perfectly up to the task when it comes to the representational relation meant to occur between subjective and generative models. Isomorphisms may not be sufficient to guarantee that a given structure is a univocal representation of some other structure, but this does not have to do with any of the features of the two structures *per se*. On the other hand, weaker morphisms are not sufficient because the structures connected by such functions are not suitable for the modelling practice in question. In terms of generative models, it seems that subjective generative models are such that they can be adequately captured by the objective generative models only if there is a function at least as strong as isomorphism in place, although the existence of such function may not be all is needed overall.

To conclude, I wish to reiterate that this is not to say that creating FEP-models requires structuralism to be globally true as an account of scientific representation, let alone that all scientific representations need to be isomorphic to their targets. Nonetheless, the assumption made in debates over the modelling practice involved in the FEP appears to be that the relation between objective and subjective generative models needs to be conceived in isomorphism-based structuralist terms. The generative models FEP-scientists construct must be isomorphic to the generative models entailed by the free-energy minimising strategies adopted by cognitive systems. This is a necessary, although in all likelihood not sufficient, condition.

3. Against the realist stance

Let us take stock. In the first section I have presented the Free-Energy Principle, according to which living organisms manage to stay far from thermodynamical equilibrium by engaging in active inference, i.e. by engaging in strategies apt to minimise free-energy, a proxy for surprise. The separation of a given system from its environment as well as the system's probabilistic "beliefs" about the way its sensory states are generated are what FEP-models are meant to capture. More exactly, according to the FEP, the free-energy minimising strategies adopted by the relevant systems are the manifestation of a subjective generative model. It is then these subjective generative models that modellers try to reconstruct by elaborating their (objective) generative models, which should be understood as purporting to be isomorphic to the subjective generative models. It is now time to examine more closely how talk of "embodying a generative model" and "being delimited by a Markov blanket" are to be interpreted. This is what this section sets out to do, by applying the points raised in the second section to the particular case of the modelling practice guided by the Free-Energy Principle. The picture that will emerge discourages adopting a realist stance on FEP-models, because of the issues related to isomorphism-based representational processes.

As I mentioned earlier, it is no secret that isomorphism-based structuralism struggles to deal with misrepresentation. Broadly speaking, insofar as representations in general, and scientific models in particular, purport to represent their targets by containing information about their targets' features, they can end up misrepresenting their targets in three ways. First, they may fail to include some more or less important features of the target system, in which case they would be incomplete representations; this is the case, say, of a scale model of a building, which, differently from the real building, may not have windows, or a detailed inside. Second, they may contain information that does not correspond to actual properties possessed by the target system; this is the case of a planisphere, which features fictional lines indicating parallels and meridians. Third, they may misrepresent the target system because of a combination of the first two ways to misrepresent; for instance, a toy model of the solar system may leave out features of the real system such as the presence of moons around Jupiter, while it might have thin metallic sticks keeping the planets suspended at fixed distances from each other (which obviously do not correspond to anything in the real solar system).

Misrepresentation is not inconsistent with partial representation in all three cases. The tiny building still represents the real building even if it is lacking on some details (it is incomplete), the planisphere still represents the world even if it represents non-existing lines (it is inaccurate), and the toy solar system still represents the solar system even if it does not represent Jupiter’s moons and even if the real planets are not kept in place by sticks (it is incomplete and inaccurate).

Nonetheless, it may happen that the representational process fails entirely because of any of the three cases presented. Imagine, as an illustrative analogy, that you ask me to draw a dromedary, and I draw a camel with two humps on its back. Dromedaries are camels with only one hump, so, in virtue of having drawn a camel with one hump too many, I have just drawn a generic camel, but I failed to represent a dromedary. Or, conversely, imagine that I am asked to draw a unicorn, and I draw what seems to be a normal horse. In virtue of lacking a crucial feature, my drawing fails to represent a unicorn. In both cases, the intuitions intended to be elicited are to the effect that I end up representing something else than what I intended to represent. I am not representing my targets at all, even if my representation is meant to represent them, and even if the first drawing is a complete and almost entirely accurate representation, while the second is an accurate and nearly complete representation.

