Synthese (2023) 202:58
https://doi.org/10.1007/s11229-023-04292-2

ORIGINAL RESEARCH

Incorporating (variational) free energy models into mechanisms: the case of predictive processing under the free energy principle

Michał Piekarski[1]

Received: 25 January 2022 / Accepted: 29 July 2023 / Published online: 10 August 2023

Abstract
The issue of the relationship between predictive processing (PP) and the free energy principle (FEP) remains a subject of debate and controversy within the research community. Many researchers have expressed doubts regarding the actual integration of PP with the FEP, questioning whether the FEP can truly contribute significantly to the mechanistic understanding of PP or even undermine such integration altogether. In this paper, I present an alternative perspective. I argue that, from the viewpoint of the constraint-based mechanisms approach, the FEP imposes an important constraint, namely variational free energy, on the mechanistic architecture proposed by PP. According to the constraint-based mechanisms approach, high-level cognitive mechanisms are integral parts of extensive heterarchical networks that govern the physiology and behavior of agents. Consequently, mechanistic explanations of cognitive phenomena should incorporate constraints and flows of free energy as relevant components, given that the implemented constraints operate as long as free energy is available. Within this framework, I contend that the FEP provides a relevant constraint for explaining at least some biological cognitive mechanisms described in terms of Bayesian generative models that minimize prediction errors.

Keywords Predictive processing · Mechanisms · Explanation · Constraints · Free energy principle · Variational free energy

1 Introduction

It has been established that proponents of the PP framework seek mechanistic explanations and that the various models of cognitive functions developed via PP are aimed at

✉ Michał Piekarski
m.piekarski@uksw.edu.pl; m.a.piekarski@gmail.com

[1] Institute of Philosophy, Cardinal Stefan Wyszyński University in Warsaw, Wójcickiego 1/3 St, 01-938 Warsaw, Poland

this kind of account (Friston et al., 2018; Gładziejewski, 2019). In line with this view, it has been argued that PP provides a sketch of a mechanism (Gładziejewski, 2019; Gordon et al., 2019; Harkness, 2015; Harkness & Keshava, 2017; Hohwy, 2015), i.e., an incomplete representation of a target mechanism in which some structural aspects of a mechanistic explanation are omitted (cf. Piccinini & Craver, 2011). Understood in this way, the sketch is defined in terms of functional roles played by the respective components, disregarding to some extent their biological or physical implementation. This raises the important question of how to understand the causal structure responsible for predictive mechanisms. It can be a simple multi-level hierarchy from simple neural levels of, e.g., pattern recognition, edge detection, color perception etc. (implemented in the early sensory system), to high-level neural representations (implemented deep in the cortical hierarchy [Sprevak, 2021]), to increasingly abstract and general levels related to Bayesian beliefs and concerning the general properties of the world; or it can be a subtler structure implemented by several different, partially independent mechanisms responsible for various phenomena.[1]

The key to this type of practice is the recognition of cognition in the categories of mechanistic causal relations (cf. Gładziejewski, 2019, p. 665). Gładziejewski suggests that sketches of mechanisms provided by PP should be understood in the sense that these models "share common core assumptions about relevant mechanisms" but do not describe a single cognitive structure (mechanism). This means that "there are a couple of ways in which a collection of mechanisms that fall under a common predictive template could provide a schema-centered explanatory unification" (Gładziejewski, 2019, p. 666). This author points to four possible research heuristics which, by providing sketches, may allow the identification of actual mechanisms:

1. There are separate neural mechanisms that follow the same predictive scheme;
2. Different levels within one hierarchy can explain different cognitive phenomena;
3. Various aspects of PP mechanisms are explanatory, which means that for a given mechanism, certain aspects of its functioning may explain specific phenomena;
4. The ways in which distinct PP mechanisms become integrated may play explanatory roles (Gładziejewski, 2019, pp. 666–667).

Regardless of which of the indicated heuristics is actually employed by PP researchers (whether it be one or a combination of several), there is no doubt that many supporters of PP seek mechanistic explanations.

As can be seen, the thesis about the mechanistic nature of PP is already reasonably well-founded, but it seems that in light of the view advocated by some mechanists (cf. Bechtel, 2019; Winning & Bechtel, 2018), (at least) some mechanistic explanations should include constraints and flows of free energy as their constitutive component. This view, which I will refer to in this paper as the constraint-based mechanisms approach,[2] could be of great importance to many debates about PP and FEP theories

[1] Regardless of how to understand the exact causal basis of the implementation of predictive mechanisms, the mechanistic strategy of reconstructing these mechanisms by providing their sketches certainly corresponds to the actual practice of PP researchers (cf. Gordon et al., 2019; Keller & Mrsci-Flogel, 2018).

[2] This approach is based on the recent papers of William Bechtel and colleagues and, in a sense, unifies their views as presented in various papers. The very concept of constraint-based mechanisms approach has not appeared in the literature so far and, as such, is a novelty. The same is the case with heuristics of constraint-based mechanisms, which can be taken as a distinctive feature of this approach.

as it allows for a rethink of the relationship between PP, FEP, and FEP-based Active Inference.

The possibility of a mechanistic integration of PP and the FEP has already been raised by researchers. Some responses have also been offered. There are authors who share the viewpoint that the FEP carries mechanistic implications for PP, asserting that the FEP can be treated as a heuristic guide or regarded as a regulatory principle. Supporters of the first position include Paweł Gładziejewski who, in his paper *Mechanistic unity of the predictive mind*, states that the FEP is "a powerful heuristic guide for the development of PP" but "only puts extremely general constraint on the causal organization of organisms, perhaps to the point of lacking any non-trivial commitments about it" (Gładziejewski, 2019, p. 664). Another supporter, Dominic Harkness, claims that "the upshot of this criticism lies within the free energy principle's potential to act as a heuristic guide for finding multilevel mechanistic explanations" (Harkness, 2015, p. 2). Jakob Hohwy supports the second position, claiming that the "FEP can be considered a regulatory principle, 'guiding' or 'informing' the construction of process theories" (Hohwy, 2020, p. 39), meaning that the FEP provides "distinct process theories explaining perception, action, attention, and other mental phenomena" (Hohwy, 2020, p. 47).

However, some researchers are not convinced by the FEP or its explanatory relationship with PP. For example, Daniel Williams in his recent paper *Is the brain an organ for free energy minimization?* argues that "the claim that the FEP implies a substantive constraint on process theories in cognitive science—namely, that they must describe how the brain's mechanisms implement free energy minimization—rests on a fallacy of equivocation" (Williams, 2021, p. 8). Similarly, Mateo Colombo and Patricia Palacios in their paper *Non-equilibrium thermodynamics and the free energy principle in biology* note that "because of a fundamental mismatch between its physics assumptions and properties of its biological targets, model-building grounded in the free energy principle exacerbates a trade-off between generality and biological plausibility" (Colombo & Palacios, 2021, p. 2). Colombo defends a slightly different position in a paper co-written with Cory Wright, where they take into account that the analysis carried out by the FEP's supporters can be treated as sketches of mechanisms in the sense of Piccinini and Craver (2011). They do, however, only treat them as weak explanatory idealizations: "Some of the confusions in recent debates surrounding the FEP, organicism, and mechanism depend on indulging this sort of metaphysics without carefully considering the epistemic and pragmatic roles that 'rampant and unchecked' idealizations, like those involved in FEP, play in science" (Colombo & Wright, 2021, p. 3486).

In this paper I will take a different starting point. I want to demonstrate by reference to the constraint-based mechanisms approach, that the FEP offers an explanatory relevant (variational) constraint for the causal organization of any and all systems equipped with generative models, explained mechanistically by PP. In other words, I will claim that the FEP provides a constraint which determines PP's scheme of mechanism.

This paper has the following structure: in Sect. 2, I present an overview of the PP and FEP frameworks and explain why, when analyzing predictive mechanisms, one should take into account the quantity described in the literature on the FEP as variational free

energy (VFE). In Sect. 3, I sketch the new mechanical philosophy and its characteristic systems tradition, describing explanations in terms of the identification and decomposition of mechanisms. I also present the recent position based on mechanism, which I refer to as the constraint-based mechanisms approach and—characteristic for this approach—the so-called heuristics of constraint-based mechanisms. In the following part, I formulate a mechanistic interpretation of PP and wonder if it can meet the norm defined by the heuristics of constraint-based mechanisms. The context of the question is set by the discussion on the FEP and its explanatory relationship with PP. In Sect. 4, I discuss two main possible interpretations (realistic and instrumental) of the statement that self-organizing systems minimize VFE. Discussing them is important because it provides an initial answer to whether the FEP determines the energetic (in information-theoretic sense) constraint for mechanistic PP. In Sect. 5, I articulate the position of mechanistic realism, which asserts the feasibility of employing heuristics based on the constraint-based mechanisms approach. I argue that the interpretation of the FEP, which I called moderate realistic, is compatible with mechanistic realism. In Sect. 6, I discuss Karl Friston's argument from Bayesian mechanics that VFE coincides with thermodynamic free energy (TFE). If Friston's perspective is accurate, the FEP serves a similar explanatory role in elucidating living organisms as thermodynamics does in explaining physical systems. However, in this section, I reject Friston's argument because of its instrumental character, which precludes mechanistic realism and the application of the heuristics of constraint-based mechanisms. As a result, in Sect. 7, I present an argument in favor of moderate realism regarding the FEP and FEP-based PP. This argument is supported by empirical evidence from investigations into neural computations and the thermodynamics of information. Next I discuss the ontological commitments of this position, and I also formulate a provisional response to the objections of those authors who deny explanatory value to the FEP. In the *Conclusion*, I summarize the analyses carried out.

2 Predictive processing and variational free energy

PP is a process theory of the brain that provides a computational model of cognitive mechanisms and core processes that underwrite perception and cognition. Some advocates of PP believe that it can be used to unify the models of perception, cognition, and action theoretically (Clark, 2013; Hohwy, 2015; Seth, 2015). Specific versions of PP are grounded in the same process of precision-weighted, hierarchical, and bidirectional message passing and error minimization (Clark, 2013; Hohwy, 2020). In this framework, perceptual and cognitive processes are conceived as being the result of a computational trade-off between (hierarchical) top-down processing (predictions based on the model of the world) and bottom-up processing (prediction errors tracking the difference between predicted and actually sensed data). A characteristic feature of this view is the assumption that, in order to perceive the world, the cognitive system must resolve its uncertainty about the 'hidden' causes of its sense states. This is because the causes of the sensory signals are not directly recognized or detected, but instead must be inferred by a hierarchical, multi-level probabilistic (generative) model. In PP,

the activity of the brain (or cognitive system) is understood as instantiating or leveraging a generative model (cf. Clark, 2016), which is, generally speaking, a model of the process that generated the sensory data of interest. In short, PP purports to explain the dynamics of the brain by appealing to hierarchically organized bidirectional brain activity, cast as instantiating a generative model.

The generative model is defined as the joint probability of the "observable" data *e*—sensory state, and *h*—a hypothesis about these data (trees, birds, glasses etc.). In other words, a generative model is the product of *p(h)* (priors over states) *and* $p(e|h)$ (likelihood of evidence probability if the hypothesis is true). This means that the generative model is a statistical model of how observations are generated (strictly speaking, a description of causal dependencies in the environment and their relation to sensory signal). It uses prior distributions *p(h)* (which determine the probability of hypothesis before evidence) that the system applies to the environment about which it makes inferences.

The model minimizes the so-called prediction errors, i.e., the differences between the expectations of the organism—its "best guess" about what would be the case (what caused its sensory states) and what the organism factually observes. To minimize prediction errors, the generative model continuously creates statistical predictions about what is happening or can happen in the world. This means that updating the likelihoods and priors based on prediction errors is a mechanism that can be described in terms of Bayesian inference, i.e., a statistical inference in which a Bayesian rule is used to update the probability for a hypothesis as more evidence or data becomes available.

