Synthese (2021) 199:7801–7833
https://doi.org/10.1007/s11229-021-03140-5

Modelling ourselves: what the free energy principle reveals about our implicit notions of representation

Giovanni Pezzulo[1] · Matt Sims[2]

Received: 10 March 2020 / Accepted: 25 March 2021 / Published online: 13 April 2021

Abstract
Predictive processing theories are increasingly popular in philosophy of mind; such process theories often gain support from the Free Energy Principle (FEP)—a normative principle for adaptive self-organized systems. Yet there is a current and much discussed debate about conflicting philosophical interpretations of FEP, e.g., representational versus non-representational. Here we argue that these different interpretations depend on implicit assumptions about what qualifies (or fails to qualify) as representational. We deploy the Free Energy Principle (FEP) instrumentally to distinguish four main notions of representation, which focus on organizational, structural, content-related and functional aspects, respectively. The various ways that these different aspects matter in arriving at representational or non-representational interpretations of the Free Energy Principle are discussed. We also discuss how the Free Energy Principle may be seen as a unified view where terms that traditionally belong to different ontologies—e.g., notions of model and expectation versus notions of autopoiesis and synchronization—can be harmonized. However, rather than attempting to settle the representationalist versus non-representationalist debate and reveal something about what representations are *simpliciter*, this paper demonstrates how the Free Energy Principle may be used to reveal something about those partaking in the debate; namely, what *our* hidden assumptions about what representations are—assumptions that act as sometimes antithetical starting points in this persistent philosophical debate.

Keywords Free energy principle · Active inference · Predictive processing · Generative model · Internal representation · Action-oriented representation · Enactivism · Forward model · Markov blankets

"Hold it up sternly! See this it sends back! (Who is it? Is it you?)" Walt Whitman

✉ Giovanni Pezzulo
giovanni.pezzulo@istc.cnr.it

[1] Institute of Cognitive Sciences and Technologies, National Research Council, Rome, Italy

[2] School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Edinburgh, UK

'A Hand Mirror' (1860).

1 Introduction

There is growing consensus in computational and systems neuroscience around the idea that the brain is a *predictive machine*, which uses internal (generative) models to continuously generate predictions in the service of perception, action and learning. One theory that is centred around this idea and which is rapidly acquiring prominence—especially within the field of philosophy of mind and cognitive science—is *predictive processing* (PP) (Clark, 2013; Hohwy, 2013, 2020). The *Free Energy Principle* (FEP), a normative proposal, extends PP, providing it with a fundamental principle of adaptive self-organization (Hohwy, 2020). FEP and related predictive processing are rapidly acquiring prominence outside of neuroscience, especially within the field of philosophy of mind (Clark, 2013; Hohwy, 2013, 2020). Despite this fact, some of FEP's fundamental implications for epistemology and philosophy of mind remain widely debated. Perhaps the most prominent discussion concerns the representational or non-representational nature of FEP, with various proposals that FEP supports internalist (Hohwy, 2013; Kiefer & Hohwy, 2017), action-oriented (Clark, 2013, 2016), or enactivist and non-representationalist claims (Bruineberg et al., 2016; Gallagher & Allen, 2016; Kirchhoff et al. 2018; Kirchhoff and Robertson 2018).

Whether or not living organisms have (or need) internal representations is an old and fiercely debated topic (Merleau-Ponty, 1945; Ryle, 1949; Dreyfus, 1979; Newell & Simon, 1972; Fodor, 1975; Gibson, 1979; Thelen & Smith, 1994, Clark, 1998; Ramsey, 2007). Such debate is reiterated within the predictive processing view, from a more specific and mechanistic angle, i.e., assuming that FEP is a good model of living organisms, does it entail the notion of internal representation or not? Given that FEP has been implemented in a family of computational models that are by definition fully observable, it may seem paradoxical that it has been interpreted in so many ways. This is less surprising if one considers that FEP unites notions that are generally considered antithetical—most notably, the notion of internal (generative) model (von Helmholtz, 1866), which is usually associated with representationalist theories, and the notion of autopoiesis (Maturana & Varela, 1980), which is usually associated with non-representational enactive approaches. Can FEP help us advance, or even resolve, the long-lasting debate on internal representation in philosophy of mind?

Here, we will argue that, even if FEP cannot solve this debate, it can play an invaluable role in revealing our hidden assumptions about the very notion of representation and to create some common ground to discuss and negotiate them. Our general strategy here is to use FEP for the conceptual clarification of different notions of representation: we work backwards from either representational or non-representational interpretations of FEP's constructs to the various notions of representation that motivate those interpretations. It will emerge from our analysis that FEP has been (or can be) used to implement various kinds of computational models, which satisfy the requirements of certain theories of representation. Hence, the question of whether or not FEP entails representations depends on what notion of representation one uses in the first place.

The rest of this article is structured as follows. Section 2 briefly introduces FEP, its theoretical constructs and underlying assumptions. Section 3, after presenting some of the most representative interpretations of FEP in philosophy of mind, turns to look at how these interpretations highlight four distinct, yet overlapping, aspects of internal representations: *organizational aspects* (e.g., having some variable inside a system that is separated from that which it represents outside that system), *structural aspects* (e.g., having representational vehicles that are structurally similar to the state of affairs in the world that they stand in for), *content-related aspects* (e.g., having internal models that either encode environmental contingencies or sensorimotor contingencies; specification or description of how the world is taken to be in turn analysed in terms of correctness or truth conditions) and *functional role aspects* (e.g., supporting vicarious use before or in the absence of external events) of internal variables of a model. Section 4 briefly considers the evolutionary importance of functional role aspects and how such considerations favour representational versus non-representational interpretations of FEP. Section 5 focuses on explicating some of the relationships that hold between the four aspects of representation. We conclude by highlighting the heuristic power of FEP to advance our understanding of the notion of internal representation.

2 The free energy principle (FEP): a short summary

The Free Energy Principle (FEP) is an integrative proposal on adaptive self-organizing systems that remain far from thermodynamic equilibrium (Friston, 2010, 2019). It starts from the premise that, in order to survive, living organisms that engage in reciprocal (action-perception) exchanges with their environment must do what they can in order to remain within a neighbourhood of viable (i.e., physiologically-friendly) states which make up or comprise their phenotypes; an organism's phenotype being relative to its adaptive, ecological niche (e.g., a fish should never go out of the water). An organism's remaining within its neighbourhood of viable states means that it remains far from thermodynamic equilibrium. This idea is then cast in Bayesian terms, by assuming that the ecologically adaptive states constitute the *priors* of the agent (i.e., the states that an agent 'prefers' to visit become the states that it expects a priori to visit). Surviving then consists in ensuring that one keeps visiting the usual, non-surprising states despite environmental disturbances—where surprise is a measure of discrepancy between expected and actually occupied (i.e., observed) states. Hence, minimizing surprise or its long-term average (entropy) becomes a primary imperative for living organisms, as it permits them to counteract an otherwise unavoidable process of dispersion and loss of integrity (see Fig. 1).

However, minimizing surprise is challenging for a number of reasons, hence motivating a number of additional assumptions of FEP. First, the agent can only know about the (hidden) external states it encounters indirectly, via its sensations; hence it has to engage in (Bayesian) inference to infer the hidden states (von Helmholtz, 1866). In the same way, it can only indirectly modify the hidden states—by acting—and even its adaptive courses of actions (or policies) need to be inferred (Friston, 2011). Second, to perform the inferential steps that support perception and action, the agent needs to learn (hierarchical) generative models that essentially describe two things: what the agent expects a priori and how its sensations are generated based on externally postulated events and/or the agent's actions. It is by inverting the generative model that posterior inferences about hidden states (and expected action outcomes) are generated. It is only in virtue of the finetuning of the generative model (i.e., learning occurring at both ontogenetic and phylogenetic timescales) so as to achieve a continued degree of fit between it and the causal and statistical structure of the environment that inferences generated via model inversion allow the organism to bring about the conditions for its own existence. In other words, a generative model is entailed by both an organism's continued existence and its environment-regulating behaviour (Friston, 2011, 2012). This follows the "good regulator" theorem (Conant & Ashby, 1970), or the fact that in order to control the environment, an agent has to have (or to be) a model of the environment.[1] Essentially, then, this requirement creates a separation between the agent's *generative model* and the external dynamics (called *generative process*); we will discuss later the fact that this separation can be expressed using the statistical construct of Markov blankets.

Third, exact inference using the generative models is usually intractable because it would require having knowledge and computational resources that are not available to the agent; and hence the agent can only engage in approximate inference. The proposed approximate form of inference is called *variational Bayesian inference* and it rests on a number of assumptions that cannot be fully discussed here (Buckley et al., 2017). Under these simplifying assumptions, Bayesian inference becomes mathematically analogous to the process of variational free energy minimization, as studied in statistical mechanics. Minimizing variational free energy is roughly analogous (more properly, an upper bound) to the original problem of minimizing

[1] The difference between *having* a model versus *being* a model may seem puzzling. In this context, "being" a model (as opposed to "having" a model) implies that some aspects of the statistical and causal structure of the external world are statistically harnessed (e.g., over evolutionary timescales) by the dispositional activity of the agent's physiology and morphological structure, so that the features of an agent adaptively complement homeostatic- relevant aspects of the environment. One example is the fact that the visual system seems to include two streams, a dorsal and a ventral stream, which are specialized to process "where" versus "what" information (Ungerleider & Haxby, 1994). In Bayesian terms, this may correspond to the fact that the internal generative model supporting vision is factorized (i.e., segregated) into two processing streams, and this factorization mimics some "true" statistical independencies of the external world (i.e., the same object can be seen in different parts of the world; and the same part of the world can contain different objects). This factorization may be evolutionarily hard-wired, in such a way that an anatomical separation in the brain embodies (or "is", as opposed to "has") a model of statistical independencies in the world. The field of morphological computation offers some additional examples of how the body of living organisms is well adapted to (and in this interpretation, "is a model of") some important characteristics of the environment we live in, e.g., gravity. Body design can be leveraged to alleviate on even replace control demands, as in passive (brain- and control-less) walking robots (McGeer, 1990). For a recent experimental and computational demonstration of how a slime mould's hierarchical network morphology may be seen as modelling its nutrient environment at fast timescales see Kramar and Alim (2021). See Sect. 3.2 on potential limitations of *being* (as opposed to *having*) a model.