Now, if, as I have argued, FEP-modelling is based on isomorphism, it faces this sort of problem. If being isomorphic to its target is a necessary condition for an objective generative model to represent the relevant subjective generative model, then failure to establish an isomorphism between a model and its target amounts to the objective model’s failure to represent its intended target. Injectivity without surjectivity (i.e. accuracy without completeness), surjectivity without injectivity (i.e. completeness without accuracy), bijectivity (i.e. injectivity and surjectivity together) without “relation-preserving-ness”, or “relation-preserving-ness” without bijectivity; none of these options will do. Each of these four ways in which a function may fail to be an isomorphism between objective and subjective generative models, and which may result, in turn, in one of the three aforementioned ways in which misrepresentation might occur, is enough for the representational process to fail entirely, as far as isomorphism is concerned. This issue has important consequences with respect to the stance we should adopt on models based on the FEP. That is, if, for any given data-set obtained from the observation of a cognitive system’s behaviours there only corresponds one subjective generative model, and failure to establish an isomorphism between that model and the modellers’ objective one leads to failure to represent the former, then we cannot be realist about the content of our objective generative models.

Let me elaborate. My criticism of realism with respect to FEP-models moves along different, weaker lines than other extant views. For instance, because of an observed systematic ambiguity between subjective and objective generative models, van Es (2021) denies that we should be realist about subjective generative models, in that reflections upon their objective counterparts does not warrant their reification. This is similar to, but subtly different from, what I maintain, as I do not take the distinction between subjective and objective generative models to be blurred and thus unable to warrant a realist stance on the former. What I do maintain is that, while the distinction may still be a meaningful one, we should not take our own (objective) models of such (subjective) models to perform their intended representational function. This is because the link between subjective and objective generative models is severed by the overwhelming likelihood that the required isomorphisms backing it up fails to obtain. In other words, I am not questioning the in-principle viability of realism, but only the actual effectiveness of the means by which such realism is meant to be bolstered. What I take issue with is not the possibility of making a realist move based on the sharpness of the distinction between subjective and objective models. I am, for the sake of the argument, granting that there are sufficiently solid conceptual grounds for this distinction, so much so that realism is not precluded. However, the realist move fails nonetheless, because the representational link which it requires breaks down.

In short, I am not ruling out in principle that the FEP may still ultimately turn out to get things right about how life and cognition work in general, including the fact that cognitive systems' behaviours really do entail, in a realist and theorist-independent sense, (subjective) generative models. What I am denying is that this potentiality is enough to warrant taking our own (current) models to represent what really goes on when cognitive processes unfold. Hence, instrumentalism. I will return to this point in a moment.

Perhaps many will find this line of reasoning a little unusual. Generally speaking, antirealist arguments tend to deny that our theories "get things right" as opposed to just being empirically adequate, because the existence of some specified connection between theory-informed models and the target phenomena is necessary but not sufficient to warrant realism about the core claims of the theory in question. What I am claiming here proceeds in the opposite direction. Even if it really were the case that we are ultimately right about what the theory generally says concerning the kind of target phenomena, that would not be sufficient to take our theory-informed models to entertain the required sort of link with their target phenomena. To use an everyday analogy: even if there is a cat on my bed, not every perceptual experience of a cat on my bed would thereby be veridical, as it may be a hallucination. Consequently, even if true, the realist belief that there is a cat on my bed would not be warranted. Similarly, even if it were not ultimately wrong to be realist about subjective generative models, this would not entail that the subjective generative models isomorphic to the corresponding objective generative models are the ones that we should be realist about. Consequently, even if true, the general claims afforded by the theory should not construed in a realist way, as that would not be warranted.

It is important to point out that this, however, does not make the FEP a hopelessly empirically empty mathematical framework (for some critical discussions: Colombo and Wright, 2021; Andrews, 2021), or at least not completely. If a theory can be said to have empirical content insofar as the claims it affords apply to the world because of the possibility of constructing models of the relevant phenomena, then the FEP, once coupled with some process theory, can meet this requirement in two cases. First, in case we overcome the difficulties undermining the isomorphism meant to hold between an objective generative model and its subjective counterpart. This is obviously a virtually impossible task for our scientific community, as it would require finding an objective, ideal way to conceptualise cognitive systems, so as to solve the previously discussed issues related to drawing Markov blankets, for instance.

The second and significantly more viable possibility consists in embracing instrumentalism. That is, it consists in accepting that the relation between objective and subjective generative models is different from what it is currently thought to be. Under an instrumentalist reading of the modelling practices carried within the FEP framework, target systems may or may not actually engage in active inference. In either case, what warrants the modelling of the relevant target systems and phenomena by means of the FEP's array of conceptual tools is not the fact that objective generative models (and Markov blankets) correspond to the way their targets are. Rather, it is the fact that they are empirically adequate, in the sense of being explanatorily, descriptively, and predictively effective, at least to some degree.