Technically speaking, according to the Bayesian rule

$$p(h|e) = \frac{p(e|h)p(h)}{p(e)},$$

the generative model $p(h|e)$ calculates the posterior probability $p(e|h)$, which in practice allows the system to assume the most probable hypothesis explaining the nature and causes of the sensory signal, taking into account the available sensory data.[3] This hypothesis enables the minimization of the long-term average prediction error (Hohwy, 2020). Moving from $p(h|e)$ to $p(e|h)$, i.e., inverting the likelihood mapping, allows one to update beliefs from prior to posterior beliefs (Smith et al., 2022, p. 3). Proponents of the PP framework argue that the model approximates Bayesian inference rather than computing it exactly (cf. Clark, 2013). In PP, the model implements an algorithm that computes Bayesian inferences so that the prediction error is gradually minimized, which maximizes the posterior probabilities of the hypotheses.

This way, when the model minimizes the prediction error, it also minimizes a certain *quantity* that is always greater than or equal to the surprisal—negative log probability of an observation/outcome—the surprisal model itself cannot be minimized directly due to ignorance of the underlying causes of the sensory signals (Friston, 2009, p. 294). This quantity refers to the objective function that is known as VFE or an evidence lower

[3] In this sense, the model update proceeds in a rational manner.

bound (cf. Winn & Bishop, 2005). The introduction of VFE helps to convert exact Bayesian inference into approximate Bayesian inference.[4]

Why is this important? Approximate Bayesian inference uses VFE minimization, which can be described as the difference between the approximate posterior distribution of the model and the target distribution. The introduction of an approximate posterior distribution over states, denoted $q(e)$ (such that each $q(e) \in Q$ is a possible approximation to the exact posterior distribution), makes simplifying assumptions about the nature of the true posterior distribution. By iteratively updating the approximate posterior (initially arbitrary), one can find a distribution that approximates the exact posterior. The next step is to measure the similarity between approximated $p(h|e)$ and the true posterior $p(e|h)$. Formally, this means minimizing the so-called Kullback–Leibler divergence (KL-divergence). It is important that KL-divergence cannot be directly estimated, and therefore the model must optimize a different function (i.e., VFE) which bounds the model evidence. The smaller the VFE, the smaller the KL-divergence. When KL-divergence is zero, then the distributions match. It gets larger the more dissimilar the distributions become. In variational inference, the model iteratively updates approximate posterior $q(e)$ until it finds the value that minimizes VFE at which $q(e)$ will approximate the true posterior $p(e|h)$ (Smith et al., 2022; cf. Buckley et al., 2017).

The association of PP with VFE helps explain how the generative model minimizes prediction errors by Bayesian inference approximation, which can be interpreted as the way in which neural information processing mechanisms perform variational inference. This remark is crucial for further analyses.

To sum up: predictive mechanisms can be described in terms of the realization of variational principles (cf. Friston et al., 2017). In research practice, this means that in order to be able to concretize any variational inference algorithm, we must define the forms of the variational posterior and the generative model, which in the case of PP means (relying on the Laplace assumption) that posterior probability densities are normal (Gaussian). With this assumption in place, free energy can be viewed as the sum of the long-term average prediction error, which is supposed to be linked to the FEP (cf. Friston, 2010). It means that in the context of PP, the process involves the minimization of long-term average prediction error through the model's optimization of the statistics of an approximate posterior distribution. Modelers postulate and refine this distribution to align with the desired target distribution (Millidge et al., 2021, p. 7). This is an important observation for the very understanding of PP because it allows us to think about the normative function of the predictive mechanisms, which is the long-term average precision-weighted error in terms of free energy minimization.

[4] VFE was introduced by Richard Feynman to solve an intractable inference problem in quantum electrodynamics (Feynman, 1998, cf. Friston et al., 2006, p. 221). Minimization of a computable objective function will approximate the minimization of the evidence. This evidence is always upper bounded by VFE. This means that by introducing VFE, an intractable integration problem was converted into a tractable optimization problem; namely minimizing VFE (Dayan et al., 1995; Friston, 2011). Thus, in variational inference, the model does not directly compute the intractable true posterior. Instead, it optimizes a tractable upper bound on this divergence, called the VFE. VFE is a tractable quantity because it is the discrepancy between two qualities (which we know as modeling subjects) i.e., the variational approximate posterior and the generative model. And because VFE is an upper bound, minimizing it brings us closer to true posterior.

At this point, however, difficulties arise regarding the linking of the PP framework with the research framework motivated by the FEP. Before discussing them (cf. §4), it is necessary to at least briefly explain what the FEP is.

The FEP was introduced by Karl Friston and colleagues as a mathematical framework that specifies the objective function that self-organizing systems need to minimize in order to change their relationship with the environment and maintain thermodynamic homeostasis (Friston, 2009, 2010, 2012; Friston & Stephan, 2007; Friston et al., 2006; cf. Andrews, 2021). Originally, the FEP was a principle explaining how the sensory cortex infers the causes of its inputs and learns causal regularities. What distinguished the FEP from other theories of inference (cf. Gregory, 1966; Rock, 1983) is the fact that all cognitive processes and functions, not only perceptual, can be explained in terms of one unifying principle, which is the minimization of free energy (Bruineberg et al., 2021, p. 3; cf. Friston, 2010). Later, the validity of the FEP was extended from perception and action to organization of all self-organizing systems: from unicellular cells to social networks (cf. Friston, 2009, p. 293; 2013; Wiese & Friston, 2021).[5]

According to the current formulation of this principle[6] any self-organizing system that is at a nonequilibrium steady-state (NESS) with its environment must minimize its free energy.[7] In other words, any "thing" that achieves NESS can be construed as performing a Bayesian inference with posterior beliefs that are parameterized by the thing's (model's) internal states. In other words, the FEP offers an interpretation of mechanical theories of systems *as if* they possess (Bayesian) beliefs (Ramstead et al., 2023, p. 2). This is related to the fact that the state flow of a given self-organizing system can be described as a function of their NESS density. The system, if it exists, can be described in terms of a random dynamic system (in terms of Dynamic System Theory—DST) that evolves, which means that it can be said to change over time, subject to random fluctuations. It must be added that any self-organizing system that is at NESS, i.e., one that has an attracting set, can be described in terms of Markov blankets (Friston, 2013; Friston et al., 2020; Wiese & Friston, 2021).[8]

[5] In the light of the analyses carried out, one can invoke Jakob Hohwy's observation that the FEP as a mathematical principle is a regulatory principle. Hohwy is probably right when he states that the FEP itself does not imply cognitive architecture (Hohwy, 2021, p. 47). However, it is important to answer whether the FEP is a regulatory principle or has a specific explanatory power in the explanation of neurocognitive mechanisms modelled by the PP framework.

[6] I use the term "current" because the FEP and the Active Inference framework are constantly modified by their proponents. This can of course be explained by the internal dynamics of the theory development, but for this reason, for the opponents of using this research framework "FEP can appear like a moving target, each time introducing new constructs that make the previous criticism inapplicable" (Bruineberg et al., 2021, p. 2).

[7] The notion of NESS comes from statistical mechanics, in which it denotes the energy dynamics between the system and the surrounding heat bath. NESS is best understood as a breach of this balance.

[8] The full presentation of Markov blankets goes beyond these considerations, so I will only discuss them to the extent necessary for further analysis. The concept of Markov blankets comes from research on Bayesian inference, Bayesian networks, and graphical modeling (Pearl, 1988; cf. Bruineberg et al., 2021), and basically means a set of random variables which "shield" another set of random variables from other variables in the system. One set of variables (we can call them states) makes states internal to the blanket conditionally independent of external states. For a Bayesian network (described in terms of a directed acyclic graphical model) the Markov blanket comprises the parents, children, and parents of the children of a state. Markov blankets allow for the division of blanket states into internal and external states via their conditional

NESS density means a certain probability of finding it in a particular state when the system is observed at random (Friston et al., 2020, p. 4). In this sense, everything that exists is characterized by properties that remain unchanged or stable enough to be measured over time. In other words, this means that the states of a given system behave *as if* they are trying to minimize exactly the same quantity: the surprisal of states that constitute the thing, system, and so on. That is, everything that exists will act *as if* to minimize the entropy of its particular states over time. Thus, open systems that are far away from equilibrium resist the second law of thermodynamics (Friston & Stephan, 2007; cf. Davies, 2019; Ueltzhöffer, 2019). What exists must be in a sense self-evidencing, meaning that it must maximize a particular model evidence or equivalently minimize surprisal (cf. Hohwy, 2016). This way, according to Friston and colleagues, it is possible to interpret the flow of (expected) autonomous states of the model as a gradient flow on something what we know as VFE,[9] and at the same time allows us to think of systems that have Markov blankets as "agents" that optimize the evidence for their own existence. In this sense, their internal states with the blanket surrounding them are (in some sense) autonomous (Kirchhoff et al., 2018, p. 2; cf. Friston et al., 2020). Autonomy understood in this way allows us to think of "agents" as adaptive systems, where adaptivity refers to an ability to operate differentially in certain circumstances. This means that a system that is not adaptive, suggesting that it does not have a Markov blanket and cannot exist.[10]

On the basis of the conducted analyses, it can be concluded that the FEP, as a formal statement—the existential imperatives for any system that manages to survive in a changing environment—can be treated as a generalization of the second law of thermodynamics to NESS (Parr et al., 2020). In that sense, the FEP is true for any bounded stationary system that is far from equilibrium, because the FEP applies to all self-organizing systems at NESS (meaning that the FEP applies to all systems equipped with the generative model because NESS density can be described in the terms of generative model [Friston, 2019, p. 89; cf. Sakthivadivel, 2022]).[11]

Footnote 8 continued
independence. Then the blanket states can be further divided into sensory and active states where sensory states are not influenced by internal states, and active states are not influenced by external states. Internal and external states can only influence each other through a blanket (Friston, 2013). Understanding of Markov blankets proposed by Friston differs from that introduced by Pearl. The latter understands blankets in an instrumental way, as a mathematical construct. According to Friston, they gain an "ontic" interpretation that is not "philosophically innocent" (Bruineberg et al., 2021; see also: Beni, 2021). Without going into detail, I emphasize that in these analyses, I refer to Markov blankets in a Fristonian manner.

[9] Information geometry is also related to the parameterizing states. Information geometry offers a formalism for describing the distance between probability distributions in an abstract space. In this space, each point represents a possible probability distribution. According to Friston (2019), all systems with NESS distribution and Markov blankets can be described in terms of information geometry (cf. Friston et al., 2020, pp. 9–11). The analysis of this issue, however, goes beyond the scope of this paper.

[10] Not all existing self-organizing systems are alive. The FEP also applies to such systems—non-biological agents—which have a certain degree of independence from the environment (Wiese & Friston, 2021, p. 3).

[11] This corresponds in some way to the concept of living organisms defended by mechanists as autonomous dissipative structures, i.e., those "that […] actually use the second law of thermodynamics to their advantage to maintain their organization" (Winning & Bechtel, 2018, p. 3; cf. Friston & Stephan, 2007; Kirchhoff et al., 2018; Ueltzhöffer, 2019).

3 Systems tradition of mechanistic explanation and the constraint-based mechanisms approach

In §1, I drew attention to the fact that many researchers either have doubts about the actual integration of PP with the FEP—where the FEP would offer an explanatory significant contribution to the mechanistic PP (cf. Gładziejewski, 2019; Harkness, 2015; Hohwy, 2020), or even negate such a possibility (cf. Colombo & Palacios, 2021; Colombo & Wright, 2021; Williams, 2021). In this paper, I propose a different research perspective, according to which the FEP imposes an explanatory relevant informational constraint (i.e., VFE) on the mechanistic architecture postulated by PP. In order to justify this view, I will refer to the position I call the constraint-based mechanisms approach. Before I develop my argument, however, it is necessary to explain, albeit briefly, what this approach is.

Scientific research can be described in terms of discovering and describing mechanisms. In many fields of science, it is assumed that in order to formulate a satisfactory explanation of the phenomenon under study, one needs to provide a decomposition of its mechanism. Mechanistic explanations are used with great success in neuroscience as well as in biological, physical, and social sciences (cf. Glennan & Illari, 2018). This new mechanistic explanatory program became the dominant view across many debates in the philosophy of science (Bechtel, 2008; Bechtel & Richardson, 1993/2010; Craver, 2007; Craver & Darden, 2013; Machamer et al., 2000).