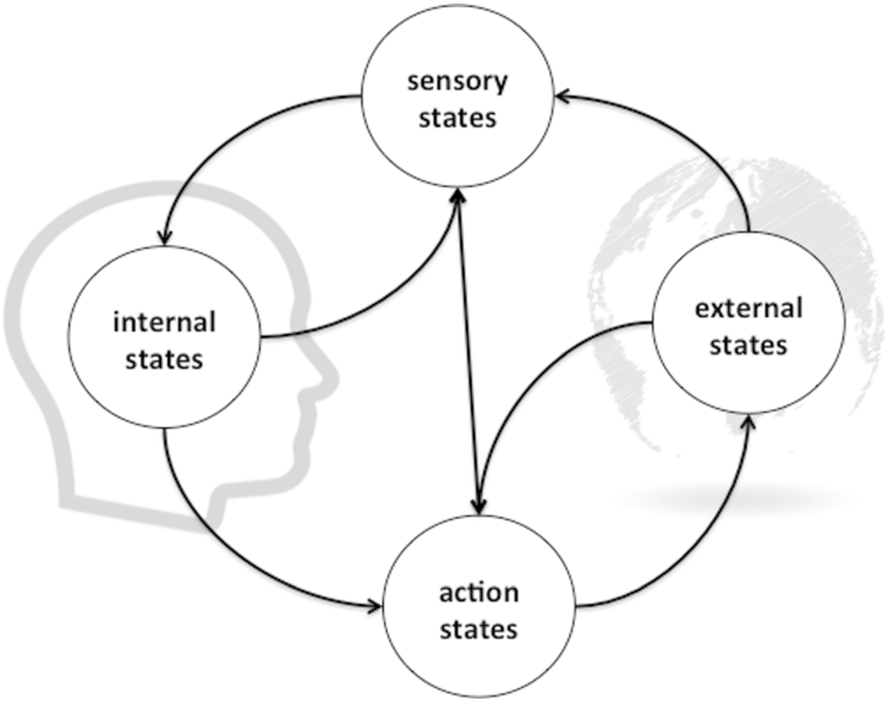

Fig. 1 Schematic of the reciprocal exchanges between an active inference agent (left) and the environment (right). Here states of the environment (hidden external states) are inferred via approximate posterior estimates (internal states) conditioned upon the activity of an agent's sensory receptors (sensory states). Internal states also infer the evolution of actuators (action states) so as to bring about the kinds of changes to external states that are consonant with the sensory states that an agent—given its phenotype—expects to observe

surprise but is tractable, based on the local information that the agent possesses. In variational inference, such information comprises the agent's *generative model G* and the agent's *recognition density Q*. The generative model includes two components: a prior and likelihood. The former is a summated distribution of the agent's expectations about environmental states while the latter is a conditional distribution specifying the agent's expectations of how environmental states map onto sensory observations. The recognition density Q, also described sometimes as a recognition model, corresponds to the approximate posterior estimate of (i.e., a Bayesian 'belief' about) hidden causes of sensory states[2]; it is a conditional distribution specifying the probability of environmental states given sensory observations.

Note that FEP actually involves two distinct but complementary processes of free energy minimization: the first, *variational free energy minimization*, corresponds to the optimization of perception and action, on the basis of present and past

[2] It is important to note that "belief" as used in FEP is a technical term denoting a probability distribution. These probabilistic beliefs should not be confused with personal level beliefs of folk-psychology.

information; the second, *expected free energy minimization*, supports the selection of action sequences (or policies), on the basis of future information—which by definition cannot be observed but only predicted (hence the term "expected"). Here, the idea is that an active inference agent should preferentially select policies that are expected to minimize more free energy in the future. Interestingly, *expected free energy* can be decomposed into two factors; the former measuring how distant the agent is from its preferred states and the latter measuring the dispersion or entropy of its beliefs, respectively. In other words, FEP assumes that adaptive action selection must necessarily have two components: a utility-maximization (or pragmatic) and an uncertainty-reduction (or epistemic) part. This formulation is distinct from (and subsumes) classical utility-maximization schemes in decision theory and has important implications for understanding human and animal behaviour (Friston, Levin, et al., 2015; Pezzulo et al., 2016).

FEP has a number of implications from theoretical and neurobiological perspectives. From a theoretical perspective, it suggests that the agent's imperative to survive (and avoid dispersion) can be mapped into the imperative of minimizing free energy; this formulation may work equally well at different levels of biological organization, from simple life forms (Friston, 2013; Friston, Levin, et al., 2015) to complex animals like ourselves and can even be applied to symbiotic associations (Sims, 2020), extended societies, and ecosystems (Clark, 2016; Ramstead et al. 2017). The aspect of FEP that has attracted more attention is the fact that perception, learning, and action all do the same thing—minimize free energy—but yet in different ways. Perception has a mind-to-world direction of fit: it operates by modifying internal states to make them more compatible with what is sensed. Learning has a mind-to-world direction of fit, too, but it operates at a different timescale from perception, by finessing the (parameters of the) agent's generative model rather than its current internal states. Finally, action has a world-to-mind direction of fit: it operates by modifying the external world to make it more compatible with what is expected. Hence, all aspects of adaptive self-organizing systems as described by FEP conjoin to minimize free energy.

Put in another way, active inference agents are "self-evidencing" (Hohwy, 2016). This is to say that because minimizing free energy over time is equivalent to maximizing Bayesian model evidence or "self-evidence" (Friston, 2013), in engaging in long term active inference agents seek out or generate those sensory states that "maximize the lower bound on the evidence for an implicit model of how their sensory samples were generated" (Friston, 2013 p. 2). As such, active inference agents author evidence for their own continued existence via free energy minimizing model optimization.

Finally, another remarkable aspect of free energy minimization is that it boils down to three basic (gradient descent) updates—for state estimation, action and precision[3]—that are consistent with known mechanistic features of brains and correspond to processes of perception, action selection and gain-modulation (attention in

[3] Precision is a technical term that denotes the inverse of the variance of a probability distribution (e.g., Gaussian).

perceptual domains and affordance in action domains), respectively. This means that one can use the normative principle of FEP to realize a "process theory" of cognition that has biological face validity and makes a number of important empirical predictions (Friston et al., 2016a).[4] PP is one such process theory.

In sum, the FEP is an integrative proposal on adaptive self-organizing systems that suggests living organisms manage to survive by forming internal generative models of the causes of their sensations and using them to minimize a measure of (roughly) surprise—or in other words, to ensure that they remain in the ecologically "adaptive" states that they should inhabit. The FEP is based on a number of assumptions—e.g., the fact that perception, action and learning minimize free energy—and it introduces a number of theoretical and mathematical constructs—e.g., the notions of generative model, generative process, priors, prediction errors and Markov blankets (see below)—which have attracted attention in philosophy of mind but have also been interpreted in different and sometimes antithetical (e.g., representational or non-representational) ways.

Below we review some of the most prominent philosophical interpretations of FEP, and in particular in relation to (strong and weak) representational versus non-representational arguments. We will argue that the different interpretations of FEP start from different meta-theoretical assumptions of what a representation is (or does), and the relative importance of its *organisational*, *structural*, *content-related* or *functional* aspects—which we will use as the axes of our review and conceptual analysis below.

3 What the philosophical debate on FEP says in relation to organisational, structural, content-related, and functional aspects of representations

Philosophical discussions—or "representation wars" (Clark, 2015; Downey, 2018; Williams, 2018)—about FEP have typically been focused on four main aspects of the notion of representation: *organisational, structural, content-related* or *functional.* These criteria overlap to some extent but are a useful starting point to organize a systematic review. Below we briefly summarize some of the most important arguments advanced in relation to these four topics, in both representational and non-representational camps.[5]

[4] A process theory is a theory that attempts to explain what a *system does* and hence 'how' a particular phenomenon comes about. This may be contrasted to a normative principle which describes what a system *should do* and hence 'why' a more general phenomenon should arise given certain assumptions. In the case of the FEP, this latter description takes the form the objective function of variational free energy (see Schwartenbeck et al., 2013). Because Free energy minimization over the long run approximates Bayesian inference, and because Bayesian inference is intrinsically normative, prescribing optimal belief update, FEP gains the status of a normative principle (Hohwy, 2020).

[5] There is a lot more that can be said about how the four aspects that we cover relate to various analyses of representations in the wider philosophical literature. Given limited space, however, we can only refer the interested reader to the work of both (Ramsey, 2007) and (Shea, 2018).

3.1 Focusing on organisational aspects of representation

The FEP rests on the idea that agents need to be separated from the rest of the environment, see Fig. 1. One can formalize this idea using the construct of a Markov blanket (Pearl, 1988).[6] While originally used in the context of graph theory to describe any set of random variables with a conditional dependent structure, in the context of FEP, the Markov blanket formalism is deployed to mathematically describe the boundaries of self-organized systems and their dynamic exchanges with the environment (Friston 2019; Wiese & Friston, 2021).[7] These system-environment interaction dynamics are specified in terms of four kinds of states and their relations: *internal states* of the agent, *external states* of the environment, and *action* and *sensory* states that mediate the reciprocal exchanges between internal and external states. Crucially, internal and external states cannot influence each other directly, but only via action and sensory states—which hence form a "blanket" that separates them. It is under this condition, when the agent's internal states are separated (or statistically independent) from external states, that the former appear to be *a model of* or *infer* the latter, and vice versa (Parr et al., 2020).

Take a cell for example. When applying the Markov blanket formalism to such a system its intracellular states may be cast as playing the role of internal states. Internal states influence and are influenced by external dynamics in virtue of the states of the mediating cell membrane (here, playing the role of intermediate states of the Markov blanket). It is the environmentally coupled activity of the states that are cast as the system's Markov blanket (e.g., cell membrane) which allow the cell to maintain its integrity in the face of environmental disturbances, while remaining separated (in statistical terms) from its environment.[8]

This organizational requirement—the fact that FEP requires internal states (i.e., states that encode the recognition model) that are statistically separated from the external reality—motivates one of the most prominent representational interpretations of FEP: the *internalist* view of Hohwy (2013, 2017) and Kiefer & Hohwy (2017). This view, in short, holds that because the internal states are segregated from the "external" world via the Markov blanket, and because the brain's activity may be cast as internal states, the brain must infer and represent what it does not have direct access to. Given that it is sensory evidence rather than the causes of that sensory evidence that the brain has access to, the thought is that it is only by inferring (explaining) the unfolding sensory states that the brain can infer the states of

[6] The notion of conditional independence of inner states from external states partitioned by a Markov blanket may be understood in terms of *information gain*; knowledge of the sensory and motor states is sufficient for predicting the behaviour of internal states of a system. And as such, there is no additional information about internal state behaviour that would be gained from knowledge of external states.

[7] See Bruineberg et al. (2020) regarding the current debate on whether or not Markov blankets—what they suggest calling 'Friston Blankets'—are best understood instrumentally or ontologically.

[8] It is important to note that the Markov blanket formalism is scale-free. As such, Markov blankets may be placed both at the level of subordinate systems 'inside' a superordinate agent and at the level of superordinate systems (e.g., systems of multiple agents or agent-niche systems) to which a subordinate agent belongs (see Palacios et al., 2020; Kirchhoff et al., 2018; Ramstead et al., 2017).

the hidden causes of those sensory states. This speaks of an inferential seclusion of the recognition model Q, where there is an "evidentiary boundary" between internal states (the recognition model) on one side, and environmental causes of sensory states on the other (Hohwy, 2016). Importantly, the location of this evidentiary boundary, Hohwy argues, "determines what is part of the representing mind and what is part of the represented world" (2016, p. 268). In other words, because the causal structure of the hidden environment must be inferred, perception and cognition are localized processes that are realized strictly by the activity of the central nervous system.

However, philosophers of the ecological-enactivist persuasion have approached the Markov blanket structural formalism in a manner that suggests a non-representational view of FEP (Bruineberg et al., 2016). They begin by noting that, falling out of the very notion of coupling, any two coupled systems must be separated by a Markov blanket (i.e., the notion of coupling requires at least two organizationally distinguished systems). The fact that partitioning of two coupled systems (e.g., the agent and its environment) can be described by deploying the Markov blanket formalism does not, however, necessarily imply that the behaviour of those systems is best explained by inferences generated by an internal model, the structure of which represents that which it is a model of. Bruineberg et al. suggest that the relationship between two such coupled and statistically partitioned systems is more parsimoniously understood in terms of their achieving high (relative) mutual information via the non-representational process of *generalized synchrony* (Huygens, 1673)—of the same kind that occurs between two coupled pendulums.[9] Despite the fact that one can formally describe each pendulum as a model of the other, actively inferring the other's behaviour, the question for the enactive-ecological views remains whether or not there is any added explanatory benefit of describing such inference (that falls out of the structural partitioning of Markov blankets) in representational terms (Kirchhoff et al. 2018). The alternative is framing the organizational separation that a Markov blanket introduces within dynamical systems theory—something that attempts to dispense altogether with representational constructs (Kelso, 1995; Port & van Gelder, 1995).

