This paper is under review. All comments are welcome.

Title:

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

Author:

Michał Piekarski, PhD

Author affiliation:

Cardinal Stefan Wyszyński University in Warsaw, Poland

m.piekarski@uksw.edu.pl

ORCID: 0000-0002-9482-526X

Acknowledgements:

I am grateful to Majid D. Beni and Maxwell J. D. Ramstead for helpful comments and discussions on this paper. Draft of this paper was also discussed 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 Marcin Miłkowski and his research group for the invitation and discussion.

Abstract

There is a view emerging in the philosophy of science that research practices in science can be characterized in terms of discovering and describing mechanisms. Mechanistic explanations are based on the identifying the underlying mechanisms that generate a target phenomenon and strategies understood as decomposition of these mechanisms. Recently, there has been a discussion among mechanists about the necessity to include constraints and free energy flows into the explanations, as constitutive components of mechanistic explanations. This is directly related to the existence of control mechanisms that are non-autonomous and entail the existence of heterarchical networks. I refer to this as the 'constrained mechanisms approach'. This paper examines the extent to which this approach can be applied to the predictive processing framework, which is now an influential process theory, offering a computational description of perceptual and cognitive mechanisms in terms of hierarchical generative models approximating Bayesian inference. In other words, I examine whether the constrained mechanisms approach can be applied to the framework in which control mechanisms play an important explanatory role. I will argue that predictive processing models based on the free energy principle are amenable to this approach. In practice, this means that free energy principle offers a normative explanatory framework for predictive processing, and that in turn, this framework offers a biologically plausible account of the manner in which the principle is implemented in terms of hierarchical generative models and heterarchical active mechanisms. These analyzes are of great importance for those approaches that undermine the explanatory status of the free energy principle.

Keywords: predictive processing; mechanisms; constraints; free energy principle; explanation; normativity.

1. Introduction

A growing number of researchers in the philosophy of science agree that explanatory practice in sciences is based on discovering and describing the mechanisms that underwrite studied phenomena (Bechtel & Richardson, 1993; Bechtel, 2008; Craver, 2007; Craver & Darden, 2013; Glennan & Illiari, 2019: Machamer, Darden & Craver, 2000).[1] According to the 'systems tradition' of research (Craver, 2007), the explanation of a given phenomenon is based on the so-called mechanistic decomposition: we list a set of parts, operations, and their organization that are to be responsible for that phenomenon, and so constitute a causal mechanism for the emergence of that phenomenon (Bechtel & Richardson, 1993; Bechtel, 2008; Craver; 2007; Illiari & Williamson, 2012).

Recently, this approach to mechanism has sparked a discussion about the need to consider constraints and free energy flows in the explanations (of at least some phenomena). Supporters of this approach (Bechtel, 2018; 2019; 2020; Bechtel & Bich, 2021; Bechtel & Bollhagen, 2021; Bich & Bechtel, 2021; Winning, 2020; Winning & Bechtel, 2019) argue that the decomposition strategy assumes that mechanisms are organized in terms of composition or/and causal production. However, this is an incomplete approach, as it ignores two features of many cognitive mechanisms. Firstly, the relation between these mechanisms often exist as control, and not just composition and causation; secondly, many cognitive mechanisms are components of a complex network of heterarchically organized control systems. These observations leads to the conclusion that the mechanisms are organized not only in terms of production and composition but also in terms of control. Control mechanisms act on production mechanisms (which perform the work required to produce a given phenomenon) altering their conformation and also the work they perform. To regulate production mechanisms so that they perform their work, control mechanisms must measure relevant conditions in the organism or/and its environment or charge on information from other control mechanisms which make such measurements (Bechtel, 2021). What this means in practice is that, explaining them, one should account for the constraints in which they operate and their constitutive flows of free energy. Therefore, by constraint I understand the factors that reduce the degree of freedom of a given system with regard to the variability or the possibility of changing its parameters, components and behaviors (Umerez & Mossio, 2013). Constraints

[1] I assume that mechanisms is "a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (Bechtel & Abrahamsen, 2005, 423; Bechtel, 2008, 13).

understood in this way, can be a consequence of specific material structures, such as particles, membranes, or, for example, machines. These structures are static, that is, to some extent dependent on the laws of nature, but their behavior can only be explained by pointing to their time-dependent constraints.[2] In this sense the constraints are not only limiting but also enabling, such as the pipe not only restricts the flow of water, but also allows the water to reach places it would not reach without the use of the pipe (Hooker, 2013; Nowakowski, 2017, 2). This means that "constraints act as conditions that allow systems to exhibit behaviors that could not be exhibited without the presence of these constraints" (Raja & Anderson, 2021, 10).