The introduction of a new mechanism comes with the assumption that a distinction should be made between explanations which are componential or constitutive and etiological explanations, which explain a phenomenon by describing its antecedent causes. Constitutive explanations detail a phenomenon by describing its underlying mechanism, i.e., the relation between the behavior of a mechanism as a whole and the organized activities of its individual components is constitutive (cf. Salmon, 1984).[12] The latter's explanations assume a strategy of decomposing high-level cognitive capacities into components that are responsible for various information processing operations, and then using various computational models, showing how these operations together explain a given phenomenon. Decomposition is a characteristic determinant of the 'systems tradition' (Craver, 2007; cf. Bechtel & Richardson, 1993/2010; Cummins, 1975; Fodor, 1968; Simon, 1969). In this tradition, explanation is understood as a matter of decomposing systems into their parts to show how those parts are organized in such a way to emphasize the explanandum phenomenon.

Systems tradition is currently the dominant approach to explanations formulated in biology, system research, and cognitive neuroscience, while decomposition is the central heuristic strategy in mechanistic explanations besides the identification of mechanisms (Bechtel & Richardson, 1993/2010; cf. Bechtel, 2008; Craver, 2007; Illari & Williamson, 2013). However, the mechanistic view of explanation has met with controversy (cf. Koutroufinis, 2017; Silberstein & Chemero, 2013). Moreover, some

[12] In this paper, by "explaining" I mean "constitutive explanations".

authors defend dynamical explanation as an alternative to mechanistic explanation (cf. Stepp et al., 2011).[13]

3.1 What about constraints?

Some researchers (cf. Bechtel, 2018, 2019, 2021; Bechtel & Bollhagen, 2021; Winning, 2020; Winning & Bechtel, 2018) point out that the decomposition strategy, as understood by mechanism, assumes that there is a composition or causation relationship (i.e., causal production) between processes present in mechanisms (where one process, an organized set of causal processes is "responsible for" the implementation of another). Such a view, however, ignores two important features of cognitive mechanisms:

1. Mechanisms of this kind primarily act to control production mechanisms, i.e., mechanisms which are responsible for bodily movement and physiological processes. This type of relationship can be called control, and it is as important for the understanding of the nature of mechanisms and their explanations as the relationships of causation and composition (Winning & Bechtel, 2018, p. 2). These are, therefore, mechanisms that help to maintain the internal environment of the given organisms. The analysis of control mechanisms is important because they allow organisms to quickly adapt to their environment. Therefore, they perform an important adaptive function and are responsible for the autonomy of the individual, as they contribute to the maintenance of the existence of a given organism. In this sense, they are normative because they contribute to the self-maintenance that is the norm of autonomous living systems (cf. Bickhard, 2003). Self-maintenance is the norm (what is good or bad for the system) in the sense that it "is not externally interpreted or derived from an adaptive history but defined intrinsically by the very organization of the system" (Barandiaran & Moreno, 2006, p. 174);
2. High-level cognitive mechanisms are components of a highly developed and complex network of heterarchically organized control systems whose aim is to perform a given cognitive task (Bechtel, 2019, p. 621, cf. Pattee, 1991). By heterarchical organization, I mean a such distributed causal network in which a given (production) mechanism is regulated by multiple (control) mechanisms without these control mechanisms being themselves subsumed under a higher-level controller. This means that their organization is horizontal and not vertical, as is the case with hierarchical organization (cf. Bechtel & Bich, 2021).[14]

[13] My goal here is not to argue with models of explanations that are alternative to mechanism, or to discuss their validity, especially since there are strong arguments that dynamic models are ultimately mechanistic (cf. Bechtel & Abrahamsen, 2010; Kaplan & Bechtel, 2011; Zednik, 2008). I am rather interested in the discussion that took place within mechanism about the limitations of this view (cf. Bechtel, 2018, 2019, 2021; Bechtel & Bollhagen, 2021; Winning & Bechtel, 2018; Winning, 2020).

[14] "In both machines and human institutions, control mechanisms are often organized hierarchically. In a hierarchy, individual control mechanisms are themselves controlled by higher-level control mechanism, with a single controller ultimately in charge. The system is organized as a pyramid. In living systems, however, control mechanisms are typically organized heterarchically" (Bich & Bechtel, 2021, p. 2). The notion of heterarchy first introduced McCulloch (1945). See also Cumming (2016).

These features (1) and (2) are extremely important and their omission in explaining cognitive mechanisms makes these explanations incomplete, violating the standard of mechanistic explanations (Craver & Kaplan, 2018). This may result in "incorrect accounts of cognition" (Bechtel, 2019, p. 621).[15] Taking account of these two aspects of cognitive processes, i.e., their function in the production of control mechanisms and their non-autonomous character, leads to the conclusion that their explanation should also cover other components (some of which are flexible and able to be operated on and altered by other mechanisms) than those previously considered.[16] This means that the mechanisms are organized not only in terms of production and composition, but also in terms of control. Such a view thus presupposes a revision of the systems tradition in which "processes are controlled by other processes, and mechanisms are controlled by other mechanisms, often hierarchically" (Winning & Bechtel, 2018, p. 3).

A drift from the classical understanding of systems tradition does not mean a departure from the norms of mechanistic explanations, but rather their extension and the recognition that the concept of constraint is also important from the explanatory perspective. The concept of constraint comes from classical mechanics. It was used to describe the reduction of the degree of freedom available to components organized into macroscale objects. Constraints define some limits on independent behavior but also create possibilities (Hooker, 2013). For example, in contexts where there is a source of (thermodynamic) free energy, constraints can be used to direct the flow of this energy. This means that elements of biological mechanisms can be used to limit the flow of available free energy so that work is done (which can be used to generate particular phenomena). Some (control) mechanisms are therefore systems of constraints that restrict the flow of free energy to perform work. Therefore, the operation of control mechanisms leads to such behaviors or physiological processes that would not be possible if not for the changes that constraints make in the mechanisms of production. Controlling the production mechanisms is essential because they are constrained to do work as long as free energy is available. The same is true for artifacts. For example: turning on the on/off switch enables the user of a given machine to control it so that it can use energy and carry out its design activities (Bechtel, 2019, p. 623).[17]

Constraints understood in this way do not only (or at all) function as the context or background conditions in which a given mechanism is implemented, but most of all they are its constitutive (in the sense of being responsible for producing a given

[15] This is not to say that the systems tradition does not recognize the importance of constraints (cf. Craver, 2007; Darden, 2006). I do claim, however, that it treats constraints as background conditions or as factors that limit the space of possible mechanisms. In the constraint-based mechanisms approach, the constraints are primarily control mechanisms.

[16] Certain mechanists have engaged in discussions regarding specific control mechanisms, such as circadian mechanisms (Bechtel & Abrahamsen, 2010) and feedback mechanisms (Bechtel, 2008, Ch. 7). Nevertheless, they did not talk much about the effects of these mechanisms on others within certain complexes.

[17] The concept of constraint, as used in this context, was originally proposed by Howard Pattee (1972) and David Marr (1982). Marr drew attention to the fact that specific processes can be defined by indicating and separating physical or natural constraints. The importance of Marr's observation was not duly noted by mechanists at first, but in recent years, several authors have advocated the necessity of referring to various types of constraints, either in explaining neuronal mechanisms (cf. Weiskopf, 2016) or in explaining wide cognition (Miłkowski et al., 2018).

phenomenon, *resp*. mechanism) component because "mechanical systems inherently contain a 'thicket' of constraints" (Winning, 2020, p. 20).[18]

Bechtel (2018, 2019, 2021), Bechtel and Bollhagen (2021), Winning and Bechtel (2018), and Winning (2020) emphasize the need to refer to constraints, linking them with the necessity to include both constraints and energy flows as those elements which, apart from entities and activities, are relevant for the explanation of mechanisms at higher levels of organization.[19] It is the constraints and the flows of free energy that make living organisms "dissipative structures",[20] which means "that they actually use the second law of thermodynamics to their advantage to maintain their organization" (Winning & Bechtel, 2018, p. 3; cf. Moreno & Mossio, 2014). This way, living organisms—unlike most "things"—develop while maintaining their autonomy, rather than being degraded by the flow of energy and interaction with the environment.[21]

Biological mechanisms derive their causal efficacy from being constrained systems: "An active causal power exists when a system within a larger system is internally constrained in such a way as to externally constrain under certain conditions" (Winning, 2020, p. 28). In other words, constraints determine the causal powers of mechanisms in such a way that they direct the flows of free energy so that biological systems may remain in a state of energy non-equilibrium with the environment. Such mechanisms are part of a heterarchical network of controllers that guarantees the biological autonomy of a given system. Based on this, mechanisms are systems of constraints that restrict the flow of free energy to perform work (Bich & Bechtel, 2021, p. 2).

Mechanisms are active and serve to maintain the autonomy of biological systems as a result of the constrained flows of free energy. Including these kinds of constraints in the explanation of activities means breaking with the standard account of mechanistic explanation (systems tradition).[22] If the energetic dimension is ignored, "at some point, such research typically bottoms out" and "this process leaves the active nature

[18] It is important that such a view to constraints is conditioned by the research perspective. However, an explanatory strategy that favors certain constraints at the expense of others must be distinguished from the fact that these constraints exist and define a given organism or structure (Pattee, 1972).

[19] "Higher-level activities, just as those at the bottom-out level, depend upon the release of energy. Higher-level entities also constrain those at the bottom level, determining how energy released in molecular motors, ion pumps, etc. results in activities at higher levels" (2021, p. 21).

[20] Far from the equilibrium state, these are stable stationary states, the formation of which is accompanied by an increase in order.

[21] For the purposes of the analyses, I assume that biological autonomy and the related self-organization and integrity (which enable living organisms (systems) to achieve, maintain, and propagate a high degree of complexity) define the "situatedness" of biological systems in their environment and their "grounding" in thermodynamics. Thanks to this, biological systems do not disintegrate: they construct, maintain and replicate themselves in a changing environment. It means that an organism lives as long as it remains in an energetic non-equilibrium with the environment (cf. Friston & Stephan, 2007; Moreno & Mossio, 2014). A paradigmatic example of such a system is a living cell that uses metabolic processes to convert energy and materials from the environment into chemical energy and organic molecules, which are essential for the processes that keep the cell alive. All living autonomous organisms "must procure matter and energy from their environment and use these to construct and repair themselves" (Bich & Bechtel, 2021, p. 1).

[22] Earlier, Darden (2006, p. 272) drew attention to this, claiming that the process of decomposition of selected mechanisms consists in constructing, evaluating and revising them in relation to empirical and experimental limitations. In other words: constraints limit the space of possible mechanisms to a specific area that the model is to reconstruct (cf. Craver, 2007).

of activities unexplained" (Bechtel & Bollhagen, 2021, p. 17) because "a completely unconstrained system will have no behaviors; it would simply be a disorganized motion of particles" (Winning & Bechtel, 2018, p. 7). The approach that takes into account the need to refer to constraints and flows of free energy will be referred to as the 'constraint-based mechanisms approach' and its postulate as heuristics of constraint-based mechanisms. It is important to emphasize that this approach is not so much a break with the systems tradition, but its significant modification.[23]

3.2 What about predictive processing?

In §1, I have already discussed the mechanistic commitments of PP. We can now take the next step. From the point of the view of the constraint-based mechanisms approach we should note that, if PP explains its phenomena mechanistically, then it is legitimate to ask whether the mechanistic explanations based on the PP framework include constraints and the energy dimension as their constitutive component. This is not a trivial or secondary question, because, according to the heuristics of constraint-based mechanisms, mechanistic PP *should* also include energy processes. This case is not obvious. Let us note, however, that there are indications that the above heuristic is used by researchers working in the PP framework.

On the one hand, many of PP's supporters use the term "constraint" in their considerations to refer to perceptual inference in the brain. For example, "the only constraint on the brain's causal inference is the immediate sensory input" (Hohwy, 2013, p. 14), but "immediate sensory input is not the only constraint; there are, in addition, general beliefs about the world, specific hypotheses about the current state of the world, and ongoing sensory input" (Anderson, 2017, p. 3) and "perceptual experience is determined by the mutual constraint between the incoming sensory signal and ongoing neural and bodily processes, and no aspect of that content can be definitively attributed to either influence" (Anderson, 2017, p. 17). It is also worth adding that the levels of bidirectional hierarchical structure are constraints for each other (Clark, 2013, p. 183; cf. Gordon et al., 2019). Conversely, some have suggested that "without independent constraints on their content, there is a significant risk of post hoc model-fitting" (Williams, 2020, p. 1753). However, it is not clear in what sense these authors use this term and whether they use it in the same way.[24]

These various uses of the concept of constraint are difficult to relate directly to the understanding of constraints as control mechanisms, which I defend in this paper. The constraints discussed by these authors, however, reveal the non-trivial commitment of PP. Namely: the functioning of predictive mechanisms depends on the existence of

[23] This modification assumes the need to analyze (at least) some mechanisms in terms of heterarchical organization and network organization of constraints.