The crucial point is that, in an instrumentalist setup, the representational link between objective and subjective generative models cannot be problematically severed, because such link is not established in the first place. This is because there is no commitment to the reality of subjective generative models. Subjective generative models are not the actual aspect of target systems objective generative models are meant to capture. Instead, they are a fictional (instrumental) conceptualisation of the target system, playing a role subordinate to empirical adequacy. That is, objective generative models aim at being empirically adequate models of their target systems, and this agenda is facilitated by their approximating the subjective generative models stemming from a construal of the target systems "as if" they were engaging in active inference. But, to reiterate, the failure of an objective generative model to be isomorphic to its subjective counterpart is not an issue. For, this does not amount to failing to be a model of the real target system.

One may argue that my understanding of realism commits what Kirchhoff, Kiverstein and Robertson (2022) have labelled the "literalist fallacy". According to Kirchhoff and colleagues, realist approaches to the FEP have been misguidedly criticised based on an overly demanding understanding of what realism maintains. In their view (which is akin to the one upheld by Godrey-Smith, 2003, 2009), we should be realist about FEP-models as generalised models[12], that is, as capturing a "family" of phenomena. Following Weisberg (2006, 2007), Kirchhoff and colleagues maintain that this understanding is immune to the sort of issues related to misrepresentation (in the sense I have been using the term in the present paper), insofar as the ultimate goal is to expunge them from our modelling practices. To claim otherwise, and accordingly criticise realism and upholding instrumentalism (as I did) is to commit the literalist fallacy; that is, it is to take realism to depend on the ability of our current models to be perfect, literal representations of their targets. This, in their view, is a mistake. Realism is not committed to such an unrealistic (pun intended) claim. Rather, realism is to be understood as maintaining that while our theories, *broadly speaking*, get things right about their targets, they will get the *details* right only on the long run, by eliminating, or reducing as much as possible, the use of idealisations.

[12] This terminology is originally introduced by Weisberg (2013).

As I understand this view, the gist is that FEP-models are *not literally false*[13]; rather, they are *approximately true*, and their presently being only approximately true does not constitute a problem for long-term realism. Simultaneously, this undermines attempts at putting forward instrumentalist readings of the FEP, as stronger reasons than current inaccuracy are required to bolster instrumentalism. However, although a full discussion of the form of realism they endorse is beyond the scope of the present paper, some remarks in response to Kirchhoff's and colleagues' view need to be made.

First, it seems to me that their argument can be turned on its head. As mentioned above, instrumentalism differs from other forms of antirealism as it does not make the positive claim that scientific theories (and the FEP as such) *are not* true. They may or may not[14] be. What instrumentalism says is that, in the absence of irrefutable reasons to maintain their truth, scientific theories should be assessed purely in terms of their empirical (phenomenal) adequacy, without any commitment as to whether they "get things right", or even approximately right, with respect to what their targets really are like. In particular, the way I have argued for instrumentalism does not hinge on the falsity of current (and, likely, future) FEP-models. Rather, it hinges on their failure to establish the desired representational link with their targets. As a consequence, my case for instrumentalism should not be read as a manifestation of impatience, so to say, and unwillingness to wait for better, more precise models. Indeed, for a theory to be even just approximately true, as realists claim, that theory needs to represent its objects in the first place. Hence, if what I have been arguing for so far is correct, what realists need to do to block my instrumentalist argument is showing that there is a way for FEP-models to represent their intended targets in a way that accommodates misrepresentation. But, given what I have argued in section 2.2, it seems to me that the most plausible account of representation applying to the link between the objective and subjective generative models, in a realist context, is the isomorphism-based one I have been discussing.

In summary, by abandoning a realist approach to models based on the FEP, and adopting an instrumentalist one as an alternative, it is possible to avoid the problems associated with taking the representational relation between objective and subjective generative models to be based on isomorphism. In the next, final section I will consider

[13] *Contra* Klein (2018, pp.2253-54).

[14] In Van Es's and Hipolito's (preprint, p.16) words: "instrumentalism in itself is characterized by ontological agnosticism with regards to what actually makes a system tick".

how this position would reflect on the usefulness of the FEP as a theory from which to derive precise accounts of cognition specifically, or “marks of the cognitive”.

4. Implications for marks of the cognitive

In the light of the argument presented in the previous section, it seems that the models of cognitive systems based on the FEP should not be construed in a realist way. This is because, I maintain, they would plausibly be meant to be isomorphic to the generative models (and Markov blankets) they describe. Hence, given that approximations, idealisations, and in general variably arbitrary design choices are virtually impossible, and indeed undesirable, to expunge from modelling practice, the required isomorphisms would systematically fail to obtain. As a consequence, even though scientific representation in general may not depend on isomorphisms to succeed, the representational attempts fail to go through in the case of FEP-models. In other words, insofar as FEP-models are understood in a realist way, they cannot succeed in representing their targets. This leaves only one option to FEP-theorists, namely conceiving of their models in an instrumentalist way. This, as we will see shortly, has interesting consequences for attempts at formulating an account of cognition (a mark of the cognitive) in terms of the Free-Energy Principle.