In this paper, I will consider the extent to which the constrained mechanisms approach is justified in relation to the currently influential framework of predictive processing (PP) (Clark, 2013; 2016; Hohwy, 2013; 2020a; Wiese & Metzinger, 2017). PP offers a computational model of perceptual and cognitive mechanisms in terms of (variational or approximate) Bayesian inference premised on a hierarchical generative model. Many researchers agree that PP provides a mechanism sketch (i.e. an incomplete representation of a target mechanism that specifies some of the relevant entities, activities, and organizational features[3]) (cf. Gładziejewski, 2019; Hohwy, 2015; Harkness, 2015; Harkness, & Keshava, 2017), which helps to explain several different mechanisms in predictive terms. This approach is associated with the free energy principle (FEP) (Friston, 2010; 2019; Friston & Stephan, 2007; cf. Andrews, 2021). The FEP is important because it explicitly assumes that the minimization of prediction errors in PP is a special case of minimizing variational free energy (VFE). This view, however, is somewhat controversial (cf. Williams, 2021). What is relevant in this context is the question of whether the FEP provides a significant constraint to PP, binding it to flows of free energy, and thus allows PP researchers to realize the desiderata of the constrained mechanisms approach. I will devote my analyzes to this issue.

The presented paper has the following structure: in §1, I discuss the new mechanical philosophy and its characteristic systems tradition, describing explanations in terms of the identification and decomposition of mechanisms. §2 concerns the recent position based on mechanism, which I refer to as the constrained mechanisms approach. What is characteristic

[2] For this reason, they can also be referred to as modal patterns (Winning, 2020), because they are "modally loaded" (Mumford, 2004), i.e. they relate to what can happen in relation to the system or mechanism that is being constrained.
These remarks allow us to think of mechanisms in terms of dynamical systems (Winning & Bechtel, 2018, 7; cf. Van Gelder & Port, 1995). I will not elaborate on this thread here.

[3] A satisfactory explanation of the analyzed phenomenon is based only on the formulation of a mechanism scheme that relates to the actual components of the mechanism and its organization. The scheme of the mechanism is actually a complete model of a given phenomenon (Craver, 2007, 113–114).

for this approach is the so-called heuristics of constrained mechanisms. According to this approach, there are a number of mechanisms that cannot be satisfactorily explained by using decomposition as understood by the systems tradition. Cognitive mechanisms—which, depending on their type, may be understood in terms of control mechanisms, active mechanisms, homeostatic mechanisms or feedback mechanisms— to be explained adequately, they must be considered together with constraints and flows of free energy as their constitutive components. §§3-4 are further devoted to the notions of constraints, free energy and autonomy, which are key to my analysis. In §5, I discuss the PP framework and its basic concepts. I would like to point out that, according to PP, the generative model by which the supporters of this approach describe the cognitive activity of the brain does not perform explicit Bayesian inference, but rather, that its dynamics can be described as approximately Bayes optimal inference. This is an important distinction for my further analysis. In §6, I present the mechanistic commitments of the PP framework and I wonder to what extent it meets the heuristics of constrained mechanisms. These analyzes lead me to introduce the notion of VFE and the FEP (in §7). I argue that the approximation of Bayesian inference that minimizes prediction errors does in fact minimize VFE. This is an important conclusion because, as I will argue in §8, the association of PP with VFE minimization makes it possible to implement the mechanistic heuristics of constrained mechanisms. This means that the mechanistic decomposition of generative models minimizing the long-term average prediction error should refer to VFE minimization as a constitutive constraint for these mechanisms. According to this approach, the FEP not only provides constraints on the space of possible algorithms and models for PP, but also indicates energetic constraints for the causal organization of any and all autonomous systems equipped with generative models, explained mechanistically by PP. In final §9, I discuss some of the implications stemming from my view of the relationship between the PP and the FEP and refer to some critical views.