Yet there is a potential problem with this interpretation when it comes to accounting for the performance of *counterfactual* inference, something that is required for planning under FEP. Counterfactual inference requires engaging models that have some temporal depth, hence going beyond the immediate coupling to predict the consequences of future courses of actions or policies and—most importantly—selecting amongst them. Counterfactual inference provides sophisticated active inference agents (i.e., those with temporally deep models) with a higher degree of autonomy and disengagement from the current situation, which is necessary for adaptive action selection, beyond what synchronized pendulums can do (i.e., pendulums can synchronize, but they cannot select which pendulum they want to synchronize with). Furthermore, and importantly, during counterfactual inference (e.g.,

[9] For a criticism of Bruineberg et al.'s dynamical interpretation of inference, see Colombo & Wright (2018).

planning) the generative model is used to generate possible future observations—and the possibility to engage vicariously with "what is not present" has been considered a hallmark of representation at least since Piaget (1954); we will return to this point in Sect. 3.3.

Does counterfactual inference require recourse to representational states? If the answer is affirmative, then it seems that coupled generalized synchrony stops short of accounting for the behaviour of the kinds of adaptive self-organizing systems that are the targeted *explanandum* of FEP (we will return to representational interpretations of counterfactual inference in Sect. 3.4). On the other hand, if a non-representational account of counterfactual inference—or something that similarly explains adaptive action selection—is available, then something other than counterfactual inference must motivate a representationalist perspective on FEP.

An example of ecological-enactive (non-representational) approach to counterfactual inference is the theory of "adaptive active inference" of Kirchhoff et al. (2018). Adaptive active inference is the ability of adaptive self-organizing systems to selectively behave in response to environmental perturbation, actively seeking out the conditions that are compatible with their continued existence. Adaptive active inference is distinct from *mere* active inference (e.g., generalized synchrony of pendulums)—in that it requires agents to engage generative models with some amount temporal (i.e., counterfactual) depth, allowing for future-oriented selection over different courses of action. It is expected free energy minimization that is predicted to accompany particular action outcomes which drives future-oriented action selection. If something like Kirchhoff et al.'s account is tenable, then neither general considerations regarding organizational aspects nor more specific considerations regarding adaptive action selection seem to provide a sustainable motivation for a representationalist view of FEP. This being said, just how the kind of future-oriented posterior beliefs that drive adaptive active inference are accounted for without recourse to representation is not made explicit by these authors; if counterfactual inference is adaptive active inference, the task of explaining how the former can come about without the use of representations has just been pushed back onto adaptive active inference. Thus, the jury is still out on whether adaptive active inference lends itself to a non-representational account of future oriented action selection. Let us now turn to the structural aspects of FEP.[10]

3.2 Focusing on structural aspects of FEP

Rather than focusing on organizational aspects of FEP (e.g., the statistical separation described by Markov blanket formalism) one may instead focus on *structural* aspects of FEP. When adopting this perspective, the question to be addressed is whether generative models are structurally similar to their targets, or accurate descriptions of external reality, *or* whether models needn't be accurate but merely adequate enough (i.e., satisficing) to leverage for adaptive behavioural control.

[10] We are indebted to an anonymous reviewer for pushing us to clarify these points.

As discussed above, one aspect of FEP that lends itself to representational interpretations is the fact that active inference distinguishes between *internal states* and *external states*—and it is tempting to assume that internal states are about (or structurally similar to) external states. One can dissect the mechanisms of active inference in more detail to ascertain which ones are, or are not, structurally similar to environmental dynamics. Recall, active inference agents have two probability densities: the agent's internal generative model (*generative density* G) and the current estimate of the value of the hidden variables (*recognition density* Q). In principle, one can consider either (or both) of these densities as having representational aspects. For example, the generative density includes a *prior* term that describes the agent's preferences over observations and the recognition density encodes the *posterior probability*, or the agent's current best-guess of the causes its sensations, which is continuously updated during inference, and corresponds to what we have previously called the internal states of the Markov blanket. It is tempting to interpret these probability densities G and/or Q as the agent's representations of *desired* and *estimated* hidden states, respectively (Hohwy, 2013; Kiefer & Hohwy, 2017; Clark, 2016; Gladziejewski, 2016; Williams, 2018). This idea would link to a long tradition of identifying cognition with the usage of small-scale models or cognitive maps (Craik, 1943; Tolman, 1948)—where an internal representation is a *vehicle* that stands for, or represents, something different from it (possibly, some entity in the external environment).

Furthermore, one can argue that the generative model needs to have some *structural resemblance* with the generative process, in order to be useful for control purposes (Gladziejewski, 2016; Kiefer & Hohwy, 2017; Williams, 2018). In other words, without some form of structural and exploitable resemblance between controller and controlled system (independent of the actual correspondence between internal and external states), control would fail. Finally, predictions and prediction error signals could be considered as mental states with specific reference to external entities and their truth conditions. Successful predictions would therefore indicate a good representation of what is out there, whereas prediction error would indicate a *misrepresentation* that needs to be fixed by revising one's beliefs or by acting.[11]

There are however some counterarguments to the above representational interpretation. One line of argument is that one of the formal constructs required in active inference—specifically the generative density G—has been systematically misunderstood; and when understood correctly, it would turn out to be non-representational

[11] Kiefer and Hohwy (2017) suggest that the Kullback–Leibler (KL) divergence can be understood as a measure of misrepresentation. The KL divergence is a method of averaging over all the log ratios in probability vectors that lets one mathematically compare one probability distribution to a reference probability distribution. Optimally, one minimizes the KL divergence from the reference distribution. They suggest taking the KL divergence between the generative model's posterior distribution and the causal structure of the world as an internal proxy for the objective notion of misrepresentation (mismatch between the generative model and the world). See Kirchhoff and Robertson (2018) however who propose that the KL-divergence between prior and true posterior can be cast as a measure of relative entropy, which—over multiple timesteps—turns out to be a measure of covariance. This potentially blocks Kiefer & Hohwy's move from KL-divergence to structural (mis)representation.

(Ramstead et al., 2019). The authors suggest that generative models do not meet the requirements of structural representations, where some internal structure replicates some structure of the generative process. This being said, if one applies the same criterion, then the recognition density Q would however have properties ascribed to structural representations, as they need to have some "exploitable structural resemblance" with their targets in order to be leveraged for adaptive behavioural control (Ramstead et al., 2019). Indeed, the Q density would need to encode information (a posterior belief of states, including control states from which actions are selected) that is both exploitable for action guidance and shares structural features with environmental states. Interestingly, the authors go on to argue that this process of forming exploitable structural representations (i.e., posterior beliefs) would nonetheless have an enactive flavour, as it is not a passive perceptual inference but enabled by active inference.[12]

Another line of argument is that internal states in FEP would fail to accurately (at least to some degree) capture the structure of its target, the possibility of which some consider an important prerequisite for representation (Kiefer & Hohwy, 2017)." In some cases, accurate control requires internal states to *systematically misrepresent* what is out there (Bruineberg et al., 2016). For example, consider that an active inference agent's preferences are encoded as prior beliefs over future observations. Performing a goal-directed action requires agents to have a prior expectation that some goal state is met; that is, to misrepresent its actual state (note that other similar misrepresentations may be required for adaptive control, in addition to the aforementioned priors). In other words, under certain circumstances, the traditional requirements of representational systems (e.g., maintaining an accurate estimate of the external milieu) conflict with the control demands of FEP; one example is the fact that (for technical reasons) FEP agents need to maintain a sort of *optimism bias* in order to act adaptively. In these cases, control demands have priority, given that action—rather than perception—does the heavy lifting of free energy minimization (i.e., only action can reduce *surprisal* and make the world compatible with the prior). As a result, there are cases in which the agent's perceptual and state estimation processes become systematically biased (e.g., by the agent's goals). These models, which do not (always) maintain an accurate estimation of external variables, can be considered either representational—by appealing to the fact that temporary misrepresentation is a nuance of goal-directed behaviour—or non-representational, by noticing that accuracy of representation is constitutively less important than adaptive action (Bruineberg et al., 2016).

On the other hand, at least three features of FEP suggest that it entails some epistemic grasp (and not only a pragmatic grasp) over the external world. First,

[12] The idea that the posterior beliefs parameterizing the recognition density Q are structural representations and that they arises from a system's coupled activity with its environment is interesting for the following reason: since the Q density is encoded by internal states of the Markov blanket, it is possible that the kind of recourse to coupled dynamics—something which has traditionally been used to as an explanatory alternative to representationalism—from the perspective of organizational aspects (Sect. 3.1) could nonetheless imply elements of structural representation. We are indebted to an anonymous reviewer for pushing us to make this point explicit.

perceptual inference can be cast as a process in which the brain attempts to maximize the empirical adequacy of the (approximate) belief Q with respect to current sensory observations (Hohwy et al., 2008). This process is continuously optimized, and the model complexity actively regulated to avoid overfitting to sensory data (Hobson & Friston, 2012), which testifies that FEP carefully considers the necessity of maintaining some adaptive level of accuracy. Second, in FEP it is possible to select actions that serve properly epistemic functions, as they are aimed at exploring the environment to reduce long-term uncertainty about its causal structure (Friston, Rigoli, et al., 2015; Pezzulo et al., 2016)—as opposed to having pragmatic functions, such as obtaining a reward. Hence in FEP, epistemic imperatives are fundamental, over and above pragmatic imperatives. Third, on the PP view, the brain stores metacognitive estimations of the precisions of its incoming sensory signals, and adjusts those estimations to match environmental volatility, thus regulating the learning rate (Mathys et al., 2014). The best way of making sense of these three features is in terms of the brain (or cognitive system) actively maintaining the generative model, so as to maximize accuracy (or a certain level of accuracy-to-complexity ratio) and hence maintain some epistemic grasp on the external world.

Lastly, the evaluationist view provides yet another manner of thinking about the notion of representational accuracy. This view holds that a prediction error signal is an internally accessible evaluation or assessment of how ready or prepared the organism (or receiver at a given level in the processing hierarchy) is in relation to anticipated future environmental and organismic states in light of its most recent behaviour and goals. Prediction error signals relay to the organism "its state of readiness for goal-directed, adaptive activity" (MacKay, 1969)—that is, how well or poorly adjusted it is to how various states of the world appear to be unfolding given how it expects them to unfold.[13] Contextualizing the evaluationist perspective within the notion of intentionality or the "aboutness" of mental states as directed at the world (Brentano, 1874/1995) marks an important distinction with traditional views. Whereas traditional views take for granted aboutness-as-representation across the board, the proposed evaluationist alternative posits aboutness-as-evaluation for prediction error signals and accepts aboutness-as-representation for generative models.