[24] One can also point to the "model" understanding of the concept of constraint concerning the very architecture of model building in PP (Millidge et al., 2020). It is worth adding that Sprevak has recently drawn attention to the difficulties faced by PP regarding the inclusion of the explanation of constraints: "In general, it is not obvious how predictive coding should reconcile two opposing forces: (i) permitting the implementation to be complex, idiosyncratic, and varied in ways that we do not yet understand; and (ii) imposing some constraints on which physical states do and do not implement the model in order to render the view empirically testable" (Sprevak, 2021, p. 26).

various types of constraints, which on the one hand limit the content of the generative model, and on the other hand, enable its adaptation to the environment, making it an effective adaptive tool to maintain the autonomy of the organism. The perspective I defend allows us to specify the functions of constraints in PP and to study them in a more systematic way. What is important is the question of how certain constraints are constitutive of predictive mechanisms. In other words, the point is to demonstrate how such and such organization of predictive mechanisms constrains free energy so that it is possible to perform the work required to generate particular phenomena, *resp*. predictions.

On the other hand, broadly speaking, we have to note that the findings within the FEP and NESS mathematics (expressed in the language of DST)—according to which, if something exists then it must exhibit properties *as if* it is optimizing a VFE—*look like* they coincide with the heuristics of constraint-based mechanisms whereby mechanisms are active and serve to maintain the autonomy of biological systems as a result of the constrained release of free energy. It seems that mechanistic PP should take into account the energetic dimension of predictive mechanisms. Is it really so? The full answer to this question depends on further empirical solutions, and it is certainly not only an a priori answer. Nevertheless, I argue that if the arguments presented above are correct, then it should be asked if FEP-based PP meets the requirements of the constraint-based mechanisms approach and allows one to think of predictive mechanisms as constitutive control mechanisms for autonomous systems armed with a generative model. I will devote my further analysis to answering this question.

4 What does it mean for the system that it minimizes variational free energy?

The connection between PP and the FEP raises a number of doubts, which can be reduced to two main issues: (1) the very interpretation of the FEP as a principle of modeling self-organizing systems armed with generative models; (2) the question of how the FEP determines the energetic (in the information-theoretical sense) constraint for the mechanistic PP. Let me start by outlining the first difficulty. I will devote another section to the second.

I stated earlier that under the mathematical framework of the FEP, PP looks like it coincides with the heuristics of constraint-based mechanisms. But why do I use the terms "looks like" and "as if"?[25] I do it because this is how some proponents of the FEP define its application to autonomous systems: "physical systems that look as if they encode probabilistic beliefs about the environment"; "self-organising system that looks as if it is modelling its embedding environment" or "all systems that look as if they engage in inference" (Ramstead et al., 2023, pp. 1, 2, 18) and so on. What does the phrase "as if" mean? Simon McGregor defines its use as follows: "To say that something behaves 'as if' it has property X usually implies that it does not, in fact, have property X. However, there is clearly a sense in which a system possessing property X must also behave as if it had property X; it is in this, less restrictive, sense that we

[25] I use those two terms interchangeably in this context.

intend the phrase 'as if'. In other words, we classify both the regulation of temperature by a thermostat, and also the pursuit of prey by an eagle, as 'as if' agency" (McGregor, 2017, p. 72). McGregor distinguishes between two senses of "as if". In the first one ("instrumental"), the system can be described *as if* it had a given property, even though it does not actually have it, and in the second ("realistic"), it can be described *as if* it had a given property precisely because it has it.[26]

This duality allows us to see that the use of the phrase "as if" in relation to systems that are supposed to minimize VFE can be interpreted in at least several ways: from the realistic interpretation, where VFE is a quantity (or means a quantity) that is minimized by biological systems that maintain their organization – in this approach, VFE cannot be reduced to researches' construction or explained only in terms of the practice of modeling[27]; to various anti-realistic or instrumental interpretations in which the FEP is a construction devised by scientists to describe the dynamics of any self-organizing system that is at NESS with its environment without any implications for their actual causal structure. In this approach, VFE looks like a quantity that relates to the models made by scientists, while the FEP serves to designate a model structure on the basis of which specific models are constructed (cf. Andrews, 2021).[28]

The discussion so far concerning the ontological and epistemological commitments of the FEP is rich. It is worth mentioning the papers of Andrews (2021, 2022), Bruineberg et al. (2021), Kirchhoff et al. (2022), Ramstead et al. (2022) or Van Es (2021). I will not discuss them here. However, I would like to draw attention to the fact that the mechanistic perspective adopted in this paper is realistic (see §5) and therefore imposes certain theoretical commitments on the understanding of the FEP and VFE, which bring my positions closer to a realistic interpretation of the FEP, which I will call moderate.[29] It is moderate in the sense that it assumes that systems can be described as if minimizing VFE, because they implement some causal mechanism that can be

[26] Indeed, in the latter sense (as one reviewer pointed out to me), there is no need to describe a system as behaving "as if" it had a given property if this is indeed how the system behaves. However, we can still relate the phrase "as if" to our best models or simulations and assert that the given model or simulation behaves "as if" it had a certain property of the target system. In this context, the term "as if" serves to acknowledge the use of models or simulations as approximations or representations that imitate certain aspects of the target system's behavior.

[27] This interpretation assumes that systems can be described *as if* minimizing VFE, because they implement some causal mechanism that can be described (approximately) in terms of minimizing VFE, resp. long-term average prediction error. In other words, there is a definite causal pattern that is the object of scientific interpretation.

[28] In addition to the above-mentioned positions, one should also take account of the views of authors such as Williams (2021), Colombo and Wright (2021) or Colombo and Palacios (2021), who treat the FEP as (at best) a general idealization. Their views can be collectively described as eliminativism about the FEP.

[29] The notion of moderate realism I have proposed can be related to some extent to Hilary Putnam's (1978, cf. Hacking, 1981) distinction between metaphysical realism and internal realism. The first position assumes that the world contains a specific set of objects that exist independently of the human mind and the ways of its conceptual articulation. Our theories are true if they denote what the world is like. The position of internal realism (i.e., the moderate realism I propose) assumes that objects in the world depend on accepted conceptual schemes. Thus, there may be different objects, depending on the conceptual schemes adopted. This means that there is no fixed set of objects that exists independently of conceptual schemes.

described (approximately) in terms of minimizing VFE, *resp*. long-term average prediction error (see §6).[30]

Therefore, considering the goal I have set for myself in this paper, which is to apply the heuristics of constraint-based mechanisms in relation to PP and determine the energetic constraint for the mechanistic architecture proposed by this framework, it is important to acknowledge that the FEP provides a relevant (variational) constraint for the causal organization of all autonomous systems equipped with generative models, as explained mechanistically by PP. If this is true, than the FEP provides a relevant constraint for PP's scheme of mechanism.[31]

5 Mechanistic realism and the free energy principle

Many mechanists emphasize that there are objective structures in the world that are in some sense richer than mere aggregations of causes. Entities, their hierarchical-heterarchical organization and the operations binding them, produce mechanisms. The task of scientists is to identify and decompose them (cf. Bechtel, 2008; Craver, 2007; Craver & Darden, 2013; Machamer et al., 2000). This view can be called mechanistic realism. It is not a clear-cut position, as recently demonstrated by Dewhurst and Isaac (2023), because its ontological commitments are unclear. There is no space in this paper to discuss this issue in more detail, but I believe it is reasonable to say that the architecture implied by the heuristics of constraint-based mechanisms assumes a certain mechanistic realism in relation to the causal patterns present in the world (cf. Winning, 2020). In other words, the fact that production mechanisms are limited and activated in one way or another by specific constraints and flows of free energy suggests that the causal relationships between specific patterns or, in Bayesian modeling terminology, sensory signal statistics cannot be described merely as an aggregation of causes (cf. Craver & Bechtel, 2007). This means that there must be some facts about the structure of mechanisms that explain them and determine what mechanisms *should be* and what components and operations will appear at a given level of their hierarchical-heterarchical structure (Dewhurst & Isaac, 2023; cf. Craver, 2013).

[30] This interpretation should be distinguished from the approaches that treat self-organizing systems as literally minimizing VFE, while the use of the phrase "as if" implies that systems behave as if they were minimizing VFE, because in fact they implement the mechanism of VFE minimization, *resp*. long-term average prediction error (in this view, the phrase "as if" is redundant – see footnote 26). In this sense, my analysis corresponds to the critique of what Kirchhoff, Kiverstein & Robertson (2022) call the literalist fallacy. The fallacy is that the instrumentalist position is accepted or adopted due to the belief that FEP-based models are not literally mapped onto real target systems.

[31] It is important to bear in mind that based on the difference between realistic and instrumental, *resp*. antirealistic approaches to the FEP, one can distinguish between free energy minimizing systems that use gradients (VFE-users) and systems that are just minimizers of those gradients (VFE-minimizers) (Kuhn, 2022, pp. 94–95). Consequently, if there are any VFE-users that exist, they must actually minimize VFE and not just be described as minimizing VFE. This would mean that the FEP indicates an energetic constraint that has significant causal powers necessary for the implementation of specific mechanisms regulating the work of those systems. From this perspective, it is reasonable to claim that the human brain armed with the generative model is actually a VFE-user (cf. Kuhn, 2022, p. 95).

Because of this realistic nature of the mechanistic explanations, I argue that if the heuristics of constraint-based mechanisms can be applied to VFE-constrained predictive mechanisms, then realism must be assumed for the FEP. Instrumentalism imposes no commitments on the explanations regarding the architecture of the mechanisms, and treats the mechanisms themselves as useful fictions. The heuristic interpretation of the FEP defended by Gładziejewski (2019) and Harkness (2015), while not excluding realism in relation to mechanisms, denies any explanatory power to the FEP.[32] In this sense, it does not allow VFE flows to be treated as significant for the functioning of predictive mechanisms.

I can now present the realistic interpretation of what it means to say that self-organizing systems minimize VFE. The moderate realistic interpretation, which I defend, does not impose strong commitments on mechanistic architecture that would involve committing the literalist fallacy (cf. Kirchhoff et al., 2022). Moderate realism assumes that the concepts implied by Bayesian modeling are not precisely mapped to the target phenomena. Thus, they can be treated as approximations (cf. Laudan, 1981; Weisberg, 2007). Bayesian formal structures are rather non-arbitrary (in the instrumental sense) interpretations of causal patterns in the world, which, according to mechanistic realism, have specific structures that cannot be reduced to being aggregates of causes.

I argue that the proper interpretation that allows PP to be integrated with the FEP framework in accordance with the heuristics of constraint-based mechanisms, follows a moderate realistic approach to the FEP. Why? In order to answer this question, it is necessary to look at the arguments that concern the possibility of linking VFE with TFE.

It seems that the first step in demonstrating that VFE is a relevant constraint for predictive architecture has been made. To sum up: in accordance with the realistic approach to the FEP, VFE is not only a scientists' construct, but in a sense models the actual property of the target phenomena, which do not have to be treated as exact representations of formal structures.

We thus come to the second difficulty, which I indicated at the beginning of §4: does the FEP determine the energetic (in the information-theoretical sense) constraint for the mechanistic PP, and to what extent?

6 Is variational free energy the same thing as thermodynamic free energy?

Let us first cite the observation of William Bechtel, who explicitly states that "The notion of free energy invoked in mechanical action is distinct from the free-energy principle articulated by Friston (…). The conception of free energy required in the account of mechanisms is that appealed to in mechanics to explain work of any form"

[32] Let us recall: in line with the classic view of Herbert Simon, heuristics strategies allow researchers to limit their investigations to particular regions within a given space (cf. Simon, 1977). However, it is important to emphasize that heuristics as such "cannot itself provide evidence for any particular hypothesis over an empirically equivalent alternative" (Zednik & Jäkel, 2016, p. 3969). "They are not adequate explanations" and "often provide only the illusion of understanding a mechanism" (Craver, 2006, pp. 361, 373).