4.1 The Free-Energy Principle is not a theory of cognition

One aspect of the FEP that the community of scholars variously concerned with it has acquired awareness of in very recent years is the fact that the FEP as such does not constitute a theory of cognition. Or, at least, it does not constitute a theory of cognition *specifically*. In fact, the claim that adaptive systems engage in active inference, thus increasing their likelihood of entering in unsurprising states and consequently remaining far from thermodynamical equilibrium (i.e. death) does not specify anything that may distinguish cognition from other phenomena such as digestion, or even life itself[15]. While

[15] The worry that the scope of the FEP may end up being too broad may arise here. Such worry has been addressed by Kirchhoff et al. (2018), and by van Es and Kirchhoff (2021), by drawing a distinction between “mere” and “adaptive” active inference.

this may seem at first an interesting characteristic of the FEP, as an important connection between life and cognition is established[16], such connection may lead, in some philosophical areas, to conceptual problems. If to be alive just is to engage in active inference, and if to be a cognitive system just is to be a system engaging in active inference, then there is no difference between being a cognitive system and being a living organism. But, given the current state of our knowledge and our conception of the natural world, the distinction between cognition and life is one that, at least for some philosophical purposes, ought to be preserved. This, of course, does not mean that the FEP as such needs to accommodate this distinction. Rather, it means that, if one intends to apply the FEP's conceptual framework to tackle philosophical issues that specifically have to do with cognition as a distinct natural phenomenon, then one needs to supplement it, at the level of the process theories paired with it, with some cognition-specific elements.

In short, one reason why the distinction between life and cognition is important is the interest that many philosophers of mind and the cognitive sciences have in the nature of cognition as a specific natural phenomenon. Indeed, one of the effects of the development of the 4E views on cognition (embodied, embedded, extended and enactive cognition) has been the increased felt need of a clear account of what cognition is. This need, clearly voiced for example by Adams (2010) and Wheeler (2019) (but for a sceptical take see Clark, 2019), has led to the formulation of a number of proposals. Some proposals (Adams and Aizawa, 2008; Rowlands, 2010; Adams and Garrison, 2013) are directly linked to the debate over extended cognition stemming from Clark and Chalmers (1998), while others are more closely related to the literature over the contrast between anthropogenic and biogenic approaches to cognition (Lyon, 2006; Van Duijn, Keijzer and Franken, 2006; Keijzer, 2021).

To obtain a theory specifically of cognition based on the FEP one needs therefore to add further constraints to the core claims constituting the FEP. Differently put, one needs to show how cognition enables adaptive systems to minimise their free-energy, while at the same time differentiating it from other free-energy minimising phenomena. This is what, for instance, Kiverstein and Sims do, arguing that cognition ought to be

[16] Kirchhoff, Froese (2017); Kirchhoff (2018); Bruineberg, Kiverstein and Rietveld (2018). This idea appears also in the literature on autopoiesis (Maturana and Varela, 1980) and on enactivism (Thompson, 2007; Di Paolo, 2009)

understood in terms of allostatic control, which is “prospective behaviour directed at avoiding the anticipated divergence from homeostatic setpoints” (2021, p.25). Their criterion draws on the FEP: insofar as homeostasis is achieved by entering states associated with low free-energy, allostatic control results in free-energy minimisation. Moreover, what Kiverstein’s and Sims’s account stresses is the proactive nature of the behaviours involved in allostatic control. That is, such behaviours must not be purely reactive, but also, and crucially, anticipatory. This allows genuine instances of cognition to be discerned from other free-energy minimising processes. For instance, the circulatory system of some creature plausibly should not be considered responsible for any cognitive activity that creature may be said to perform, as its behaviours would be purely reactive, instead of proactive.

Now, I am not concerned here with the assessment of the strength of Kiverstein’s and Sims’s proposal specifically, nor of any other alternative account[17]. What I intend to point out is just that it is possible to elaborate an account of cognition in line with the FEP, as it has indeed been done. But in order to do so, one needs to specify some distinctive characteristics of cognition that other free-energy minimising strategies do not possess: cognition is just one of the ways in which adaptive systems stay attuned with their environment in a suitable way for their survival, but it is not the only one, and it may need to be distinguished from the others.