3.3 Focusing on content-based aspects of FEP

The arguments for or against casting FEP as a representational framework presented in the previous section have been implicitly based on the question of whether internal states need to be somewhat similar to external states, or share some structural resemblance with them, to represent those states. Now let us turn to a closely related but distinct question that has been instrumental in considering the status of FEP as

[13] For further illumination, compare the evaluationist proposal with Millikan's (1995) pushmi-pullyu representations and Shea's (2012) use of them in this context. These are representations that "face in two directions at once" (Millikan, 1995, p. 186) in that they have two types of representational content, one indicative or descriptive and one directive or imperative (borrowing her familiar bee-dance example, "There is nectar at location X" and "Fly to location X").

representational or non-representational: whether the generative models need to explicitly model the ways external states produce sensations (aka environmental models) or the ways actions produce sensations (aka sensorimotor models).

To date, most published studies using active inference use generative models that closely mimic generative processes; or in other words, whose internal states (and their relations) are largely analogous to external states (and their relations). This is often done for illustrative purposes (Friston et al., 2016a) and few would assume that this condition could be met in biological organisms, where generative models and internal states can be at best a very impoverished version of external, environmental states and dynamics.

However, this raises a more profound question about the content of the generative models in FEP: is it really important that internal states *resemble* (closely or loosely) external states, or is it sufficient that—whatever they are—they *afford accurate action control*? This question, which concerns contentful aspects of representations, motivates a family of *action-oriented* theories, which de-emphasize the similarity between internal and external states, arguing that the most important requirement of models and internal states is affording accurate control—and this does not necessarily imply that they also provide a precise account of the world. In action-oriented and naturalistic models, predictions and prediction error signals are useful as bases for acting adaptively without necessarily representing external entities. These might be seen as pragmatic representations or representations of affordances (Pezzulo & Cisek, 2016). They are *about* some precondition for action (using the notion of aboutness of Brentano (1874/1995), or about one's relation with something external, but without necessarily representing it (nor does the object of thought need to exist).

One such influential action-oriented view, championed by Clark (2013, 2016), is that FEP (or predictive processing more specifically) is representational but in a sense that is more compatible with embodied, action, or control-oriented theories of cognition. Accordingly, although FEP requires generative models, these can fall within a range, spanning from rich to quite frugal models. Importantly, rather than maximizing the accuracy of representations, the role of generative models is to secure behavioural control; and hence they only need to incorporate those aspects of the external world and its "true" dynamics to the extent that such aspects satisfy control demands. As Clark puts it, "the role of such models is to control action by predicting and bringing about the complex plays of sensory data" (2016, p. 4). This means that in most practical cases, models do not need to encode deep aspects of the external world (e.g., a full 3-D reconstruction of the visual scene) but use information efficiently in the service of adaptive action. Action-oriented representations as such, are a means to engage the world, rather than to portray it in a way that is impartial to action. If this is correct, then the traditional view of models that see them *exhaustively* as world-depicting "mirrors of nature" (Rorty, 1979) divorced from action is a substantial oversimplification if not outright erroneous.

The embodied cognition literature has shown some practical examples of how good control can be realized using fast-and-frugal solutions and very simple models. One popular example is the "baseball outfielder problem", or the fact that catching a moving ball may not require a full model of the ball's position, velocity and direction (allowing for trajectory prediction), but a simpler control mechanism that only keeps the image of the ball stationary on the retina (McBeath et al., 1995). The idea behind Clark's (2013, 2016) action-oriented representational view is that it is possible to incorporate within FEP this and other lessons of embodied and ecological cognition, by considering *frugal* internal models, the main aim of which is managing adaptive agent-environment interaction as opposed to representing the external reality as accurately as possible (Cisek, 1999, 2007; Pezzulo & Cisek, 2016). A more nuanced view is that different kinds of models can coexist within different hierarchical layers of the same FEP agent, with lower hierarchical levels that encode relatively simpler and cheaper models (or heuristics) that can be contextualized by higher hierarchical levels to incorporate more sophisticated models or knowledge (Clark, 2016; Pezzulo, et al., 2015). This hierarchical view affords habit formation when lower hierarchical levels acquire sufficient precision or when environmental uncertainty is fully resolved (Friston et al., 2016b). Along similar lines, one can consider that extensive learning implies strong information compression, thus favouring the selection of behaviours that are both adaptive and cheap from an informational viewpoint (Globerson et al., 2009).

Another manner of approaching the question of whether resemblance of inner states to target states *or* the capacity of inner states to guide satisficing action control is most relevant to considerations of representational content begins from the notion of generative models. A formal requirement of FEP is having a generative model that describes (in statistical terms) the ways incoming sensations are generated. However, an FEP agent can "explain" or "model" the way its sensations are generated in at least two different ways: by appealing to entities or dynamics in the external environment or by only considering the effects of its own actions. In other words, an FEP agent can be equipped roughly with two kinds of generative models: *environmental* and *sensorimotor* models. The first kind of model—environmental models—captures the contingencies between a given external state or event (e.g., the presence of an apple) and the ensuing sensations (e.g., seeing something red, feeling something hard after grasping). Examples of this family of models include hierarchical perceptual models that represent simpler-to-more-complex visual features in their lower-to-higher hierarchical layers (Rao & Ballard, 1999; Lee & Mumford, 2003) and motor control models that are equipped with a hierarchy of motor representations, from long-term intentions to short-term motor programs (Hamilton & Grafton, 2007).

The second kind of model—*sensorimotor models*—captures the contingencies between the agent's actions (e.g., moving the finger forward) and the ensuing sensations (e.g., feeling something hard). These latter models code for sensorimotor contingencies (O'Regan and Noe 2001; Ahissar & Assa, 2016; Seth, 2014) relevant for behaviour, without necessarily hypothesizing external entities or providing precise accounts of the outside world. Consider for example an FEP agent (a simulated rodent), which senses the external environment through use of its whiskers. A sensorimotor model may encode (for example) the contingencies between the movement, velocity and angle of the whisker and the ensuing (expected) sensation, and a stimulation of the touch sensor. This model would afford some form of adaptive action, such as for example deciding whether to move forward (if no touch stimulation is sensed or expected following a whisking), change direction (in the opposite case) or jump (if some touch stimulation is absent and yet expected following a whisking with a certain angle)—without explicitly assuming external entities such as corridors, walls or holes in the ground. One can then consider that these are not models of external entities, but of one's own sensory-motor cycle—and that one actively samples sensations (e.g., the way whisking results in touch sensations).[14] See (Bastos et al., 2012; Friston, Rigoli, et al., 2015; Pio-Lopez et al., 2016) and (Friston, 2013; Baltieri & Buckley, 2017; Buckley et al., 2017) for more detailed examples of the two kinds of models, respectively.

One immediate question that may be posed for sensorimotor models is how they might afford sophisticated cognitive processing? On the one hand, one can argue that the notion of sensorimotor model can stretch to any kind of complexity (Bruineberg et al., 2016), as sensorimotor loops become able to incorporate increasingly distal consequences. Furthermore, the sensorimotor dynamics created by sensorimotor models may be progressively *internalized* to support mental operations detached from the sensory-motor cycle (Hesslow, 2002; Ito, 2008; Buzsáki et al., 2014; Pezzulo, Donnarumma, et al., 2017; Pezzulo, Kemere, et al., 2017; Pezzulo et al., 2020; Stoianov et al., 2020).

An alternative proposal is that sensorimotor models pave the way to increasingly more sophisticated models. To give an account of this, one approach which dates back at least to *schema theory* (Drescher, 1991; Roy, 2005), consists in starting from *sensorimotor* (action-effect) models but progressively extending them to incorporate extra variables that describe external causes of sensations (e.g., objects)—which therefore become akin to *environmental* models. The key to this form of schema learning is postulating external objects as the common cause of several action-effect pairs. For example, one can postulate the existence of a "coffee cup" because it explains several sensorimotor processes (e.g., lifting X results in feeling something light-weighted; drinking from X, results in something hot; looking at X from various positions results in seeing something cylindrical; grasping X results in feeling something cylindrical, etc.).[15] Schema learning thus constitutes a way to gradually pass from sensorimotor to environmental models, and to let them co-exist afterwards. In contrast to enactive views of sensorimotor contingencies, in schema learning the common cause of coherent sets of contingencies is reified in the form of a novel set of internal variables that are partly action-independent—hence producing environmental models.

With this being said, how might the distinction between *environmental* and *sensorimotor* internal models allow for a more detailed understanding of the different representational and anti-representational perspectives on FEP? In principle, both

[14] One interesting question that might arise is how this distinction between environmental and sensorimotor models maps onto the two-visual systems account of perception and action, according to which the anatomical branching of the visual system into the ventral or 'what' and dorsal 'how' pathways underwrite semantic and pragmatic processing (see Goodale & Milner, 1992).

[15] How sensorimotor schema learning leads to object postulation is clearly expressed by Drescher when he writes: "The coordination of hand motions and eye motions, of seeing and feeling, begins to describe the nature of objects and space; sight and touch begin to be known as coordinated properties of external objects" (2002, p. 135).

environmental and *sensorimotor* models can be interpreted in representational and anti-representational terms. Following a long tradition in cognitive science (Craik, 1943; Tolman, 1948) environmental models have been often interpreted in representational terms, given that they play the role of vehicles that refer to external entities, and they include internal (hidden) variables bearing some *similarity* to the external entities, e.g., increasingly complex features of an apple or a face for perceptual models. However, it is possible to assign different representational status (representational versus non-representational, or strong representational versus weak or minimal representational) to internal dynamics that accurately encode the key aspects of external dynamics (the generative process) and to more frugal environmental models; for example, models that afford catching a ball by modelling its trajectory or by maintaining a stable visual angle. The latter (frugal) kinds of model do not necessarily lend themselves to representational interpretations. It is not even necessary to interpret them as "models" in the sense specified by FEP: at least in certain conditions, the FEP scheme can include simple stimulus–response mechanisms (Friston et al., 2016b).

Sensorimotor models have also been interpreted as being either representational or non-representational. For instance, from radical enactivist or ecological perspectives, sensorimotor models do not involve the use of representations; they mediate the reciprocal exchanges between the agent and the environment without involving internal representation of the latter (Bruineberg et al., 2016; Gallagher & Allen, 2016). Moreover, when taking the radical enactivist or ecological perspectives on FEP, the notion that FEP agents *are endowed with* sensorimotor models (or environmental models for that matter) may be interpreted more broadly as the claim that the agent *is* a model of its environment. In other words, the agent's phenotype embodies evidence of the very environmental dynamics that it and its progenitors over both phylogenetic and ontogenetic timescales have successfully adapted to. Since being a model of the environment does not imply any need to represent the environment, the central place that models play in FEP does not itself entail a representationalist view of FEP. One can push this argument even further, and argue that the statistical notions (e.g., Bayesian inference, estimation) used by FEP are stretched beyond explanatory usefulness given that FEP applies just as equally to non-neural bacteria as it does to more complex organisms with nervous systems. Descriptions of adaptive behaviour which make use of inference as such are thus seen as uninformative and may be more fruitfully replaced with explanations involving the idea that certain states of the system become dynamically coupled with certain states outside the system via generalized synchrony (Bruineberg et al., 2016; Korbak, 2019).