(Bechtel, 2019, p. 634; cf. Bich & Bechtel, 2021, p. 52). This claim seems to exclude the idea of using VFE as a constraint for mechanistic PP, at least in the sense that Bechtel and colleagues propose. However, it seems that it is doubtful, however, whether Bechtel rightly excludes Fristonian VFE. In the quoted paper, he refers to a 2010 piece by Friston. In this work, free energy is understood as "an information theory measure that bounds or limits (by being greater than) the surprise on sampling some data, given a generative model" (Friston, 2010, p. 127) and as such it is distinguished from the thermodynamic free energy referred to by Bechtel (cf. Moreno & Mossio, 2014). However, in more recent papers, Friston argues, based on the mathematical relationships between non-equilibrium dynamics, variational inference, and stochastic thermodynamics, that VFE is the same as TFE, because VFE "is consistent with the notion of free energy as the thermodynamic energy available to do work when an ensemble is far from equilibrium" (Friston, 2019, pp. 66–67; Parr et al., 2020).[33] This statement, as I will soon show, may raise reasonable doubts and ultimately does not justify the belief that VFE is a constraint for mechanistic PP.

What is Friston's argument for equating VFE with TFE, and why is it important? I will start with the second point. Let us recall: the fact that cognitive mechanisms are active and can serve to maintain the autonomy and self-organization of biological systems is a result of the constrained flows of free energy. It is important to explain "how that free-energy is converted into a specific activity" (Bechtel & Bollhagen, 2021, p. 3). It seems that Friston goes a step further: mechanisms are constrained and made active not only by the energy in the thermodynamic sense, but also the energy in the information-theoretical sense (i.e., VFE) that the system optimizes to achieve NESS (cf. Friston, 2013; Wiese & Friston, 2021). If Friston is right, then there are some phenomena that need to be explained by taking into account the energetic constraint of VFE. This means that there are mechanisms that are implemented because they minimize the VFE quantity.

Let us now return to the identification of VFE with TFE. If VFE coincides with TFE, then it looks like the FEP (as a framework for explaining minimization of VFE) is fundamental to explaining many biological and cognitive mechanisms by analogy with the scientific importance of explanations using statistical mechanics and the concept of TFE. In the latest papers, Friston and colleagues introduce the concept of Bayesian mechanics, which "is a probabilistic mechanics, comprising tools that enable us to model systems endowed with a particular partition (i.e., into particles), where the internal states (or the trajectories of internal states) of a particular system encode the parameters of beliefs about external states (or their trajectories)" (Ramstead et al., 2023, p. 1). In other words, according to these authors, Bayesian mechanics is exactly the same as all these other mechanics but with the added variational energy constraint (i.e., the assumption of Markov blankets) (Friston, 2019, p. 122). We will therefore take a closer look at Friston's argument in favor of equating VFE with TFE. I will call it an argument from the Bayesian mechanics.

In this perspective, the assumption of the compatibility of TFE and VFE can entail both ontic commitments characteristic of the realistic interpretation of the FEP and

[33] Friston earlier integrated predictive coding with the FEP (Friston, Kilner, & Harrison, 2006) by identifying the Rao and Ballard's energy function (Rao & Ballard, 1999) with VFE.

epistemic commitments characteristic of instrumentalism. In the latter case, instrumentalism would imply treating TFE and VFE as constructs of scientists, or useful fictions. As I will argue in the remainder of this paper, however, that from the point of view of the constraint-based mechanisms approach to mechanistic PP, instrumentalism cannot be reconciled with mechanistic realism. This means, therefore, that the application of the heuristics of constraint-based mechanisms to mechanistic PP is possible only in two cases: either when the compatibility of TFE and VFE is justified in realistic terms, or when both of these quantities are treated as independent interpretations of such and such patterns or causal structures present in the world (cf. Weisberg, 2013).

6.1 The Bayesian mechanics argument

According to Friston and colleagues, the concept of VFE can only be applied on the basis of Bayesian mechanics: "At the core of Bayesian mechanics is the variational free energy principle (FEP)" (Ramstead et al., 2023, p. 2). This belief, however, reveals a deeper assumption about the nature of mechanics: every kind of mechanics has its own reified constructs (such as thermodynamic energy, temperature, second law or very VFE). It means, Friston claims, that the existence of this type of construct is justified by a given type of mechanics (classical, statistical, quantum, or Bayesian). For example, from the point of view of quantum mechanics, the temperature construct has no object reference. According to Friston, recognizing the existence of this type of reified constructs presupposes the so-called ensemble assumption (that all particles in your ensemble are exchangeable) which entails a weak coupling between fast and slow modes (Friston, 2019, p. 47).[34] In statistical physics or thermodynamics, the ensemble assumption is an idealization according to which there are collections of a very large number of systems in different (quantum) states with common macroscopic attributes. The ensemble is distinguished by which thermodynamic variables are held constant (cf. Gibbs, 1902). This means that their properties result from the laws of classical or quantum mechanics. The ensemble assumption, Friston argues, translates into a weak coupling between internal particles and their Markov blanket, which means that the states of the ensemble are partitioned so that the states of each constituent particle can be identified with the homologous states of another. This makes it possible to associate the NESS density with an ensemble density. This means that instead of describing the probability of a given particle appearing in a certain state over time, the NESS density describes a greater number of particles that occupy the same (or adjacent) states (Friston, 2019, p. 64).

So, how does the use of the ensemble assumption in Bayesian mechanics differ from its use in other mechanics? Friston claims that Bayesian mechanics adds a variational energy constraint (i.e., assumption of Markov blankets). With this additional constraint in place, one can speak of states of something as relative to something else, which is directly applicable to living organisms or neural structures. According to Friston, only

[34] For example: mechanisms that underwrite self-organization rest upon bottom-up causation and top-down causation, which means top and bottom-up causation is necessary in the sense that it defines what variables and relevant variables (in the language of the renormalization group) matter (define the coupling and the shape of the coupling). Top down causation means that these variables also have a very slow dynamic, and crucially contextualize and constrain the dynamic at the lower faster level (cf. Ellis, 2012).

Bayesian mechanics can do this (cf. Parr et al., 2020). The other types of mechanics assume that a Markov blanket and the states outside the blanket can be ignored, which is related to, for example, talking about a heat bath or a thermal reservoir in terms of statistical mechanics (Friston, 2019, p. 122).

This is where the important question arises as to why only Bayesian mechanics should allow the separation of internal and external states. Why should this not be possible to achieve through, for instance, a constraint-based mechanistic approach as understood by Bechtel and colleagues or the biological autonomy approach as characterized by Barandiaran, Moreno, Varela and so on? According to Friston and colleagues, it is important to bear in mind that the above-mentioned approaches already assume solutions that are only enabled by the Bayesian mechanics. The new mechanical philosophy of neural mechanisms or an account of biological autonomy based on autopoiesis are only possible on the basis of the ensemble assumption with a variational constraint. To be more precise: the statement that there are some mechanisms, presupposes the mechanistic nature of certain phenomena. Friston claims that without the ensemble assumption, it seems impossible. Nevertheless, it is not difficult to see that the ensemble assumption follows from the assumptions of statistical mechanics, thermodynamics, or even mechanistic realism (cf. Dewhurst & Isaac, 2023). However, Friston argues, it is only on the basis of Bayesian mechanics that one can recognize active mechanisms (e.g., information processing neuronal mechanisms) that are characteristic for the organization of living systems such as, for example, bacteria and our brains (Friston, 2019, p. 1). In other words, only Bayesian mechanics allows us to explain why biological systems "*exist* the way they do" (Sakthivadivel, 2022, p. 2), i.e., indicate what physical mechanisms and constraints enable biological systems to be what they are, rather than being inanimate matter.

According to Friston, VFE can be applied only in the realm of Bayesian mechanics and thus refers to autonomous or active things, while TFE can only be applied in the realm of the statistical ensemble. Thus, both of these mechanics are based on quantum mechanics (cf. Friston et al., 2022, pp. 5–6). For this reason, it can be said that VFE and TFE are two consequences or expressions of the same *thing* of a more elemental mechanistic or quantum nature.[35]

6.2 Instrumental interpretation of the Bayesian mechanics argument

Note that, if the Bayesian mechanics argument is valid, then VFE is an explanatory relevant constraint for the PP's mechanistic architecture. This means that according to the heuristics of constraint-based mechanisms, predictive mechanisms are active because they are a result of the constrained release of free energy (both in terms of TFE that crucial for physical mechanisms and VFE as constitutive for information processing neural mechanisms). From my point of view the main difficulty in accepting this argument lies in its instrumental interpretation defended by Friston and colleagues (cf. Friston, 2019; Friston et al., 2022).

[35] "The ensuing Bayesian mechanics is compatible with quantum, statistical, and classical mechanics and may offer a formal description of lifelike particles" (Friston, 2019, p. 1).

In instrumental interpretation this argument assumes that TFE and VFE turn out to be two sides or aspects of some more primal dynamics, which, depending on the measurement tools, in one case reveals properties are thermodynamic, in another variational. In this sense, "Bayesian and stochastic mechanics are equivalent formulations of the same thing. One can either regard Bayesian inference is a necessary consequence of thermodynamics (i.e., gradient flows on a thermodynamic potential). Alternatively, Bayesian mechanics is a corollary of thermodynamics" (Friston, 2019, p. 119). As Friston claims, each type of mechanics posits a different kind of reified constructs, and what they all have in common are random dynamic systems. This means that VFE and TFE can be understood as constructs that are relativized to the description and method of measurement, and each type of mechanics is a complementary description of the behavior of dynamic systems.

Therefore, we should distinguish the map (models developed by science) from the territory (what the models represent) (cf. Friston, 2019, p. 123; Andrews, 2021). In the instrumental interpretation, the FEP allows for the construction of "a map of that part of the territory which behaves as if it were a map" (Ramstead et al., 2022, p. 8). In this sense, VFE is a tool that is used to explain the dynamics of self-organizing systems (given the state of our knowledge) (Ramstead et al., 2022, p. 17) without making any ontological commitments regarding the representational or architectural properties of these systems. Therefore, the FEP is only a tool for modeling phenomena. It is arbitrary in the sense in which the choice of measurement tools or labels to name objects is arbitrary.

From this perspective, FEP-based models address the causal structure of the world in the sense that they are epistemically useful. Their use in modeling some empirical data may speak in their favor (cf. Smith et al., 2022). However, it is difficult to talk about their mechanistic character in this case (at least if mechanisms are understood ontically). This means that the constructions postulated by the FEP or PP can be treated as useful fictions (cf. Ramstead et al., 2020; van Es, 2021; van Es & Hipólito, 2020).

We have to conclude that instrumental interpretation does not allow for a satisfactory mechanistic integration of the FEP and PP from the perspective of the constraint-based mechanisms approach, because instrumentalism imposes no mechanistic commitments regarding the causal structures under study. Therefore, it is challenging to regard it as compatible with the earlier-discussed mechanistic realism, which I deemed normative for the using of heuristics of constraint-based mechanisms (cf. §5). In such a situation, the only possible position justifying the mechanistic integration of the FEP and PP seems to be moderate realism about the FEP. Is it really so?

7 Moderate realism about the free energy principle and predictive processing

According to the moderate realistic interpretation of the FEP, the system minimizes VFE because it implements *some* causal mechanism that can be described (approximately) in terms of minimizing VFE. In this sense, VFE can be treated as a constraint of such active mechanisms, which researchers explain in terms of the minimization of long-term average prediction errors. In other words, there are causal structures whose

organization cannot be reduced to an aggregation of causes and must be explained in terms of mechanisms constrained by quantity flows described in terms of minimizing VFE or maximizing mutual information between sensory states and internal states (cf. Friston, 2010; Friston et al., 2022).[36] In the moderate interpretation that I defend, this means that some mechanisms are systems of constraints that restrict the flow of information to perform work (cf. Bich & Bechtel, 2021, p. 2) in such a way that they minimize the discrepancy (i.e., prediction error) between estimate-based predictions of the system and the actual sensory stimulation coming from the input to stay at NESS. Why are these systems VFE-users and not just prediction error-users? Because, the minimization of prediction errors by the approximation of Bayesian inference happens through VFE minimization (cf. §2).