4.2 Instrumentalism and the mark of the cognitive

So, the FEP is not, *per se*, a theory of cognition, but it can be the conceptual framework in which more specific proposals for a mark of the cognitive are couched. However, based on the points raised throughout this paper, the models of cognitive systems that might be constructed in such a conceptual framework encounter a series of problems if looked at from a realist perspective.

Realism is an appealing position, especially when it comes to the characterisation of natural phenomena like cognition. Being realists about what our theories of cognition

[17] Indeed, Kiverstein’s and Sims’s work is a response to Corcoran, Pezzulo and Hohwy (2020). Corcoran and colleagues also maintain that active inference is what grounds the appearance of cognitive phenomena, but, unlike Kiverstein and Sims, they spell their account in terms of counterfactual active inference, rather than in terms of allostatic control.

have to say means taking our theories to capture, to different degrees of accuracy and completeness, what the nature of cognition is, as well as how cognitive systems are structured and work. But realism with respect to FEP-models is problematic. Consider again Kiverstein's and Sims's proposal that cognition has to be understood in terms of allostatic control. Adopting a realist stance on this view means underwriting the claim that cognitive systems display allostatic control independently from any external observer's acknowledgement of this fact, so that allostatic control is not just a theoretical construction helping us to make sense of what cognitive systems do. Furthermore, any model representing the (subjective) generative models supporting allostatic control would have to be taken as faithfully depicting the models really in play in the generation of these phenomena. However, as we saw, this risks being too optimistic a view, one which fails to take into consideration the unavoidable creeping in of theorist-dependent design choices in the elaboration of such models.

This can be noted by taking a closer look at one of the examples Kiverstein and Sims discuss, namely the case of slime mould, *Physarum polycephalum* (2021, pp.19-20). Slime mould has been observed to slow down its motion, while foraging for food, when it anticipates the periodic occurrence of a dry stimulation which would normally elicit such behaviour (Saigusa, Tero, Nakagaki, and Kuramoto, 2008). This anticipatory behaviour motivates Kiverstein's and Sims's conclusion that slime mould manifests cognition, because, by engaging in this proactive, not purely reactive, behaviour, it reduces its expected free-energy (it decreases the likelihood of entering future surprising states).

Now, for the "allostatic control" mark of the cognitive to be one that stems from the FEP, it is not enough that putative cognitive systems display some form of allostatic control: they also need to do so in accordance with the concepts the FEP is concerned with. This means that one needs to create a model of the system under examination by identifying its Markov blanket (its sensory and active states) and by specifying its generative model. But this is where the modelling issues I have been concerned with enter the scene. As the authors of the study on slime mould explain, the locomotion of the entire organism depends on multiple chemical oscillators (Saigusa, Tero, Nakagaki and Kuramoto, 2008, p.3), which means that the active states of the organism (those corresponding to its movement at different speeds) are determined by the internal states of the organism, to be understood in terms of such chemical oscillators. In turn, these internal states are influenced by other states the system enters in: the relevant periodic sensory states the slime mould enters in as a result of dry stimulation on the one hand, and other internal states on the other hand, which, over time and provided that the periodic dry stimulation is not offered in the meantime, "reset" the system to its baseline condition, so that the periodic slowing-down ceases.

All these factors need to be appropriately expressed in our FEP-informed model of the activities of the slime mould. If one fails to precisely model all the relevant chemical oscillators (an arguably difficult task), this may or may not have important consequences in terms of empirical adequacy of the resulting model, but it surely would be enough for the objective generative model in question not to be isomorphic to the subjective generative model. And this, as we have seen earlier, makes the representational relation between the two fall apart, thus warranting an instrumentalist, instead of realist, take on the objective generative model. Such a model may be empirically adequate, but it does not capture what "really" is going on in the organism under examination, from the organism's perspective.