2. Mechanistic explanation: systems tradition

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 biological, physical, and social sciences (cf. Glennan & Illari 2018). This new mechanistic approach to explanation became the dominant view across many of debates in the philosophy

of science (Bechtel & Richardson, 1993; Bechtel, 2008; Craver, 2007; Craver & Darden, 2013; Machamer, Darden & Craver, 2000).

New mechanism assumes that (1) the phenomenon to be explained (*explanandum*) must be distinguished from mechanism that underlies it (*explanans*);[4] (2) the mechanism by which a given phenomenon is to be explained is detectable and describable.[5] The key to mechanistic explanations is the 3M constraint, described by David Kaplan (model – mechanism – mapping). It assumes that the mechanistic model of a target phenomenon explains that phenomenon when (1) the variables in the model correspond to identifiable components, actions, and organizational properties of the target mechanism that produces, maintains, or underwrites the phenomenon; and (2) the (possibly mathematical) relationships among these (perhaps mathematical) variables in the model correspond to causal relationships between the components of the target mechanism (Kaplan, 2011, 347). This means that, on the one hand, the phenomenon consists of certain functional components,[6] which can be assigned appropriate operations, and which are organized in such a way as to produce the phenomenon to be explained; and on the other, that the mechanism whereby a given phenomenon is explained is decomposable into individual parts that are responsible for that phenomenon (Zednik, 2008, 1454).

This should be distinguished from etiological explanations which explain a phenomenon by describing its antecedent causes; and explanations which are componential or constitutive: they explain 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).[7] The latter 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. Strictly speaking, decomposition allows one activity to be explained in terms of other activities (Bechtel & Bollhagen, 2021, 3). This strategy allowed for a satisfactory explanation of many phenomena, such as decision-making (Ratcliff & McKoon, 2008), the formation of episodic memories (Hasselmo, 2012), syntactic comprehension (Kaan & Swaab, 2002), and so on.

[4] It should be mentioned that this assumption is common to many explanation models starting with the Hempel and Oppenheim model (1948).
[5] Generally speaking, mechanisms consist of objects (components), their actions and activities (operations), and the relationships between them.
[6] Of course, there may be some complex mechanisms that do not consist of functional components, e.g. a geometric point with no highlighted parts.
[7] In this paper, by explaining I mean constitutive explanations.

Decomposition is a characteristic determinant of the ‘systems tradition‘ (Craver, 2007; cf. Bechtel & Richardson, 1993; Fodor, 1968; Cummins, 1975; Simon 1969; Wimsatt 1974). In this tradition, explanation is understood as a matter of decomposing systems into their parts to show how those parts are organized together in such a way as to emphasize the explanandum phenomenon.

3. Constraints

The 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; cf. Bechtel, 2008; Craver; 2007; Illiari & Williamson, 2012). However, mechanistic approach is not indisputable (cf. Silberstein & Chemero, 2013). Moreover, many authors oppose the dynamic approach (cf. Stepp, Chemero & Turvey, 2011) to the mechanistic approach.

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 approach (cf. Bechtel, 2018; 2019; 2020; Bechtel & Bollhagen, 2021; Winning & Bechtel, 2018; Winning, 2020). Some researchers 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, resp. organized set of causal processes is “responsible for” the implementation of another). Such an approach, 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, 2). These are therefore mechanisms that helps 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)[8]; (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, 621, cf. Pattee, 1991).[9] 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, 621).[10] Taking account of these two aspects of cognitive processes, i.e., their function in the constitution of control mechanisms and their non-autonomous character, leads to the conclusion that their explanation should also cover other components than those previously considered.[11] This means that the mechanisms are organized not only in terms of production and composition but also in terms of control. Such an approach 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, 3).