The argument from neural computation supports the adoption of a moderate realistic interpretation of the FEP. According to this argument, there is a trade-off between neural information processing and thermodynamic energy consumption, the explanation of which makes it possible to understand how some states of biological systems have characteristically low Shannon entropy, which enables them to adapt and survive in the environment.

7.1 Argument from neural computation

Research on the thermodynamics of information clearly indicates the existence of a trade-off between neural information processing and thermodynamic energy consumption. There is an energetic cost of information processing (cf. Levitin, 1998; Niven & Laughlin, 2008; Sagava & Ueda, 2011). This energy cost can be associated both with Landauer's principle, according to which information erasure increases the entropy of the environment, i.e., energy dissipation (Landauer, 1961; cf. Sartori et al., 2014), and with Gregory Bateson's observation that information (a single bit of information) is a difference which makes a difference, which in the case of living organisms means that the power of a given process by metabolic energy depends precisely on the difference (information) contained in certain states of the organism. For this reason, Bateson claims that the mechanical interaction of muscles can be treated as a computational model (Bateson, 1987, p. 322).

It has recently been shown that the minimum energy required by a biological sensor to detect a change in an environmental signal is proportional to the amount of information processed during this event (Sartori et al., 2014). Sengupta et al. (2013) proved that minimizing VFE is a significant constraint to the tendency to maximize both metabolic and statistical efficiency in the sense that the motivation for minimizing VFE is to maintain a constant external environment that is encoded by the physical variables measured by TFE. Thus, the reference to the VFE constraint allows for the explanation of the homeostatic nature of neural processes, which mathematically means that states of biological systems have characteristically low Shannon entropy, understood—according to the ergodic theory—as the long-term average of self-information or surprise. Without reference to informational VFE, we would not

[36] Such mechanisms can be associated with the existence of systems that I previously defined after Peter Kuhn as VFE-users (see footnote 31).

be able to explain not only the homeostatic nature of neural computational mechanisms, but also their energy consumption, which is related to their ability to transmit information (cf. Laughlin, 2001). In other words, from this perspective, it follows that the use of only thermodynamics to explain the work of the brain is not fully justified.

A full explanation of how the brain works, i.e., what makes neural mechanisms active and able to perform their functions, requires taking into account information constraints that can be characterized in terms of VFE minimization. They are responsible for action potentials in the brain's sensory system, forming a neural code that efficiently represents sensory information by minimizing the number of spikes needed to transmit a given signal according to Barlow's (1961) principle of efficient coding (cf. Abbot & Dayan, 2005, pp. 123–150).

The argument from neural computation can be formally justified by the interpretation of Jarzynski equality (Jarzynski, 1997) proposed by Friston (2019). According to Friston, Jarzynski equality shows that whenever you do any belief updating by changing the information inherent in the configuration of any dynamical system (e.g., belief updating in the Bayesian generative model), there is necessarily a thermodynamic work cost. Any Bayesian belief updating involves a change in biophysical encoding of these beliefs, or any belief updating has to have a concomitant energy expenditure in terms of thermodynamic free energy. Furthermore, it is important to highlight that this thermodynamic cost we actually measure that in brain imagining using brain mapping to detect the thermodynamic activity in terms of activation foci in the brain (cf. Davatzikos et al., 2001).[37]

I argue that both empirical and formal findings will most probably determine that there are such phenomena (e.g., the neural computations performed by brains), the explanation of which, according to the constraint-based mechanisms approach, should take into account the energetic constraint of VFE. Otherwise, such an explanation fails to capture the characteristic properties that distinguish the biotic systems that are at NESS from those that can be thermodynamically described as a heat bath.

7.2 Ontological commitments of the moderate realism

If it is true that the free energy flows constitutive of the active mechanisms can be described in terms of minimization of VFE, then it seems that there are no formal obstacles to acknowledging that the mechanistic decomposition of generative models minimizing the average prediction error *should* refer to the minimization of VFE as a constitutive constraint for these mechanisms. For this reason, I argue that one should adopt moderate realism about the FEP and PP. Its legitimacy is supported by explanatory considerations, integration possibilities regarding PP and perhaps other research frameworks, as well as relatively weak ontological commitments regarding the architecture of target phenomena. Moderate realism allows one to maintain the quantity of VFE without incurring the debts of adopting instrumentalism.

[37] Jarzynski equality can be used in two ways. Either as formal support for the argument from neural computation, or, as suggested by Friston (2019), as justification for the choice of the Bayesian mechanics as the appropriate explanatory framework for systems armed with generative models which are "shielded" by Markov blankets. The latter solution leads, of course, to the difficulties I pointed out in my discussion of and instrumentalism about the FEP.

Let's take a closer look at these ontological commitments that result from adopting moderate realism about PP and FEP, *resp*. VFE. Firstly, this position assumes that formal structures such as generative models, VFE or TFE, are interpreted as part of explanations in the ontic sense, i.e., the exhibitions "of the ways in which what is to be explained fits into natural patterns or regularities … [and] usually takes the patterns and regularities to be causal" (Salmon, 1984, p. 293, cf. Craver, 2013). In this sense, moderate realism corresponds to mechanistic realism and the constraint-based mechanisms approach. In practice, this means that moderate realism does not map literally the formal structure (generative model or Bayesian network) onto the target phenomena, which would involve committing the literalist fallacy, but assumes that there are structures that cannot be reduced solely to the aggregation of causes and which implement some causal mechanism that can be described (approximately) in terms of generative models minimizing VFE, *resp*. long-term average prediction error. Therefore, it is important to assert that the formal structures (Bayesian modeling in our case) are *such and such*, because the world has genuinely causal structures, at least some of which are entities and activities organized to form mechanisms responsible for the phenomena that are described in terms of Bayesian optimization.

This view can be further elucidated through the findings of Kirchhoff, Kiverstein, and Robertson. These authors state that realism in science does not mean that all entities postulated by a given theory or model are literally true (Kirchhoff et al., 2022, p. 12). A theory may incorporate both "OK-entities" (such as electrons and similar entities) and "supposedly non-OK-entities" (such as numbers or theoretical ideals) (Psillos, 2011, p. 6). Consequently, it is important to acknowledge that each model includes parts that are fictional entities, which bear resemblance to target systems in various ways. These fictional entities facilitate the understanding of real system dynamics within the model (Kirchhoff et al., 2022, p. 13), but they do not themselves represent specific causal structures in a literal sense. The expectation of a literal interpretation of fictional entities gives rise to the literalist fallacy, as mentioned earlier. One such fictional entity is VFE. Therefore, process theories like PP should be viewed as approximations of the actual causal structures or patterns in the world. They are approximations due to the inherent complexity of target systems. Hence, I argue that moderate realism posits that a given model fits the data without a literal mapping. Instead, it is approximately true in relation to the data (cf. Kirchhoff et al., 2022, p. 16; Stanford, 2003).

Let us now delve into the relationship between the FEP and PP. Friston argues that Bayesian mechanics provides a "formal description of lifelike particles" (Friston, 2019, p. 1). This means that the Bayesian mechanics, by establishing a relationship between TFE and VFE, tells researchers something about mathematical models, i.e., formal structures, and only about them. Consequently, process theories such as PP are indispensable for addressing target phenomena. In line with the stance I advocate, the existence of control mechanisms that constrain the flow of free energy (both in terms of TFE and VFE) enables the formulation of theorems regarding the interplay between state theory (the FEP) and process theory (PP). Therefore, it is crucial to distinguish between three distinct elements: the FEP as a formal principle, PP as a computational modeling framework grounded in this formal principle, and the biological systems that PP is employed to model, which are independent of the FEP.

How, then, is the transition from the FEP to target phenomena possible? On one hand, if the view presented in this paper is correct, mechanistic PP, employing the heuristics of constraint-based mechanisms, is utilized to model control mechanisms and systems. One such control system is the brain, modeled by predictive coders as a hierarchical generative model that approximates Bayesian inference. On the other hand, the relationship between VFE and TFE established by Bayesian mechanics informs us about target phenomena because computational models of these systems in PP are constructed using the mathematics of the FEP. Ultimately, this implies that the position of moderate realism concerns not only the FEP and Bayesian mechanics themselves, but rather the application of the FEP in a specific process theory, such as PP, which is a concrete FEP-based model. It is important to note that the FEP, as a formal principle, does not imply any ontological commitments or resolutions (cf. Andrews, 2021).[38] These commitments and resolutions arise at the level of applying the FEP through a particular process theory. The use of the constraint-based mechanisms approach justifies why such an understanding of PP should be interpreted in terms of moderate realism.

There are also further benefits of the FEP and PP interpretation presented here. According to the position defended by Friston, FEP is a (normative) state theory that things may or may not conform to it, and PP is a process theory—a hypothesis on how that principle is realized (Friston et al., 2018, p. 21). It means that PP as the process theory provides "a possible (mechanistic) story about how the FEP is implemented in real-world, target systems" (Kirchoff et al., 2022, p. 6).[39] The proposed mechanistic integration of PP with FEP reveals that the FEP serves as a normative theory for PP, setting a norm that mechanistically non-trivial PP models *should* strive to meet, assuming the utilization of the constraint-based mechanisms approach and its heuristics. According to this norm, PP models *should* have an energetic component if they are to be mechanistic.[40]

The view I defend can be treated as a voice in the discussion on the status of PP and its relation to the FEP, because FEP not only constrains the space of possible algorithms for PP (cf. Spratling, 2017), but also indicates energetic constraint for the causal organization of all autonomous systems, including those that are armed with

[38] For this reason, it can be argued that there should ultimately be no moderate realistic interpretation of the FEP itself. However, if the perspective I am advocating is correct, then the integration of the FEP with the PP based on it can be seen as part of a broader scientific view that could align with a properly developed moderate realism. This perspective largely aligns with what Kirchhoff, Kiverstein & Robertson describe as scientific realism, which asserts that one reasonable goal of our best scientific theories and models is to provide descriptions and explanations of reality that are either literally true, probably true, or approximately true (Kirchhoff et al., 2022, p. 1).

[39] In the sense, that „The free energy minimizing dynamics at play are implemented by different kinds of mechanisms in different individual organisms and species, as a function of the coupling between their evolved phenotypes and biobehavioural patterns and the niches they inhabit and the scales under scrutiny" (Ramstead et al., 2017, p. 6). In this view, the FEP can be regarded as a target-directed model in the Weisberg sense (2013) (cf. Andrews, 2021; Kirchoff et al., 2022).

[40] It is worth adding that research on systems responding to a stochastic driving signal emphasizes that there is a profound connection between the effective use of information and efficient thermodynamic operation: "any system constructed to keep memory about its environment and to operate with maximal energetic efficiency has to be predictive" (Still et al., 2012, p. 1).

generative models and are or should be the subject of (mechanistic) explanations formulated on the basis of PP. In practice, this means that all autonomous systems that can be described in terms of (Bayesian) generative models realizing updating priors and likelihood based on (average) prediction error should be treated *as if* they approximate Bayesian inference constrained by VFE. In other words: FEP offers a normative framework for the PP process theory, and that the PP explains the (biologically reliable) implementation of the FEP in terms of hierarchical and heterarchical active mechanisms that implement the generative model.

7.3 Why the free energy principle is not a heuristic or a regulatory principle or an idealization

The analyses carried out in this paper allow to refer to various positions concerning the explanatory status of FEP and its relation to PP. If the approach proposed here is valid, it has certain consequences for a number of discussions among PP and FEP researchers (see §1). Due to the limited space, I can only give provisional answers to the questions raised.

Foremost, I think that the presented approach allows for a new way of describing the PP-FEP relationship. If the FEP refers to self-organizing adaptive systems, as described in DST and that are at NESS with their environment, then with the appropriate interpretation of the notion of mechanism, dynamical FEP models may in fact turn out to be descriptions of mechanisms: "dynamical models and dynamical analyses may be involved in both covering law and mechanistic explanations—what matters is not that dynamical models are used, but how they are used" (Zednik, 2008, p. 1459).[41] In this view, the FEP provides specific constraint for a PP's scheme of mechanism.