Other researchers have taken a representational perspective when interpreting sensorimotor models. A first line of argument, as discussed above, focuses on organizational aspects of internal models within the action-perception cycle of FEP. In this perspective, all FEP agents are representational, because their internal states are segregated from the states of the external milieu via a Markov blanket and the generative models are deployed in inferring the causal-probabilistic structure of external states (Hohwy, 2013). It does not really matter whether the internal models are minimal, given that even the most accurate models that one can think of are nevertheless simplifications of the "true" state of the world—after all, any models is just a useful simplification that affords estimation, prediction and control. If simpler models do a good job, they must be preferred given standard statistical arguments (model selection) that are part and parcel of free energy minimization (FitzGerald et al., 2014).

A second line of argument in favour of the representationalist interpretation of sensorimotor models is that they afford forms of internal manipulation and inference—for example, when they are used to steer sensorimotor predictions of *action consequences*—and for this they are sometimes called "action-oriented" or "embodied representations" (Clark & Grush, 1999; Grush, 2004; Pezzulo, 2008). This line of argument is in agreement with many foundational works in cognitive science, according to which internal representations are important because they can be used prior to or instead of acting in the real world—and even "vicariously", in the absence of their referent (Craik, 1943; Piaget, 1954).

This line of argument brings into focus to yet another well-acknowledged aspect of internal representations (that is distinct from organizational, structural, or content related aspects): their functional role, or what they do (as opposed to what they encode) within a cognitive architecture.

Let us now look more closely at this particular aspect of representations.

3.4 The functional role of internal models in FEP and the importance of vicarious operations and detachment

Another way to address the issue of what a representation is and what it is not is by asking what functional role representations play within a hierarchical architecture. An idea that dates back at least to Piaget (1954) is that representations should vicariously "stand for" something external in its absence and afford vicarious operations, i.e., mental operations using an internal vehicle that are executed before acting on the external referent of the vehicle (e.g., consider mentally which route to take to go home), or even when the external referent of the vehicle is absent (e.g., consider mentally whether one would enjoy eating a pizza—with no pizza is in sight).[16] Relatedly, this functional role aspect of representations has been expressed in terms of whether the agent's internal operations are detached from its action-perception cycle and hence autonomously generated, versus determined or sustained by external stimuli (Gardenfors, 1996; Pezzulo & Castelfranchi, 2007).[17]

From this perspective, to assess the representational status of FEP one could ask "what functional roles do internal models play during free energy minimization, and whether or not such minimization requires the internal manipulation of variables in ways that resemble vicarious operations in classical (e.g., Piagetian) accounts of representation?". Below we address these questions, in relation to the two forms of free energy minimization that occur in FEP: we focus first on *variational free energy* minimization (that depends on current and past observations) and then on *expected free energy* minimization (that depends on future, predicted observations).

[16] See also Tolman (1948) for a similar idea.

[17] See (Sims 2019b paper) however for a case against the idea that all off-line internal operations are decoupled across the board.

In FEP, *variational free energy minimization* starts with an internally generated prediction (or, in a hierarchical architecture, a cascade of predictions). Importantly, not only are predictions self-generated and hence at least temporarily autonomous from current stimulations, but they are also used for purely internal (prediction-matching and error-correction) operations. Specifically, the inferential scheme of FEP requires the continuous generation of two kinds of prediction errors. The first is a discrepancy between current expectations (about hidden states) and the available sensory evidence (or, in a hierarchical scheme, hidden states at lower levels). The second kind of prediction error is a divergence between current expectations (current posterior estimate of the state of hidden variables in the recognition density Q) and prior preferences (in the generative density G). The two densities are continuously compared against each other internally to the FEP agent (and in the second case, fully independent of current external stimuli). One can think of the elicitation of such expectations and the subsequent matching operations as vicarious operations, which use internal variables of the model as proxies for external entities (in the case of environmental models) or for possible actions (in the case of sensorimotor models). Hence, to the extent that these algorithmic steps of variational inference (and others, e.g., calculating the entropy of hidden states) are considered part and parcel of a process model of cognition (Constant & Clark, 2019), these internal manipulations would count as representational (Clark, 2016). However, a possible counter-argument is that the algorithmic steps of variational inference are just one way to implement FEP; and one can realize roughly the same computations by appealing to other methods that do not require this vicarious use of internal variables—for example, the synchronization between variables within and outside the generative model (Bruineberg et al., 2016; Gallagher & Allen, 2016).[18] Furthermore, one may consider that calculating an expectation is not a sophisticated process but simply boils down to (for example) calculating the mean of a probability distribution, which does not require a degree of cognitive complexity. However, the cognitive complexity increases in the case of planning and expected free energy minimization, as we describe next.

In FEP implementations of planning, candidate action sequences or policies are compared depending on their *expected free energy*; that is, the free energy that the agent expects in the future by executing the policies (Friston, Rigoli, et al., 2015). The expected free energy (or quality) of policies is calculated by internally simulating sequences of actions between action-perception cycles, without selecting actions or receiving observations in the meantime. Hence, the expected free energy calculations require an off-line usage of the generative models, to predict (action- or policy-conditioned) future states and observations—both of which by definition cannot be observed. This self-fuelling, what-if or counterfactual internal inference

[18] See (Korbak, 2019), who in defending a computational—enactivist view of FEP, argues that eliminativism with respect to algorithmic variational inference fails at certain levels of complexity at which dynamic coupling (generalised synchronicity) cannot sufficiently account for certain phenomena. And for a contrasting view that at progressively higher levels of complexity, variational inference will be considerably to be intractable without taking the agent's being environmentally embedded into account, see (van Rooij et al., 2019).

(Seth, 2014) is thus operated in the prolonged absence of the stimulus and entails a stronger detachment of the agent from action-perception cycles compared to variational free energy.[19] It therefore defies interpretations of belief updating in FEP based on coupling between internal and external states, as this coupling is broken. It also defies interpretations based on the idea an agent *is* a generative model of its environment, as opposed to the idea that an agent *has* a generative model of its environment. This is because the generative model has to be explicitly used to generate predicted states and observations. Hence the planning mechanism of FEP lends itself more easily to representational interpretations, which is in keeping with related ideas about internal, what-if simulations in computational neuroscience (Jeannerod, 2006), philosophy of mind (Clark & Grush, 1999), machine learning and robotics (Nishimoto & Tani, 2009).

In sum, the possibility to support detached operations that de-couple an agent from the current action-perception cycle has been often considered as a strong criterion for representation, in the Piagetian sense of vicarious operations, in the absence of a referent. When assuming this perspective, representation links to a functional aspect—vicarious use, momentary or prolonged detached processing—rather than organizational, structural or content-related aspects of internal models.

A possible counterargument to the idea that vicarious functions require representation is that they consist in a reuse of perceptual and motor capacities; and if these are not representational in the first place (as enactivists assume) they do not become representational just because they are reused (Hutto & Myin, 2017). There is however a problem with this rebuttal. In active inference, vicarious operations like planning require something more than a verbatim reuse. Indeed, the agent's generative model is used differently during action-perception and during planning: in the latter (but not the former) case, it is used to predict observations to feed the (expected) free energy calculations. In other words, while during the action-perception loop it is the environment that closes the loop by feeding the necessary observations to the agent (Fig. 1), during planning it is the agent's generative model that closes the loop and feeds the observations, therefore acting vicariously to replace (unobserved) environmental dynamics. In principle, one could consider the second usage of generative models as representational even if the former is not (as enactivists assume).

More broadly, one can consider the idea that "a process that starts non-representational cannot become representational" to be valid only if one focuses on *content-related* aspects of representation, but not if one focuses on *function*. In the same way that a perceptual process which does not start as a memory can become a memory afterwards, also a perceptual process that does not start as a representation may become a representation afterwards (e.g., if it is used to perform a vicarious operation "in the absence" of the initial referent). We will now turn to a related manner of thinking about function that may be used to support the claim that non-representational processes may become representational processes.

[19] For an argument that this kind of counterfactual decoupling is a marker of *bona-fide* cognition see Corcoran et al. (2020).

4 Representation as an evolutionary function

Similar to other cognitive functions like working memory, planning, cognitive control and attention; or perhaps, even to functions like flying or swimming (Cisek, 2019), representation—here intended in the sense of affording vicarious operations and detachment from the current action-perception cycle—may be an *evolutionary function*. The claim that some current trait is an evolutionary function[20] is to say that it was selected for and is now present in some organisms because in the past (or more specifically—in its recent selection history) it endowed those organisms possessing it with more adaptive advantage than those failing to have that trait (Godfrey-Smith, 1994a, b, 1996); the general notion of adaptive advantage expressing itself in the fact that organisms, because of their possessing the trait, were able to better respond to niche-specific selective pressures (e.g., avoiding predictors, competing for mates, competing for food, avoiding sickness, etc.) and, as a result, to pass on that advantageous heritable trait. In suggesting that the ability to deploy detached operations on internal variables is an evolutionary function, we are suggesting that the fact that such representational capacities are found today in some organisms can be explained (partially) by the selection pressures that the capacity of detached representation allowed such organisms to successfully respond to. Any detailed speculative account however as to what the presence of detached representational capacities in certain organisms suggests about the kinds of environmental problem spaces that they have responded to in the past falls beyond the scope of this paper (but see for example Millikan, 1989, 2004; Sterelny, 2003; Schulz, 2018; Corcoran et al., 2020).

Under the assumption that detached representation is an evolutionary function, being able to coordinate current actions with the effects of future actions seems likely to have placed those representing agents that occupied complex niches at an advantage over those failing such a capacity with respect to dealing with structured environments and predictable future events. Where complex niches might be thought of as having a high level of "structured heterogeneity" (Godfrey-Smith, 1996); those niches in which the cause-effect cycles that are relevant to a particular kind of organism unfold at long timescales such that current events may act cues for identifying

[20] Strictly speaking, the notion of *evolutionary* function is distinct from the kind of *instrumental* function which has been referred to as a criterion for representations above. In particular, instrumental functions describe the functional role played by a process in reference to the larger system (or process) that it is part of and helps to explain (Cummins, 1975). When it is said that representations play the functional role of supporting vicarious operations, it may be assumed that this role is relevant to the larger cognitive prediction error minimizing system. Evolutionary functions, on the other hand, are specified in terms of those features of organisms that have gradually arisen due to their adaptive contribution to an organism's fitness in the presence of selective pressures. The two claims that (a) the functional role of representations (models) is to allow for vicarious operations and that (b) representations as such have been selected in given the advantage that organism deploying them (historically) had in dealing with selective pressures are two distinct yet mutually supportive kinds of functional claims—as one may consider, for example, that performing vicarious operations gave organisms strong evolutionary advantages in certain niches, see the main text. For more on the distinction between evolutionary functions and instrumental functions see Godfrey-Smith (1994a, b, 1996).

and planning current action according to regular future manifestations-of organism-relevant selection pressures. Moreover, the evolutionary function of detached representation of one's own action consequences (or those of some other agent) may be seen as a particularly important adaptive benefit for social organisms, where the same behavioural cue may lead to distinct future outcomes, some of which are more supportive to the individual's survival than others (Pezzulo, Donnarumma, et al., 2017). Relatedly, the ability to represent the consequences of a given action seems to be a necessary cognitive function for being able to *actively construct* niches, the group or social environment being just one such constructed niche domain; the construction of social niches helping to ensure that selection pressures are dealt with effectively at both the individual and group level (Lewontin, 1983; Odling-Smee et al., 2013; Bruineberg et al., 2018; Constant et al., 2018).