The case of the slime mould is a relatively simple one, and one may be understandably not too worried about interpreting a FEP-informed mark of the cognitive instrumentally. But the disputes that led many to think that an account of cognition is of crucial importance make instrumentalism undesirable. Consider the debate over the extended cognition view. Is Clark's and Chalmers's (1998) well-known imaginary Otto, a person affected by Alzheimer's disease who heavily relies on the information written in his notebook to navigate the world, involved in an extended cognitive system whose boundaries encompass not only Otto himself, but also his notebook? According to Kirchhoff and Kiverstein (2019, 2021), yes, because taking this to be the case would allow us to have a better explanation for the continued existence of the "Otto/notebook" system than we would have otherwise. However, while this is consistent with an instrumentalist approach to the issue, instrumentalists of different inclinations may not feel the need to include the notebook in the picture. This is indeed a viable move, as it would only require us to consider the states Otto himself enters in as a result of interacting with his notebook as sensory states rather than internal states statistically conditioned by other internal states. In terms of the empirical adequacy of our modelling, all else being equal, there would not be much of a difference. Otto would enter the same active states, and he would tend to minimise his free-energy in similar ways. Furthermore, proponents of the "Otto/notebook" model or of the "Otto model" would have an equally questionable claim to the correctness of their conception of the cognitive system in question. For, any matter of fact able to settle the dispute would be out of reach in virtue of the overall instrumentalist framework in which the disputants would be working. As long as their models are equally empirically adequate (provided that the contenders can even agree on how to assess this), neither disputant can expect the world to tip the scale in favour of their preferred model, because both FEP-informed ways of conceiving the cognitive system would ultimately fail to represent the real cognitive system.

To summarise, instrumentalism may not prevent FEP-based accounts of cognition from being informative about the nature of cognition: there is no question that, for instance, Kiverstein and Sims make an interesting and substantial claim about what cognition is (regardless of whether it is correct or not, and of whether there is an actual need to account for cognition specifically). Nonetheless, as I have tried to show just now, going instrumentalist would make many of the philosophical disputes on cognition, which led to the felt need for an account of cognition, impossible to be solved, even if a mark of the cognitive were offered.

Conclusion

In this paper, I have argued that we should be instrumentalists, instead of realists, about the models of the Free-Energy Principle. Instead of arguing directly for this position by questioning whether adaptive systems (and cognitive systems in particular) should be literally taken to engage in active inference, I made my case using some insights coming from the literature on scientific modelling.

After having introduced the FEP and the use of models involved in it, I have argued that objective generative models should be interpreted as intended to be isomorphic to subjective generative models, and that it is in virtue of this isomorphism that the former represent the latter. But, if the representational process requires the existence of an isomorphism between the representing and the represented structure, and since specifying the representing structure is largely a matter of more or less arbitrary design choices, then it appears that one can hardly ever hope that objective generative models will represent their intended subjective counterparts at all. If this is so, then there is no reason to maintain that FEP-models (models of adaptive systems based on the FEP) are to be interpreted in a realist way. In fact, for realism about FEP-models to be warranted, such models should at the very least be representations of their targets, but that is very likely not to be the case.

Finally, I have concluded with some reflections on the FEP as a source of accounts of cognition. The FEP *per se* is not specifically a theory of cognition, but attempts at formulating a mark of the cognitive based on the FEP can and have been made. However, since we should be instrumentalists about FEP-models, while a mark of the cognitive based on the FEP may further our understanding of cognition, it will not help us to settle in any specific case the philosophical disputes whose solution is thought to need an account of cognition.

References

- Adams, F. (2010). Why we still need a mark of the cognitive. *Cognitive Systems Research*, *11*(4), 324-331. https://doi.org/10.1016/j.cogsys.2010.03.001
- Adams, F., & Aizawa, K. (2008). *The Bounds of Cognition*. Oxford: Blackwell.
- Adams, F., & Garrison, R. (2013). The mark of the cognitive. *Minds and Machines*, *23*(3), 339-352. https://doi.org/10.1007/s11023-012-9291-1
- Allen, M., & Friston, K. J. (2018). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, *195*(6), 2459-2482. https://doi.org/10.1007/s11229-016-1288-5
- Anderson, M. (2017). Of Bayes and bullets: An embodied, situated, targeting-based account of predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing (Vol. 3).* MIND Group: Frankfurt am Main.
- Andrews, M. (2021). The math is not the territory: navigating the free energy principle. *Biology & Philosophy*, *36*(3), 1-19. https://doi.org/10.1007/s10539-021-09807-0