Therefore, it is a stronger commitment than that suggested by Gładziejewski (2019) and Harkness (2015), stating that the FEP offers (only) heuristics. The approach I propose suggests that the FEP is not so much a heuristic that can aid the process of designing experiments or constructing a space of possible mechanisms, but above all points to a constitutive constraint—VFE, which is needed "not just for mechanisms to perform work, but also to maintain the mechanisms themselves" (Winning & Bechtel, 2018, p. 11). VFE as a constraint determines the causal powers of mechanisms in such a way that the flows of (variational) free energy guarantee that biological systems may remain in a state of energy non-equilibrium with the environment. Such mechanisms are part of a heterarchical network of controllers that guarantees the biological autonomy of a given system. From this point of view, biotic mechanisms are systems of constraints that restrict the flow of free energy to perform work.[42]

[41] An example of this type of practice can be found, among others, in Badcock et al., (2019, p. 105): "mechanisms involve a dynamic, bidirectional relationship between specialized functional processing mediated by dense, short-range connections intrinsic to that scale (i.e., its local integration); and their global (functional) integration with other neural subsystems via relatively sparse, long-range (e.g., extrinsic cortico-cortical) connections".

[42] "Higher-level activities, just as those at the bottom-out level, depend upon the release of energy. Higher-level entities also constrain those at the bottom level, determining how energy released in molecular motors, ion pumps, etc. results in activities at higher levels" (Bich & Bechtel, 2021, p. 21).

For the above reasons, it is also difficult to agree with Hohwy's thesis that the FEP is a regulatory principle. Surely Hohwy is right when he states that the "FEP itself (does not) implies cognitive architecture" and adds that "notions of architecture will need to build on assumptions about the particular system in question, which will constrain processes for message passing structure" (Hohwy, 2021, p. 47). However, the constraint relationship is reciprocal: on one hand, a particular system constrains flow of VFE, and on the other hand, those flows constrain the system to perform given work. Therefore, the FEP, as an explication of the dynamics of flows of VFE, possesses a specific explanatory power in the explanation of cognitive phenomena, distinct from its regulatory function. Therefore, it is agreeable to conclude, following Tomasz Korbak, that the FEP can be regarded as a functional principle that offers a general framework for understanding the mechanisms involved in free energy minimization, which can then be further specified through concrete models applied to specific phenomena (Korbak, 2021, p. 2754).

It seems that these considerations may also shed some light on a number of critical works concerning either the FEP itself or its relationship with the PP. In *Introduction*, I referred to the papers of Williams, Colombo, Palacios and Wright. Let us recall: Colombo and Palacios (2021) emphasize that there is an inalienable tension between the "physics assumptions and properties of its biological targets", which in practice makes it impossible to use the FEP to explain living organisms or, in other words, to integrate it with models developed by mechanists and/or organicists (cf. Colombo & Wright, 2021). This objection seems to be thwarted by emphasizing, as I do in my paper, the mechanistic status of explanations of biological phenomena offered in terms of constraints and free energy flows. If, for living organisms, autonomy is a constitutive property (cf. Moreno & Mossio, 2014; Ruiz-Mirazo & Moreno, 2004; Varela, 1979), then the FEP—contrary to what Colombo and Palacios claim—offers specific constraints to mechanistic explanations formulated on the basis of biology and neuroscience, in the sense that it allows one to treat descriptions, using the language of DST, as sketches of mechanisms.

From this perspective, it is also difficult to agree with the belief of Colombo and Wright that the FEP offers a weak explanatory idealization. Even if, as these authors claim, the analyses carried out by FEP supporters can be treated as (weak explanatory) sketches of mechanisms, then in the light of the constraint-based mechanisms approach and arguments presented here, sketches of free energy flow mechanisms can be used in the formulation of schemes of mechanisms with specific explanatory powers.

Finally, let's note that conducting a detailed discussion that addresses all the aforementioned positions and responds to every objection exceeds the scope of the intended framework for this analysis. Nevertheless, I believe that the general direction of the response has been set.

8 Conclusions

In this paper, I defended the view that the FEP indicates an explanatory relevant constraint (i.e., VFE) for cognitive mechanisms that can be mechanistically explained by PP. The arguments made here were based on the postulate of some mechanists about

the need to include in the explanations such constitutive components as constraints for mechanisms and free energy flows. I found that the position defined by me as the constraint-based mechanisms approach has important implications for PP, because the actual research practice in this framework corresponds to the heuristics of constraint-based mechanisms and is related to those approaches that assume the FEP to be a normative framework for the process theory realized by PP. According to the presented approach, non-trivial PP models should include an energetic component, if they are to be mechanistic. The discussion presented here has great importance for considering the relationship between PP, the FEP, and Active Inference.

The advantage of the position I defend—moderate realism about the FEP and PP—is, firstly, that it implies only minimal commitments regarding the architecture of target phenomena; and secondly, it does not reduce the constructions used by scientists to their purely instrumental functions, recognizing them, for example, as useful fictions. I argue that the approach presented here may also contribute to the formulation of a mechanism scheme, which would be defined by a common predictive template combining various mechanisms under one PP flag. Last but not least, this approach (I believe) also enables fruitful discussions with those researchers who regard the FEP as an explanatory weak heuristic, idealization or regulatory idea, as well as with those who deny any explanatory power to the FEP.

Acknowledgements I am grateful to Majid D. Beni, Stephen Fox, Karl J. Friston, Peter Kuhn, Marcin Miłkowski, Maxwell J. D. Ramstead, Noor Sajid and Wanja Wiese for helpful comments and discussions on the draft of this paper. Previous versions of this paper was discussed at the Theoretical Neurobiology meeting, during the East European Network for Philosophy of Science 2022 in University of Tartu and at the Philosophy of Cognitive Science seminar held at the Institute of Philosophy and Sociology at the Polish Academy of Sciences. I would like to thank the organizers and participants of these events for inspiring discussions. I also thank the three anonymous reviewers for this journal for helpful discussion and for their comments on previous versions of this paper.

References

Abbot, L. F., & Dayan, P. (2005). *Theoretical neuroscience computational and mathematical modeling of neural systems*. MIT Press.

Anderson, M. L. (2017). Of Bayes and bullets: An embodied, situated, targeting-based account of predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing* (Vol. 4, pp. 1–14). MIND Group.

Andrews, M. (2021). The math is not the territory: Navigating the free energy principle. *Biology and Philosophy, 36*(3), 1–19. https://doi.org/10.1007/s10539-021-09807-0

Andrews, M. (2022). Making reification concrete: A response to Bruineberg et al. *Behavioral and Brain Sciences, 45*, e186. https://doi.org/10.1017/S0140525X22000310

Badcock, P. B., Friston, K. J., & Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews, 31*, 104–121. https://doi.org/10.1016/j.plrev.2018.10.002

Barandiaran, X., & Moreno, A. (2006). On what makes certain dynamical systems cognitive: A minimally cognitive organization program. *Adaptive Behavior, 14*, 171–185. https://doi.org/10.1177/10597123060140020 8

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). MIT Press.

Bateson, G. (1987). Steps to an ecology of mind. Chicago: The University of Chicago Press.

Bechtel, W., & Richardson, R. C. (1993/2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.

Bechtel, W. (2021). Discovering control mechanisms: The controllers of dynein. In: *PSA2020: The 27th Biennial Meeting of the Philosophy of Science Association*. Baltimore, MD, 18–22 Nov 2020. Retrieved from http://philsci-archive.pitt.edu/view/confandvol/confandvol2020PSA.html

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Routledge.

Bechtel, W. (2018). The importance of constraints and control in biological mechanisms: Insights from cancer research. *Philosophy in Science, 85*(4), 573–593. https://doi.org/10.1086/699192

Bechtel, W. (2019). Resituating cognitive mechanisms within heterarchical networks controlling physiology and behavior. *Theory & Psychology, 29*(5), 620–639. https://doi.org/10.1177/0959354319873725.2020

Bechtel, W., & Abrahamsen, A. (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science, 41*(3), 321–333. https://doi.org/10.1016/j.shpsa.2010.07.003

Bechtel, W., & Bich, L. (2021). Grounding cognition: Heterarchical control mechanisms in biology. *Philosophical Transactions of the Royal Society b., 376*, 20190751. https://doi.org/10.1098/rstb.2019.0751

Bechtel, W., & Bollhagen, A. (2021). Active biological mechanisms: transforming energy into motion in molecular motors. *Synthese*. https://doi.org/10.1007/s11229-021-03350-x

Beni, M. D. (2021). A critical analysis of Markovian monism. *Synthese, 199*, 6407–6427. https://doi.org/10.1007/s11229-021-03075-x

Bich, L., & Bechtel, W. (2021). Mechanism, autonomy and biological explanation. *Biology and Philosophy, 36*(53), 1–28. https://doi.org/10.1007/s10539-021-09829-8

Bickhard, M. H. (2003). Process and emergence: Normative function and representation. In J. Seibt (Ed.), *Process theories: Cross disciplinary studies in dynamic* (pp. 121–155). Dordrecht: Springer.

Bruineberg, J., Dolega, K., Dewhurst, J., & Baltieri, M. (2021). The emperor's new Markov blankets. *Behavioral and Brain Sciences, 45*, e183. https://doi.org/10.1017/S0140525X21002351

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

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

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

Colombo, M., & Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biology and Philosophy, 36*(41), 1–26. https://doi.org/10.1007/s10539-021-09818-x

Colombo, M., & Wright, C. (2021). First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese, 198*, 3463–3488. https://doi.org/10.1007/s11229-018-01932-w

Craver, C. F. (2006). When mechanistic models explain. *Synthese, 153*, 355–376. https://doi.org/10.1007/s11229-006-9097-x

Craver, C. F. (2007). *Explaining the brain*. University Press, Oxford.

Craver, C. F. (2013). The ontic account of scientific explanation. In M. I. Kaiser, O. R. Scholz, D. Plenge, & A. Hüttemann (Eds.), *Explanation in the special sciences: The case of biology and history* (pp. 27–52). Springer Verlag.

Craver, C., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy, 2*, 547–563. https://doi.org/10.1007/s10539-006-9028-8

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press.

Craver, C. F., & Kaplan, D. (2018). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science, 71*(1), 287–319. https://doi.org/10.1093/bjps/axy015

Cumming, G. S. (2016). Heterarchies: Reconciling networks and hierarchies. *Trends in Ecology & Evolution, 31*(8), 622–632. https://doi.org/10.1016/j.tree.2016.04.009

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy, 72*, 741–764. https://doi.org/10.2307/2024640

Darden, L. (2006). *Reasoning in biological discoveries*. Cambridge University Press.

Davatzikos, C., Li, H. H., Herskovits, E., & Resnick, S. M. (2001). Accuracy and sensitivity of detection of activation foci in the brain via statistical parametric mapping: A study using a PET simulator. *NeuroImage, 13*(1), 176–184. https://doi.org/10.1006/nimg.2000.0655

Davies, P. C. W. (2019). *The demon in the machine: How hidden webs of information are solving the mystery of life*. The University of Chicago Press.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation, 7*, 889–904.

Dewhurst, J., & Isaac, A. M. C. (2023). The ups and downs of mechanism realism: Functions, levels, and crosscutting hierarchies. *Erkenntnis, 88*, 1035–1057. https://doi.org/10.1007/s10670-021-00392-y

Ellis, G. F. R. (2012). Top-down causation and emergence: Some comments on mechanisms. *Interface Focus*. https://doi.org/10.1098/rsfs.2011.0062

Feynman, R. P. (1998). *Statistical mechanics: A set of lectures*. Avalon Publishing.

Fodor, J. A. (1968). *Psychological explanation*. Random House.

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

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

Friston, K. J. (2011). What is optimal about motor control? *Neuron, 72*(3), 488–498. https://doi.org/10.1016/j.neuron.2011.10.018

Friston, K. J. (2012). A free energy principle for biological systems. *Entropy, 14*, 2100–2121. https://doi.org/10.3390/e14112100

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society, Interface, 10*, 1–12. https://doi.org/10.1098/rsif.2013.0475

Friston, K. J. (2019). A free energy principle for a particular physics. arXiv 2019, arXiv:1906.10184.

Friston, K. J., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2022). The free energy principle made simpler but not too simple. Preprint arXiv:2201.06387.

Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation, 29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912

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

Friston, K. J., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology, 100*(1–3), 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Friston, K. J., & Stephan, K. E. (2007). Free energy and the brain. *Synthese, 159*, 417–458. https://doi.org/10.1007/s11229-007-9237-y

Friston, K. J., Wiese, W., & Hobson, J. A. (2020). Sentience and the origins of consciousness: From Cartesian duality to Markovian monism. *Entropy, 22*, 516–516. https://doi.org/10.3390/e22050516

Gibbs, J. W. (1902). *Elementary principles in statistical mechanics*. Charles Scribner's Sons.

Gładziejewski, P. (2019). Mechanistic unity and the predictive mind. *Theory & Psychology, 29*(5), 657–675. https://doi.org/10.1177/0959354319866258

Glennan, S., & Illari, P. (Eds.). (2018). *The Rutledge handbook of mechanisms and mechanical philosophy*. Routledge.

Gordon, N., Tsuchiya, N., Koenig-Robert, R., & Hohwy, J. (2019). Expectation and attention increase the integration of top-down and bottom-up signals in perception through different pathways. *PLoS Biology, 17*(4), e3000233. https://doi.org/10.1371/journal.pbio.3000233

Gregory, R. (1966). *The intelligent eye*. McGrawy Hill.

Hacking, I. (1981). Experimentation and scientific realism. *Philosophical Topics, 1*(13), 71–87.

Harkness, D. L. (2015). From explanatory ambition to explanatory power—A commentary on Jakob Hohwy. In T. Metzinger & J. M. Windt (Eds.), *Open MIND, 19(C)* (pp. 1–7). MIND Group.

Harkness, D. L., & Keshava, A. (2017). Moving from the what to the how and where—Bayesian models and predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing, 16* (pp. 1–10). MIND Group.
Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.), *Open MIND, 19(T)* (pp. 1–22). MIND Group.
Hohwy, J. (2016). The self-evidencing brain. *Noûs, 50*(2), 259–285. https://doi.org/10.1111/nous.12062
Hohwy, J. (2020). New directions in predictive processing. *Mind & Language, 2*(35), 209–223. https://doi.org/10.1111/mila.12281
Hohwy, J. (2021). Self-supervision, normativity and the free energy principle. *Synthese, 199*, 29–53. https://doi.org/10.1007/s11229-020-02622-2
Hooker, C. A. (2013). On the import of constraints in complex dynamical systems. *Foundations of Science, 18*(4), 757–780. https://doi.org/10.1007/s10699-012-9304-9
Illari, P. & Williamson, J. (2013). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119–135. https://doi.org/10.1007/s13194-011-0038-2
Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Physical Review Letters*, *78*, 2690. https://doi.org/10.1103/PhysRevLett
Kaplan, D. M., & Bechtel, W. (2011). Dynamical models: An alternative or complement to mechanistic explanations? *Topics in Cognitive Science, 2*(3), 438–444. https://doi.org/10.1111/j.1756-8765.2011.01147.x
Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy in Science, 78*, 601–627. https://doi.org/10.1086/661755
Keller, G. B., & Mrsci-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron, 2*(100), 424–435. https://doi.org/10.1016/j.neuron.2018.10.003
Kirchhoff, M. D., Kiverstein, J., & Robertson, I. (2022). The literalist fallacy and the free energy principle: Model-building, scientific realism, and instrumentalism. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/720861
Kirchhoff, M., Parr, T., Palacios, E., Friston, K., & Kiverstein, J. (2018). The Markov blankets of life: Autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface, 15*, 1–11. https://doi.org/10.1098/rsif.2017.0792
Korbak, T. (2021). Computational enactivism under the free energy principle. *Synthese*, *198*, 2743–2763. https://doi.org/10.1007/s11229-019-02243-4
Koutroufinis, S. A. (2017). Organism, machine, process: Towards a process ontology for organismic dynamics. *Organisms: Journal of Biological Sciences, 1*(1), 23–44. https://doi.org/10.13133/2532-5876_1.8
Kuhn, P. (2022). The world from within: an investigation into the hard problem of consciousness from the perspective of Bayesian cognitive science. Dissertation draft. Retrieved from https://philpapers.org/rec/KUHTWF
Landauer, R. (1961). Dissipation and heat generation in the computing process. *IBM Journal of Research and Development, 5*, 183–191.
Laudan, L. (1981). A confutation of convergent realism. *Philosophy in Science, 1*(48), 19–49.
Laughlin, S. (2001). Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology, 11*(4), 475–480. https://doi.org/10.1016/s0959-4388(00)00237-3
Levitin, L. B. (1998). Energy cost of information transmission (along the path to understanding). *Physica d: Nonlinear Phenomena, 120*(1–2), 162–167. https://doi.org/10.1016/S0167-2789(98)00051-7
Machamer, P. K., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy in Science, 57*, 1–25.
Marr, D. (1982). *Vision: A computational approach*. Freeman & Co.
McCulloch, W. S. (1945). A heterarchy of values determined by the topology of nervous nets. *The Bulletin of Mathematical Biophysics, 7*, 89–93. https://doi.org/10.1007/BF02478457
McGregor, S. (2017). The Bayesian stance: Equations for 'as-if' sensorimotor agency. *Adaptive Behavior, 2*(25), 72–82. https://doi.org/10.1177/1059712317700501
Miłkowski, M., Clowes, R., Rucińska, Z., Przegalińska, A., Zawidzki, T., Krueger, J., Gies, A., McGann, M., Afeltowicz, Ł, Wachowski, W., Stjernberg, F., Loughlin, V., & Hohol, M. (2018). From wide cognition to mechanisms: A silent revolution. *Frontiers in Psychology, 9*(2393), 1–17. https://doi.org/10.3389/fpsyg.2018.02393

Millidge, B., Tschantz, A., Seth, A., & Buckley, Ch. L. (2020). Relaxing the constraints on predictive coding models. arXiv:2010.01047.
Millidge, B., Seth, A., & Buckley, Ch. L. (2021). Predictive coding: A theoretical and experimental review. arXiv:2107.12979.
Moreno, A., & Mossio, M. (2014). *Biological autonomy: A philosophical and theoretical inquiry*. Springer.
Niven, J. E., & Laughlin, S. B. (2008). Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology, 211*, 1792–1804. https://doi.org/10.1242/jeb.017574
Parr, T., Da Costa, L., & Friston, K. J. (2020). Markov blankets, information geometry and stochastic thermodynamics. *Philosphical Transactions of the Royal Society A, 378*(2164), 20190159. https://doi.org/10.1098/rsta.2019.0159
Pattee, H. H. (1972). Laws and constraints, symbols and languages. In C. H. Waddington (Ed.), *Towards a theoretical biology* (Vol. 4, pp. 248–258). Edinburgh University Press.
Pattee, H. H. (1991). Measurement-control heterarchical networks in living systems. *International Journal of General Systems, 18*(3), 213–221.
Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers.
Piccinini, G., & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese, 183*(3), 283–311. https://doi.org/10.1007/s11229-011-9898-4
Psillos, S. (2011). Living with the abstract: Realism and models. *Synthese, 180*, 3–17. https://doi.org/10.1007/s11229-009-9563-3
Putnam, H. (1978). *Meaning and the moral sciences*. Routledge & Kegan Paul.
Ramstead, M. J., Sakthivadivel, D. A. R., & Friston, K. J. (2022). On the map-territory fallacy fallacy. arXiv:2208.06924v1.
Ramstead, M. J. D., Badcock, P. B., & Friston, K. J. (2017). Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews, 24*, 1–16. https://doi.org/10.1016/j.plrev.2017.09.001
Ramstead, M. J. D., Friston, K. J., & Hipólito, I. (2020). Is the free-energy principle a formal theory of semantics? From variational density dynamics to neural and phenotypic representations. *Entropy, 22*(8), 889. https://doi.org/10.3390/e22080889
Ramstead, M. J. D., Sakthivadivel, D. A. R., Heins, C., Koudahl, M., Millidge, B., Da Costa, L., Klein, B., & Friston, K. J. (2023). On Bayesian mechanics: A physics of and by beliefs. *Interface Focus*. https://doi.org/10.1098/rsfs.2022.0029
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
Rock, I. (1983). *The logic of perception*. MIT Press.
Ruiz-Mirazo, K., & Moreno, A. (2004). Basic autonomy as a fundamental step in the synthesis of life. *Artificial Life, 10*, 235–259. https://doi.org/10.1162/1064546041255584
Sagava, E. T., & Ueda, M. (2011). Minimal energy cost for thermodynamic information processing: Measurement and information. *Physical Review Letters, 106*, 189901. https://doi.org/10.1103/PhysRevLett.106.189901
Sakthivadivel, D. A. R. (2022). Towards a geometry and analysis for Bayesian mechanics. arXiv:2204.11900v1.
Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.
Sartori, P., Granger, L., Fan Lee, Ch., & Horowitz, J. M. (2014). Thermodynamic costs of information processing in sensory adaptation. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1003974
Sengupta, B., Stemmler, M. B., & Friston, K. J. (2013). Information and efficiency in the nervous system—A synthesis. *PLoS Computational Biology, 9*(7), e1003157. https://doi.org/10.1371/journal.pcbi.1003157
Seth, A. K. (2015). The cybernetic Bayesian brain—From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. M. Windt (Eds.), *Open MIND: 35(T)* (pp. 1–24). MIND Group.
Silberstein, M., & Chemero, A. (2013). Constraints on localization and decomposition as explanatory strategies in the biological sciences. *Philosophy in Science, 5*(80), 958–970. https://doi.org/10.1086/674533
Simon, H. A. (1977). *Models of discovery*. Boston Studies in the Philosophy of Science, vol 54. Springer. https://doi.org/10.1007/978-94-010-9521-1_16

Simon, H. (1969). *The sciences of the artificial*. MIT Press.
Smith, R., Friston, K. J., & Whyte, C. (2022). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology, 107*, 102632. https://doi.org/10.1016/j.jmp.2021.102632
Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition, 112*, 92–97. https://doi.org/10.1016/j.bandc.2015.11.003
Sprevak, M. (2021). Predictive coding IV: The implementation level. [Preprint]. Retrieved from http://philsci-archive.pitt.edu/id/eprint/19669
Stanford, K. (2003). Pyrrhic victories for scientific realism. *The Journal of Philosophy, 100*(11), 553–572.
Stepp, N., Chemero, A., & Turvey, M. T. (2011). Philosophy for the rest of cognitive science. *Topics in Cognitive Science, 2*(3), 425–437. https://doi.org/10.1111/j.1756-8765.2011.01143.x
Still, S., Sivak, D. A., Bell, A. J., & Crooks, G. E. (2012). Thermodynamics of prediction. *Physical Review Letters, 109*, 120604. https://doi.org/10.1103/PhysRevLett.109.120604
Ueltzhöffer, K. (2019). Retrieved 27 Nov 2021, from https://kaiu.me/2019/10/09/life-and-the-second-law/
van Es, T., & Hipólito, I. (2020). Free-energy principle, computationalism and realism: A tragedy. Preprint.
Van Es, T. (2021). Living models or life modelled? On the use of models in the free energy principle. *Adaptive Behavior, 29*(3), 315–329. https://doi.org/10.1177/1059712320918678
Varela, F. (1979). *Principles of biological autonomy*. Elsevier.
Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy, 104*(102), 639–659.
Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford University Press.
Weiskopf, D. A. (2016). Integrative modeling and the role of neural constraints. *Philosophy in Science, 83*, 674–685. https://doi.org/10.1086/687854
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. (2020). Predictive coding and thought. *Synthese, 197*, 1749–1775. https://doi.org/10.1007/s11229-018-1768-x
Williams, D. (2021). Is the brain an organ for free energy minimisation? *Philosphical Studies*. https://doi.org/10.1007/s11098-021-01722-0
Winn, J., & Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research, 6*, 661–694.
Winning, J. (2020). Mechanistic causation and constraints: Perspectival parts and powers, non-perspectival modal patterns. *The British Journal for the Philosophy of Science, 71*, 1385–1409. https://doi.org/10.1093/bjps/axy042
Winning, J., & Bechtel, W. (2018). Rethinking causality in biological and neural mechanisms: Constraints and control. *Minds and Machines, 2*(28), 287–310. https://doi.org/10.1007/s11023-018-9458-5
Zednik, C. (2008). Dynamical models and mechanistic explanations. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the annual conference of the cognitive science society* (pp. 1454–1459). Cognitive Science Society.
Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese, 193*, 3951–3985. https://doi.org/10.1007/s11229-016-1180-3