This perspective automatically entails that representation is an attribute of biological organisms that are subject to evolutionary pressures and does not readily apply to non-biological entities (e.g., a stone does not represent gravity). However, like any other functional attribute, it needn't apply to all biological organisms—or at least to the same degree. As such, although FEP may be used to describe the self-organizing behaviour of all biological systems, not all self-organizing behaviour of biological systems requires the use of detached operations on vicarious variables. For example, homeostatic regulation may be more parsimoniously understood without evoking the notion of representations (e.g., in terms of simple reflex arcs) while the future-oriented (and hence detached) character of allostatic control in the service of essential homeostatic setpoint maintenance may indeed call out for representational explanation (Cf. Stephan et al., 2016)[21] Thus, we would like to suggest that understanding the functional role aspect of representation in the context of evolutionary function offers a manner of supporting the claim that a process that does not start as a representation may become a representation afterwards: if the functional role of a representation is to allow detached operations on vicarious variables, and it was the use of states playing this particular role that substantially contributed to the differential fitness of certain classes of living systems inhabiting increasingly complex niches, then it is plausible that the processes underpinning this function are an extension of simpler, non-representational processes that have themselves been complexified over evolutionary timescales. If this is the case, then not only can a process that does not start out as a representation become a representation at short timescales (think use/re-use accounts), but we should expect as much given some kind of evolutionary continuity between putatively non-representational processes and those detached processes that have evolved from them. In other words, full-fledged decoupled representational capacities might "shade-off" (Godfrey-Smith, 1996) into other cognitive or minimal cognitive processes, the dynamics of which are increasingly

[21] Allostatic control refers to anticipatory behaviour that by enslaving wide-range homeostatic variables allows systems to avoid deviation from essential homeostatic setpoints prior to the onset of deviation (for various and yet somewhat related views of allostatic control see Stephan et al. (2016), Corcoran et al. (2020), Kiverstein and Sims (2021).

more coupled to the environment through action and sensory feedback.[22] Therefore, when taking functional role as contextualized by evolutionary function as a criterion for identifying when recourse to representation is warranted in explanation, it is plausible to interpret at least some processes (i.e., those that involve the use of vicarious variables in the service of minimizing expected free energy) that drive the kind of self-organization that FEP describes as representational.

5 What can we learn from this debate?

This review of some of the most prominent philosophical interpretations of FEP has highlighted a wide variety of opinions. We briefly summarized that theories starting from different aspects of representations (e.g., organizational, structural, content-related or functional) focus on different constructs of FEP (e.g., Markov blankets, prediction-error-matching, environmental or sensorimotor models, vicarious use of internal models) and come to different conclusions about the representational or non-representational nature of FEP.

However, one can use the debate on (representation in) FEP the other way around: not as a way to resolve the issue at stake, but as a "mirror" to look at one's own implicit notions of representation. Figure 2 provides a simplified scheme to consider how, by starting from different implicit notions of and criteria for representation, one can arrive at different conclusions about FEP (or anything else). Note that the Figure is not meant to be prescriptive. The directions indicated by the arrows are not mandatory; they just reflect the most common pathways that we have discussed in the previous Sections, but of course alternative pathways are possible.

Representational interpretations tend to emphasize the importance of having internal generative models and environment-mediating internal states that are (statistically) separated from external states. One can interpret this separation from an internalist perspective by emphasizing that internal states are separated from the external milieu by a Markov blanket; or from a more action-based and embodied perspective by noticing that internal models can be frugal; their main role is mediating adaptive action and they needn't necessarily "resemble" external events—providing that they offer a good guidance for action. This latter argument applies particularly well to FEP agents using sensorimotor (as opposed to environmental) generative models that encode the contingencies between actions and perceptions, without necessarily appealing to external entities to explain sensations.

[22] For example, the capacity to distinguish an object from its surrounds based on some of its features (e.g., colour and shape) when an object is visually present may have evolved into a capacity to also imagine objects when they are not visually present (Pezzulo, 2008) This is exactly what should be expected from a generative model which is a machine that operates according to "analysis by synthesis" and hence is able to re-create (i.e., generate) structure. The passage to a full-fledged representational role in our opinion is not just when the object features are re-created, something which is always part of generative modelling, but when they are also successively used within some mental operation—for example, to ask the question: "do you like the chair we saw at shop 1more than the one we saw at shop 2?".

If one only appeals to the organizational aspect of representation, the presence of environmental or sensorimotor (or complex or frugal) model does not matter, insofar as the internal variables of the model are understood to be separated from external reality by a Markov blanket and the generative model is leveraged to infer the causal structure of external reality via self-evidencing. However, these differences matter if one considers structural aspects, and the degree of resemblance between hidden variables and environmental dynamics (as opposed to action or information gathering dynamics). Here, the content of representations is used to draw a further distinction (within the representational view) between an internalist (sometimes also called *intellectualist* or *encodicist*) versus an action-oriented perspective. Thus, this debate within FEP reiterates classical discussions between traditional versus action-oriented and embodied theories of cognition, which emphasize the representation of environmental regularities versus sensorimotor contingencies, respectively—although not all action-oriented theories appeal to the notions of internal models or representations; for a detailed discussion see (Pezzulo et al. 2016; Engel et al., 2013, 2016).

It is also possible to interpret FEP agents—and especially those having sensorimotor models—in non-representational terms. Non-representational interpretations seem more compelling for FEP agents that use sensorimotor models and under the assumption that some aspects of FEP—organizational (Markov blankets) and algorithmic (variational inference)—are neither mandatory nor relevant to the debate on representation. For example, it is worth noting that FEP agents make vicarious usage of internal variables before (or instead of) acting in the external environment in two situations: during on-line action-perception loops to calculate predictions and prediction errors, and off-line to propagate these predictions over time in order to covertly evaluate alternative policies or to support planning. For on-line action-perception loops, one can appeal to the fact that some algorithmic aspects of variational inference can be replaced by non-representational mechanisms (e.g., synchrony with external stimuli). However, a major challenge for non-representational interpretations of FEP (when used with a process theory) is to offer mechanistic models of detached cognitive abilities such as planning. Current active inference implementations of planning require the off-line engagement of generative models to calculate expected free energy, with a significant degree of detachment from the current action-perception loop. To the extent that detachment and vicarious use of internal variables for counterfactual inference are elevated to criteria for representation or sophisticated cognition, they would support a representational view of FEP—independent of the content or complexity of the generative models. The opposite would be true if one does not consider detachment as a criterion for representation.

One important implication falling out of this diagnosis is that when considering functional role aspects, it is often how the details of our chosen process theories are fleshed out and contextualized by the kinds of cognitive phenomena that we are attempting to account for that skew our interpretation of FEP in one direction or another. For example, one may consider that there are core aspects of FEP, such as the possession of a Markov blanket and more ancillary aspects, such as the possibility to engage in counterfactual inference (which is only required for planning)—and it is only the latter, more ancillary aspects that call for a representational

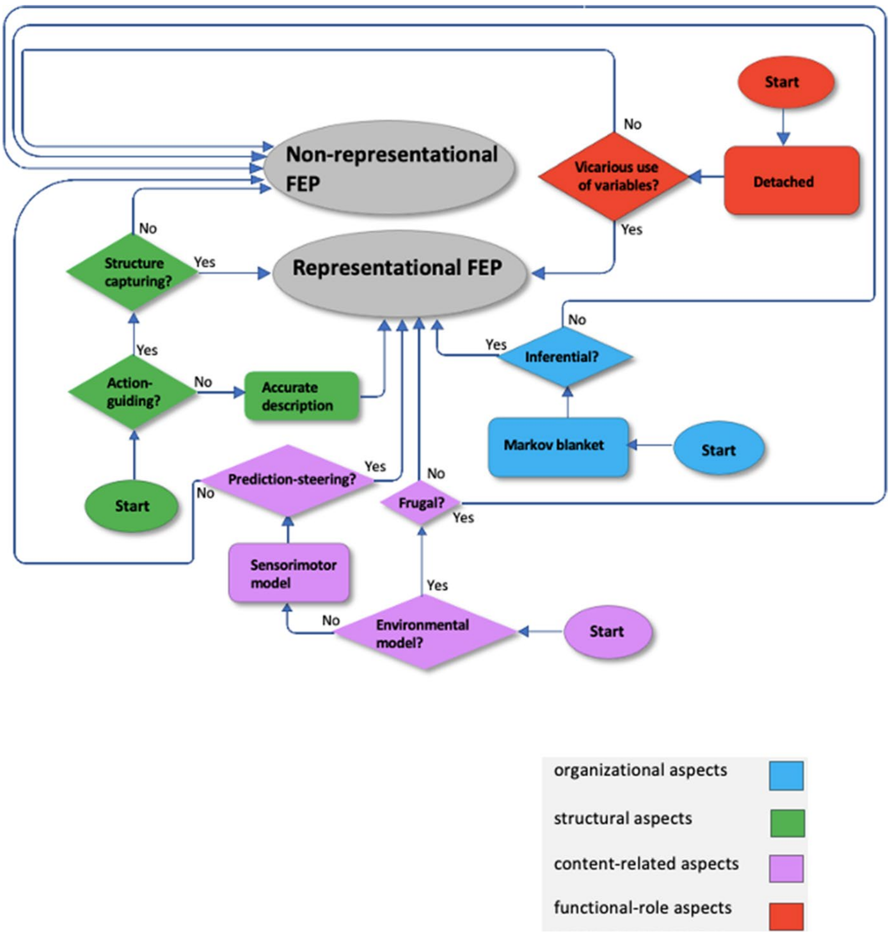

Fig. 2 A conceptual map of possible pathways from different premises to different conclusions about the status of FEP as either representational (e.g., internal states represent external states) or non-representational (e.g., internal and external states have coupled dynamics, but the former do not represent the latter). "Internal states" and "external states" are the terminology used in Fig. 1

interpretation under a given process theory. This would imply that (when using functional role as a sole criterion for representational processes) only some FEP agents—namely, those that can engage in counterfactual forms of inference—would meet the criteria for representation. It is only this subset of FEP agents that would be equated to full-fledged PP agents.

This possibility is voiced by Clark (2017), who suggest that many basal organisms (e.g., bacteria) may minimize free energy and yet fail to have deep generative models that allow for the kind of counterfactual inference he views as being core to PP. If this is the case, then representations may be something that are only available to those systems that engage in the functionally asymmetric, hierarchical, bidirectional message passing intrinsic to PP. Such a view is compatible with either of the following two claims: (1) when used with PP or without PP, FEP is non-representational; (2) at a certain level of increased complexity (i.e., meeting the increased hierarchical generative model depth that PP demands (Clark, 2017)) FEP becomes representational. It is only claim 2 which supports the assumption that FEP's representational status piggybacks upon the status of its process theory. Claim 2 however goes far beyond any of FEP's core assumptions or constructs.

Notice, however, that it is possible to adjudicate the representational status of organizational, structural, and content related aspects independently of the details of a given process theory. Because Markov blankets, recognition models, and generative models are non-negotiable constructs of FEP, the role that they play in free energy minimization is not tied to any specific cognitive *explanandum* or process *explanans*. This is one reason why FEP may be used across the (biological) board as a normative principle for all adaptive self-organizing systems at non-equilibrium steady state (Cf. Kirchhoff, 2016). The dynamic interaction of an *E. coli* with its milieu, for instance, may be modelled under FEP using the Markov blanket formalism to predict how it adaptively self-organizes despite the influence of random fluctuations. Whether or not such bacteria engage in something so specific as planning is not restriction on using FEP to model *E. coli*'s adaptive behaviour. When the organizational aspect is used as a representational criterion, then the debate about FEP's representational status is entered into from the beginning by considering how to best interpret the relationship between blanketed internal states and hidden states. The specifics of active inference (or PP), or how to account for particular cognitive phenomena are external to the debate at this more fundamental level where general principles of self-organizing adaptive behaviour are being applied to even the simplest of living systems.