- Bartels, A. (2006). Defending the structural concept of representation. *Theoria: An International Journal for Theory, History and Foundations of Science*, *21*(1), 7-19.
- Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195(6), 2417–2444. https://doi.org/10.1007/s11229-016-1239-1
- Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022). The Emperor's New Markov Blankets. *Behavioral and Brain Sciences, 45,* E183. https://doi.org/10.1017/S0140525X21002351
- Bueno, O. (1997). Empirical adequacy: A partial structures approach. *Studies in History and Philosophy of Science Part A*, *28*(4), 585-610. https://doi.org/10.1016/S0039-3681(97)00012-5
- Bueno, O., & French, S. (2011). How Theories Represent. *British Journal for the Philosophy of Science*, *62*(4), 857-894. https://doi.org/10.1093/bjps/axr010
- Callender, C., & Cohen, J. (2006). There is no special problem about scientific representation. *Theoria: An International Journal for Theory, History and Foundations of Science*, *21*(1), 67-85.
- Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* New York: Oxford University Press.
- Clark, A. (2017). How to knit your own Markov blanket. In T. Metzinger and W. Wiese (Eds), *Philosophy and Predictive Processing (Vol. 3)*. MIND Group: Frankfurt am Main.
- Clark A (2019) Replies to critics: in search of the embodied, extended, enactive, predictive (EEE-P) mind. In: Colombo M, Irvine E, Stapleton M (eds) *Andy Clark and his Critics*. Oxford University Press, pp 266–303.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7-19. https://doi.org/10.1093/analys/58.1.7
- Colombo, M., Elkin, L., & Hartmann, S. (2021). Being realist about Bayes, and the predictive processing theory of mind. *The British Journal for the Philosophy of Science*, 72(1), 185-220. https://doi.org/10.1093/bjps/axy059

- Colombo, M., & Wright, C. (2021). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese*, *198*(14), 3463-3488. https://doi.org/10.1007/s11229-018-01932-w
- Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International journal of systems science*, *1*(2), 89-97. https://doi.org/10.1080/00207727008920220
- Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, *35*(3), 1-45. https://doi.org/10.1007/s10539-020-09746-2
- Di Paolo, E. (2009). Extended Life. *Topoi*, 28, 9–21. https://doi.org/10.1007/s11245-008-9042-3
- Di Paolo, E., Thompson, E., & Beer, R. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*, *3*. https://doi.org/10.33735/phimisci.2022.9187
- Facchin, M. (2021). Extended predictive minds: do Markov Blankets matter?. *Review of Philosophy and Psychology*, 1-30. https://doi.org/10.1007/s13164-021-00607-9
- Friston, K. (2009). The free-energy principle: a rough guide to the brain?. *Trends in cognitive sciences*, *13*(7), 293-301. https://doi.org/10.1016/j.tics.2009.04.005
- Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, *11*(2), 127-138. https://doi.org/10.1038/nrn2787
- Friston, K. (2012). A free energy principle for biological systems. *Entropy*, *14*(11), 2100-2121. https://doi.org/10.3390/e14112100
- Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475. https://doi.org/10.1098/rsif.2013.0475
- Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of physiology-Paris*, *100*(1-3), 70-87. https://doi.org/10.1016/j.jphysparis.2006.10.001
- Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, *193*(2), 559-582. https://doi.org/10.1007/s11229-015-0762-9

- Godfrey-Smith, P. (2003). *Theory and Reality: An Introduction to the Philosophy of Science*. Chicago: University of Chicago Press.
- Godfrey-Smith, P. (2009). Models and Fictions in Science. *Philosophical Studies*, 143(1), 101–116. https://doi.org/10.1007/s11098-008-9313-2
- Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.
- Hohwy, J. (2016). The self-evidencing brain. *Noûs*, *50*(2), 259-285. https://doi.org/10.1111/nous.12062
- Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger and W. Wiese (Eds), *Philosophy and Predictive Processing (Vol. 3)*. MIND Group: Frankfurt am Main.
- Hohwy, J. (2021). Self-supervision, normativity and the free energy principle. *Synthese*, *199*(1), 29-53. https://doi.org/10.1007/s11229-020-02622-2
- Keijzer, F. (2021). Demarcating cognition: the cognitive life sciences. *Synthese*, *198*(1), 137-157. https://doi.org/10.1007/s11229-020-02797-8
- Kirchhoff, M. D. (2018). Autopoiesis, free energy, and the life–mind continuity thesis. *Synthese*, *195*(6), 2519-2540. https://doi.org/10.1007/s11229-016-1100-6
- Kirchhoff, M. D. (2018). Hierarchical Markov blankets and adaptive active inference: comment on "Answering Schrodinger's question: a free-energy formulation" by Maxwell James Desormeau Ramstead et al.. *Physics of Life Review*, (24), 27–28. https://doi.org/10.1016/j.plrev.2017.12.009
- Kirchhoff, M. D., & Froese, T. (2017). Where there is life there is mind: In support of a strong life-mind continuity thesis. *Entropy*, *19*(4), 169. https://doi.org/10.3390/e19040169
- 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(138), 20170792. https://doi.org/10.1098/rsif.2017.0792