It is also interesting to note that placing emphasis upon one or the other of these non-negotiable constructs of FEP results in making representations a biologically pervasive feature of all FEP agents. For example, when understood as involving inference generating, self-evidencing internal models, the organizational criterion suggests that representation is a ubiquitous feature of any biological FEP agent that maintains its structural (and functional) integrity. This kind of biological pervasiveness of representation may of course not appeal to those theorists who would prefer to reserve representations for the sophisticated cognitive capacities of "complex" creatures. Avoiding a wide scope view of representations would seem to motivate some participants in the debate to opt for the functional role criterion and, if what we have argued is correct, doing so may very manifest itself in terms of a methodological strategy—one of investigating specific higher cognitive capacities (e.g., planning) with a favoured process theory under FEP.

This brings us to a final and important more general point: even if one were to arrive at the conclusion that FEP is representational by way of taking one (or more) of the routes we have mapped out above, in evaluating FEP's representational status via considering the nature of its various aspects, one would implicitly take on one of the core tenets of embodied and enactive cognitive science; the claim that the role of coupled agent-environment dynamics is central to any adequate explanation of cognition (Schlicht & Starzak, 2021; Raab & Araujo, 2019).[23] That representational FEP takes on this tenet is interesting because many embodied and enactive cognitive science programmes often pull in the other direction, favouring non-representational explanations. This suggests that FEP may potentially offer an interesting middle ground for representationalist theories and embodied and enactive cognitive science. For example, given what the Markov blanket formalism is a description of, even if considerations about inferential seclusion of internal states lead one to hold a representational view of FEP, in adopting the Markov blanket formalism one has also committed oneself to the centrality of agent-environment dynamical exchanges in explanations of cognition (Allen & Friston, 2016); one has committed oneself not only to representing some of a system's features as internal states that infer external states but also committed oneself to representing some of a system's features as action and sensory states (blanket states) that mediate self-evidencing dynamic exchanges with the environment. Although such action and sensory states dynamics do not map onto folk psychological notions of action and perception in any direct manner (Sims, 2019a), the evolution of such states may be seen as abstractly describing coupled patterns of sensorimotor activity. This kind of causal action-perception dynamic is moreover welcomed by representationalists, whose commitment to inference by internal states is compatible with "conceiving of the mind and world as causally linked, through the causal interface of the Markov blanket" (Hohwy, 2017).

Similarly, when taking structural aspects as the route to a representationalist FEP, one's commitment to a view of generative models (or Q densities) as capturing environmental structure is compatible with the idea such models are primarily deployed in the service of guiding action as opposed to building rich internal reconstructions of the world. Recall, according to FEP only action can minimize *surprisal* and that minimizing surprisal is what underwrites an organism's continued viability (Bruineberg et al., 2016). Hence, even when models are viewed as representational, they are ultimately in the game of guiding action in ways that allow living systems to author evidence for their own existence (i.e., self-evidencing). In this way, coupled agent-environment dynamics are central to FEP's characterization of adaptive self-organization despite FEP's representational status when using the structural criterion. Again, the centrality of coupled action-perception dynamics in FEP is something that it shares with embodied and enactive theories which have typically avoided recourse to representations. Thus, when considering organizational or structural aspects, it is possible to arrive at a perspective on adaptive self-organization that potentially fuses representationalism with core aspects of traditionally non-representationalist theories, representing a kind of synthesis of say, internalism and various aspects of embodied and enactive cognitive science (Cf. Hohwy, 2017). The possibility of such a synthesis however fails to be available if one's commitments

[23] More precisely, embodied cognitive science emphasizes that the body (beyond the brain) plays a constitutive role in cognitive processes (Wilson & Foglia, 2017). Enactivism—in its most general form—emphasizes that cognition emerges via perception–action loops that dynamically couple the embodied agent to the environment (Varela et al., 1991).

stand with radical enactivism (see Baggs & Chemero, 2018). This being said, it should be noted that the possibility of representations figuring into the explanation of cognitive phenomena is consistent with non-radical forms of enactivism (Schlicht & Starzak, 2021) and embodied cognitive science (Raab & Araujo, 2019; Gentsch et al., 2016). Thus, depending upon one's previous commitments, FEP may offer a powerful overarching normative principle for enactive and embodied cognitive science.

In much the same manner as arriving at a representational or non-representational interpretation of FEP, arriving at an FEP as synthesis view depends upon which representational criterion are assumed when either considering FEP's central constructs or considering specific cognitive phenomena through the lens of a process theory under FEP. Hence, in the end, the debate about FEP may reveal more about us—our criteria for representation and our interests in particular facets of cognition—than it does about the representational status of FEP.[24]

6 Conclusion

We discussed different interpretations of FEP and explained how they depend on implicit assumptions about what qualifies (or fails to qualify) as representational. We have distinguished at least four main notions of representation, which focus on organizational, structural, content-related or functional aspects. The dispute is made more complex by the fact that FEP appeals to multiple constructs (e.g., Markov blankets, generative and recognition densities) and different process theories focus on one or the other; and furthermore, FEP models, can be constructed in various ways, e.g., one can use environmental and sensorimotor internal models. We have discussed in what ways these differences matter in arriving at a view of FEP as representational or non-representational (see Fig. 2). We believe that this debate, and the arguments and counterarguments we have reviewed, offer an opportunity to reflect both upon the importance of our implicit notions of representation—above and beyond the solution of "representation wars" in FEP—and how such notions of representation may implicitly shape the details of the process theories with which FEP is used. In other words, irrespective of its utility to offer insight into the adaptive self-organizing behaviour of biological organisms, FEP can be very heuristic for philosophy of mind: even if not so much to settle the dispute on internal representation but to unveil and dissect the hidden assumptions in the debate.

Another lesson learned from this debate is that some traditional polarizations between (for example) more classical cognitivist and enactivist perspectives may be attenuated or dissolved under a FEP treatment. Indeed, one can consider that the FEP advances a unified view where terms that traditionally belong to different ontologies—e.g., notions of model and expectation versus notions of autopoiesis and synchronization—can be harmonized. One can also consider that FEP advances these

[24] We would like to thank two anonymous reviewers for pushing us to clarify the points in this final section.

putatively disconnected ideas in novel territories: it begins with a strong enactivist flavour and a focus on action that is missing from traditional cognitive theories; but extends the scope of enactivist thinking to territories of (for example) counterfactual thinking and model selection that are rarely investigated. The extent to which this combination of enactivist and cognitive thinking is (theoretically and sociologically) possible—or desirable—remains to be seen in the future of FEP research.

Acknowledgements We would like to thank Julian Kiverstein for his very insightful feedback on an earlier draft of this paper. We are also grateful to three anonymous reviewers for the generosity and valuable feedback.

References

Ahissar, E., & Assa, E. (2016). Perception as a closed-loop convergence process. *eLife, 5,*. https://doi.org/10.7554/elife.12830

Allen, M., & Friston, K. J. (2016). From cognitivism to autopoiesis: Towards a computational framework for the embodied mind. *Synthese*, 1–24.

Baggs, E., & Chemero, T. (2018). Radical embodiment in two directions. *Synthese*. https://doi.org/10.1007/s11229-018-02020-9

Baltieri, M., & Buckley, C. L. (2017). An active inference implementation of phototaxis. https://arxiv.org/abs/1707.01806

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron, 76*(4), 695–711.

Brentano, F. (1874/1995). *Psychology from an empirical standpoint* (Trans. by A. C. Rancurello, D. B. Terrell, & L. McAlister). London: Routledge.

Bruineberg, J., Dolega, K., Dewhurst, J., & Baltieri, M. (2020). The Emperor's new Markov blankets. http://philsci-archive.pitt.edu/18467/. Retrieved 10 February, 2021.

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: The free-energy principle from an ecological-enactive perspective. *Synthese*. https://doi.org/10.1007/s11229-016-1239-1

Bruineberg, J., Rietveld, E., Parr, T., van Maanen, L., & Friston, K. J. (2018). Free-energy minimization in joint agent-environment systems: A niche construction perspective. *Journal of Theoretical Biology, 455,* 161–178.

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology, 81*(Supplement C), 55–79. https://doi.org/10.1016/j.jmp.2017.09.004

Buzsáki, G., Peyrache, A., & Kubie, J. (2014). Emergence of cognition from action. *Cold Spring Harbor Symposia on Quantitative Biology, 79,* 41–50. https://doi.org/10.1101/sqb.2014.79.024679

Cisek, P. (1999). Beyond the computer metaphor: Behavior as interaction. *Journal of Consciousness Studies, 6*(12), 125–142.

Cisek, P. (2007). Cortical mechanisms of action selection: The affordance competition hypothesis. *Philosophical Transactions of the Royal Society B, 362,* 1585–1599.

Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. *Attention, Perception, & Psychophysics*. https://doi.org/10.3758/s13414-019-01760-1

Clark, A. (1998). *Being there putting brain, body, and world together*. MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(03), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, A. (2015). *Predicting peace: The end of the representation wars* (p. 2015). MIND Group.

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

Clark, A. (2017). How to knit your own markov blanket. In Metzinger, T., and Wiese, W. (eds.), *Philosophy and predictive processing*.

Clark, A., & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior, 7*(1), 5–16.

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

Constant, A., Clark, A., & Friston K. J. (2019). Representation wars: Enacting an armistice through active inference. http://philsci-archive.pitt.edu/16125

Constant, C., Ramstead, M. J., Veissiere, S. P., Campbell, J. O., & Friston, K. J. (2018). A variational approach to niche construction. *Journal of the Royal Society, Interface, 15*(141), 20170685.

Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: Active inference, biological regulation, and the origins of cognition. *Biology and Philosophy, 35,* 32. https://doi.org/10.1007/s10539-020-09746-2

Craik, K. (1943). *The nature of explanation*. Cambridge University Press.

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy, 72*(20), 741–765.

de Hamilton, A. F., & Grafton, S. T. (2007). The motor hierarchy: From kinematics to goals and intentions. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), *Sensorimotor foundations of higher cognition*. Oxford University Press.

Downey, A. (2018). Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Synthese, 195*(12), 5115–5139. https://doi.org/10.1007/s11229-017-1442-8

Drescher, G. L. (1991). *Made-up minds: A constructivist approach to artificial intelligence*. MIT Press.

Dreyfus, H. L. (1979). *What computers can't do: The limits of artificial intelligence* (Vol. 1972). Harper & Row.

Engel, A. K., Friston, K. J., & Kragic, D. (2016). *The pragmatic turn: Toward action-oriented views in cognitive science*. MIT Press.

Engel, A. K., Maye, A., Kurthen, M., & König, P. (2013). Where's the action? The pragmatic turn in cognitive science. *Trends in Cognitive Sciences, 17*(5), 202–209. https://doi.org/10.1016/j.tics.2013.03.006

FitzGerald, T. H. B., Dolan, R. J., & Friston, K. J. (2014). Model averaging, optimal inference, and habit formation. *Frontiers in Human Neuroscience, 8,* 457. https://doi.org/10.3389/fnhum.2014.00457

Fodor, J. (1975). *The language of thought*. Harvard University Press.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K. J. (2011). Embodied inference: Or "I think therefore I am, if I am what I think". In W. Tschacher & C. Bergomi (Eds.), *The implications of embodiment: cognition and communication* (pp. 89–125). Imprint Academic.

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

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*. https://doi.org/10.1098/rsif.2013.0475

Friston, K. J. (2019). A free energy principle for a particular physics (pp. 1–148). https://arxiv.org/abs/1906.10184

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016a). Active inference and learning. *Neuroscience and Biobehavioral Reviews, 68,* 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2016b). Active inference: A process theory. *Neural Computation, 1,* 1–49. https://doi.org/10.1162/neco_a_00912

Friston, K. J., Levin, M., Sengupta, B., & Pezzulo, G. (2015a). Knowing one's place: A free-energy approach to pattern regulation. *Journal of the Royal Society, Interface, 12*(105), 20141383. https://doi.org/10.1098/rsif.2014.1383

Friston, K. J., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015b). Active inference and epistemic value. *Cognitive Neuroscience*. https://doi.org/10.1080/17588928.2015.1020053

Gallagher, S., & Allen, M. (2016). Active inference, enactivism and the hermeneutics of social cognition. *Synthese*. https://doi.org/10.1007/s11229-016-1269-8

Gardenfors, P. (1996). Cued and detached representations in animal cognition. *Behavioral Processes, 35,* 263–273.

Gentsch, A., Weber, A., Synofzik, M., Vosgerau, G., & Schütz-Bosbach, S. (2016). Towards a common framework of grounded action cognition: Relating motor control, perception and cognition. *Cognition*, *146*, 81–89.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Lawrence Erlbaum Associates Inc.

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

Globerson, A., Stark, E., Vaadia, E., & Tishby, N. (2009). The minimum information principle and its application to neural code analysis. *PNAS, 106*(9), 3490–3495. https://doi.org/10.1073/pnas.0806782106

Godfrey-Smith, P. (1994a). A modern history theory of functions. *Noûs, 28*(3), 344–362. https://doi.org/10.2307/2216063

Godfrey-Smith, P. (1994b). *Complexity and the function of mind in nature*. Cambridge University Press.

Godfrey-Smith, P. (1996). *Complexity and the function of mind in nature*. Cambridge, UK: Cambridge University Press.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), 20–25.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences, 27*(3), 377–396.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Sciences, 6,* 242–247.

Hobson, J. A., & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology, 98*(1), 82–98.

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

Hohwy, J. (2016). The self-evidencing brain. *Noûs, 50*(2), 259–280. https://doi.org/10.1111/nous.12062

Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing*. MIND Group. https://doi.org/10.15502/9783958573048

Hohwy, J. (2020). New directions in predictive processing. *Mind and Language*. https://doi.org/10.1111/mila.12281

Hohwy, J., Roepstorff, A., & Friston, K. J. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition, 108*(3), 687–701. https://doi.org/10.1016/j.cognition.2008.05.010

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

Huygens, C. (1673). *Horologium Oscillatorium sive de motu pendulorum*.

Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nature Reviews Neuroscience, 9*(4), 304–313. https://doi.org/10.1038/nrn2332

Jeannerod, J. (2006). *Motor cognition*. Oxford University Press.

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. MIT Press.

Kiefer, A., & Hohwy, J. (2017). Content and misrepresentation in hierarchical generative models. *Synthese*, 1–29.

Kirchhoff, M. D. (2016). Autopoiesis, free energy, and the life-mind continuity thesis. *Synthese*. https://doi.org/10.1007/s11229-016-1100-6

Kirchhoff, M., Parr, T., Palacios, E., Friston, K. J., & 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., & Robertson, I. (2018). Enactivism and predictive processing: A non-representational view. *Philosophical Explorations*. https://doi.org/10.1080/13869795.2018.1477983

Kiverstein, J., & Sims, M. (2021). Is free-energy minimisation the mark of the cognitive? *Biology and Philosophy*. https://doi.org/10.1007/s10539-021-09788-0

Korbak, T. (2019). Computational enactivism under the free energy principle. *Synthese*. https://doi.org/10.1007/s11229-019-02243-4

Kramar, M., & Alim, K. (2021). Encoding memory in tube diameter hierarchy of living flow network. *PNAS*. https://doi.org/10.1073/pnas.2007815118

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 20*(7), 1434–1448.

Lewontin, R. C. (1983). The organism as the subject and object of evolution. *Scientia, 77*(18), 65.

MacKay, D. M. (1969). *Information, mechanism and meaning*. Cambridge Mass.

Mathys, C. D., et al. (2014). Uncertainty in perception and the hierarchical gaussian filter. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2014.00825

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of living*. D. Reidel Pub.

McBeath, M. K., Shaffer, D. M., & Kaiser, M. K. (1995). How baseball outfielders determine where to run to catch fly balls. *Science, 268*(5210), 569–573.

McGeer, T. (1990). Passive dynamic walking. *The International Journal of Robotics Research, 9*(2), 62–82. https://doi.org/10.1177/027836499000900206

Merleau-Ponty, M. (1945). *PhÈnomÈnologie de la perception*. Gallimard.

Millikan, R. G. (1989). Biosemantics. *The Journal of Philosophy, 86,* 281–297.

Millikan, R. G. (1995). Pushmi-pullyu representations. *Philosophical Perspectives, 9,* 185–200.

Millikan, R. G. (2004). *Varieties of Meaning*. MIT Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice-Hal.

Nishimoto, R., & Tani, J. (2009). Development of hierarchical structures for actions and motor imagery: A constructivist view from synthetic neuro-robotics study. *Psychological Research, 73*(4), 545–558. https://doi.org/10.1007/s00426-009-0236-0

O’Regan, J. K., & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 883–917.

Odling-Smee, F. J., Laland, K. N., & Feldman, M. W. (2013). *Niche construction: The neglected process in evolution (MPB-37)* (Vol. 37). Princeton University Press.

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

Parr, T., Da Costa, L., & Friston, K. J. (2020). Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 378*(2164), 20190159. https://doi.org/10.1098/rsta.2019.0159

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers Inc.

Pezzulo, G. (2008). Coordinating with the future: The anticipatory nature of representation. *Minds and Machines, 18,* 179–220.

Pezzulo, G., Cartoni, E., Rigoli, F., Pio-Lopez, L., & Friston, K. J. (2016). Active Inference, epistemic value, and vicarious trial and error. *Learning & Memory, 23*(7), 322–338. https://doi.org/10.1101/lm.041780.116

Pezzulo, G., & Castelfranchi, C. (2007). The symbol detachment problem. *Cognitive Processing, 8*(2), 115–131.

Pezzulo, G., & Cisek, P. (2016). Navigating the affordance landscape: Feedback control as a process model of behavior and cognition. *Trends in Cognitive Sciences, 20*(6), 414–424. https://doi.org/10.1016/j.tics.2016.03.013

Pezzulo, G., Donnarumma, F., Iodice, P., Maisto, D., & Stoianov, I. (2017a). Model-based approaches to active perception and control. *Entropy, 19*(6), 266. https://doi.org/10.3390/e19060266

Pezzulo, G., Kemere, C., & van der Meer, M. (2017b). Internally generated hippocampal sequences as a vantage point to probe future-oriented cognition. *Annals of the New York Academy of Sciences, 1396,* 144–165.

Pezzulo, G., Rigoli, F., & Friston, K. J. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology, 136,* 17–35.

Pezzulo G., Zorzi M., & Corbetta M. (2020) The secret life of predictive brains: What’s spontaneous activity for?

Piaget, J. (1954). *The construction of reality in the child*. Ballentine.

Pio-Lopez, L., Nizard, A., Friston, K., & Pezzulo, G. (2016). Active inference and robot control: A case study. *Journal of the Royal Society Interface, 13,* 122. https://doi.org/10.1098/rsif.2016.0616

Port, R., & van Gelder, T. (1995). *Mind as motion: Explorations in the dynamics of cognition*. MIT Press.

Raab, M., & Araujo, D. (2019). Embodied cognition with and without mental representations: The case of embodied choices in sports. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2019.01825

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

Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2017). Answering Schrödinger’s question: A free-energy formulation. *Physics of Life Reviews*. https://doi.org/10.1016/j.plrev.2017.09.001

Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2019). A tale of two densities: Active inference is enactive inference. *Adaptive Behavior*. https://doi.org/10.1177/1059712319862774

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87. https://doi.org/10.1038/4580

Rorty, R. (1979). *Philosophy and the mirror of nature*. Princeton University Press.

Roy, D. (2005). Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence, 167*(1–2), 170–205. https://doi.org/10.1016/j.artint.2005.04.007

Ryle, G. (1949). *The concept of mind*. Barnes and Noble.

Schlicht, T., & Starzak, T. (2021). Prospects of enactivist approaches to intentionality and cognition. *Synthese, 198*(Suppl 1), S89–S113. https://doi.org/10.1007/s11229-019-02361-z

Schulz, A. W. (2018). *Efficient Cognition: The evolution of representational decision making*. MIT Press.

Schwartenbeck, P., FizGerald, T., Dolan, R. J., & Friston, K. J. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2013.00710

Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive neuroscience, 5*(2), 97–118.

Shea, N. (2012). Methodological encounters with the phenomenal kind. *Philosophy and Phenomenological Research, 84*(2), 307–344. https://doi.org/10.1111/j.1933-1592.2010.00483.x

Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.

Sims, M. (2019a). Minimal-perception: Responding to the challenges of perceptual constancy and veridicality with plants. *Philosophy and Psychology, 32,* 1024–1048. https://doi.org/10.1080/09515089.2019.1646898

Sims, M. (2019b). Coupling to variant information: An ecological account of comparative mental imagery generation. *The Review of Philosophy and Psychology, 11,* 899–916. https://doi.org/10.1007/s13164-019-00454-9

Sims, M. (2020). How to count biological minds: Symbiosis, the free-energy principle, and reciprocal multiscale integration. *Synthese*. https://doi.org/10.1007/s11229-020-02876-w

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A. E., Paliwal, S., Gard, T., et al. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, *10*(550), 1–27.

Sterelny, K. (2003). *Thought in a Hostile world: The evolution of human cognition*. Blackwell Publishing Ltd.

Stoianov I., Maisto D., & Pezzulo G. (2020). The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of perception and action*. MIT Press.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55,* 189–208.

Ungerleider, L. G., & Haxby, J. V. (1994). “What” and “where” in the human brain. *Current Opinion in Neurobiology, 4*(2), 157–165.

van Rooij, I., Blokpoel, M., de Haan, R., & Wareham, T. (2019). Tractable embodied computation needs embeddedness. *Reti, Saperi, Linguaggi, 1,* 25–38. https://doi.org/10.12832/94728

Varela, F., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT Press.

von Helmholtz, H. (1866). Concerning the perceptions in general. In J. P. C. Southall (Ed.), *Treatise on physiological optics* (Vol. 3). Dover.

Wiese, W., & Friston, K. J. (2021). Examining the continuity between life and mind: Is there a continuity between autopoietic intentionality and representationality. *Philosophies, 6,* 18. https://doi.org/10.3390/philosophies6010018

Williams, D. (2018). Predictive processing and the representation wars. *Minds and Machines, 28*(1), 141–172. https://doi.org/10.1007/s11023-017-9441-6

Wilson, R. A., & Foglia, L. (2017). Embodied cognition. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University.