- Kirchhoff, M. D., & Kiverstein, J. (2021). How to determine the boundaries of the mind: A Markov blanket proposal. *Synthese*, *198*(5), 4791-4810. https://doi.org/10.1007/s11229-019-02370-y
- Kirchhoff M., Kiverstein J., Robertson I. (2022). The literalist fallacy and the free energy principle: On model-building, scientific realism and instrumentalism. *Br. J. Philos. Sci*. https://doi.org/10.1086/720861
- Kiverstein, J., & Sims, M. (2021). Is free-energy minimisation the mark of the cognitive?. *Biology & Philosophy*, *36*(2), 1-27. https://doi.org/10.1007/s10539-021-09788-0
- Klein, C. (2018). What do predictive coders want?. *Synthese*, 195, 2541–2557. https://doi.org/10.1007/s11229-016-1250-6
- Lyon, P. (2006). The biogenic approach to cognition. *Cognitive Processing*, *7*(1), 11-29. https://doi.org/10.1007/s10339-005-0016-8
- Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht: D. Reidel.
- McLendon, H. J. (1955). Uses of Similarity of Structure in Contemporary Philosophy. Mind, 64(253), 79–95. http://www.jstor.org/stable/2251045
- Menary, R., & Gillett, A. J. (2021). Are Markov Blankets real and does it matter? In D. Mendonça, M. Curado & S. S. Gouveia (eds.). *The Philosophy and Science of Predictive Processing*. Bloomsbury Academic (pp. 39-58).
- Nguyen, J. (2016). On the pragmatic equivalence between representing data and phenomena. *Philosophy of Science*, *83*(2), 171-191. https://doi.org/10.1086/684959
- Nguyen, J., & Frigg, R. (2021). Mathematics is not the only language in the book of nature. *Synthese, 198*(24), 5941-5962. https://doi.org/10.1007/s11229-017-1526-5
- Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

- Pero, F., & Suárez, M. (2016). Varieties of misrepresentation and homomorphism. *European Journal for Philosophy of Science*, *6*(1), 71-90. https://doi.org/10.1007/s13194-015-0125-x
- Pincock, C. (2004). A new perspective on the problem of applying mathematics. Philosophia Mathematica, 12(3), 135–161. https://doi.org/10.1093/philmat/12.2.135
- Raja, V., Valluri, D., Baggs, E., Chemero, A., & Anderson, M. L. (2021). The Markov blanket trick: On the scope of the free energy principle and active inference. *Physics of Life Reviews*, *39*, 49-72. https://doi.org/10.1016/j.plrev.2021.09.001
- Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2020). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*, *28*(4), 225-239. https://doi.org/10.1177%2F1059712319862774
- Rowlands, M. J. (2010). *The new science of the mind: From extended mind to embodied phenomenology*. MIT Press.
- Saigusa, T., Tero, A., Nakagaki, T., & Kuramoto, Y. (2008). Amoebae anticipate periodic events. *Physical review letters*, *100*(1), 018101. https://doi.org/10.1103/PhysRevLett.100.018101
- Suárez, M. (2003). Scientific representation: Against similarity and isomorphism. *International studies in the philosophy of science*, *17*(3), 225-244. https://doi.org/10.1080/0269859032000169442
- Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind.* Cambridge, MA: Harvard University Press.
- Van Duijn, M., Keijzer, F., & Franken, D. (2006). Principles of minimal cognition: Casting cognition as sensorimotor coordination. *Adaptive Behavior*, *14*(2), 157-170. https://doi.org/10.1177%2F105971230601400207
- Van Es, T. (2021). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior*, *29*(3), 315-329. https://doi.org/10.1177%2F1059712320918678
- van Es, T., & Hipólito, I. (pre-print). *Free-Energy Principle, Computationalism and Realism: a Tragedy*.

- Van Es, T., & Kirchhoff, M. D. (2021). Between pebbles and organisms: weaving autonomy into the Markov blanket. *Synthese*, *199*(3), 6623-6644. https://doi.org/10.1007/s11229-021-03084-w
- Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.
- Van Fraassen, B. C. (2008). *Scientific representation: Paradoxes of perspective*. Oxford University Press.
- Weisberg, M. (2006). Forty years of 'The strategy': Levins on model building and idealization. *Biology and Philosophy,* 21(5), 623-645. https://doi.org/10.1007/s10539-006-9051-9
- Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, 104(12), 639-659. https://doi.org/10.5840/jphil20071041240
- Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World.* Oxford: Oxford University Press
- Wheeler M. (2019). Breaking the waves. In M. Colombo, E. Irvine, & M. Stapleton (eds.), *Andy Clark and His Critics*. Oxford University Press (pp. 81-95).