A drift from the classical understanding of the 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.[12] It is worth noting that already David Marr (1982) drew attention to the fact specific processes can be defined by indicating and separating physical or natural constraints. For example, edge detection processes are naturally constrained by spatial localization (Marr, 1982, 68-70). This means that the objects in the world that cause changes in the intensity of light are spatially located. Therefore, the explanation of the edge detection process should also address, as potential components of this mechanism, the physical constraints. Constraints can generally be understood as specific facts about the real world (cf. Shagrir 2010, 489). Such constraints should be accounted for when explaining the operation of specific mechanisms. They are

[8] "Autonomous systems allow to speak in terms of a strong sense of norm or normativity, where the nature of the norm (what is good or bad for the system) is not externally interpreted or derived from an adaptive history but defined intrinsically by the very organization of the system" (Barandiaran & Moreno, 2006, 174).

[9] "In both machines and human institutions, control mechanisms are often organized hierarchically. In a hierarchy, individual control mechanisms are themselves controlled by a 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" (Bechtel & Bich, 2021, 2). The notion of heterarchy first introduced McCulloch (1945). See also Cumming (2016).

[10] This is not to say that systems tradition does not recognize the importance of constraints (cf. Craver, 2007; Darden, 2006; 2008). However, I claim that it treats them as secondary or unconstitutional for mechanisms in the sense that Bechtel or Winning write about.

[11] It should be noted that some mechanists have discussed certain control mechanisms such as circadian mechanisms (Bechtel & Abrahamsen, 2010;) or feedback mechanisms (Bechtel, 2008, Ch. 7). Nevertheless, they did not talk much about the effects of these mechanisms on others within certain complexes.

[12] Machamer (2004) does not seem to agree with the latter.

causal and effective: they provide the necessary and sufficient conditions for the functioning of specific processes (Marr, 1982, 111-116). The importance of Marr's observation was not duly noted by mechanists at first, but in recent years, several authors have advocated for the necessity to refer to various types of constraints, either in explaining neuronal mechanisms (cf. Weiskopf, 2016) or in explaining wide cognition (Miłkowski et al. 2018).[13]

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 component, because "mechanical systems inherently contain a 'thicket' of constraints" (Winning, 2020, 20).[14]

4. Free energy and autonomy

Bechtel (2018; 2019; 2020), Bechtel & Bollhagen (2021), Winning & 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 to explanation of mechanisms at higher levels of organization. It is the constraints and the flows of free energy that make living organisms "dissipative structures",[15] which means "that they actually use the second law of thermodynamics to their advantage to maintain their organization" (Winning & Bechtel, 2018, 3; cf. Moreno & Mossio, 2014). In 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. Let's take a closer look at it.

The concept of biological autonomy has been widely discussed (Moreno & Mossio, 2014; cf. Bich & Bechtel, 2021; Rosen, 1991; Ruiz-Mirazo & Moreno, 2004; Varela, 1979). Its full explication far exceeds the present analysis. For the purposes of the analyses, I assume that biological autonomy and the related self-organization and integrity (which enable living organisms, *resp*. systems, to achieve, maintain, and propagate a high degree of complexity) define the "situatedness" of biological systems in their environment and their "grounding" in

[13] This concept is also used by Carl Craver (2007). In the theory of mechanistic explanations proposed by Craver, the space of possible mechanism is defined by such and such other entities, properties, activities and constraints that define the organization of the mechanism (e.g. gravity, the amount of energy supplied, light emission). By constraint, Craver means a discovery that either shapes the boundaries of the space of possible mechanisms or alters the probability distribution in that space, i.e. changes the probability that a model in the space of possible models accurately describes the actual mechanism.

[14] It is important that such an approach 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).

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

thermodynamics. Thanks to this, biological systems do not disintegrate: they construct, maintain and replicate themselves in a changing environment. This does not mean, however, that they are independent of the environment – on the contrary: they enter into specific interactions with it which make them organized in the way they are. In this sense, the autonomy is "the capacity of a system to manage the flow of matter and energy through it so that it can, at the same time, regulate, modify, and control: (i) internal self-constructive processes and (ii) processes of exchange with the environment. Thus, the system must be able to generate and regenerate all the constraints— including part of its boundary conditions— that define it as such, together with its own particular way of interacting with the environment" (Ruiz-Mirazo & Moreno, 2004, 240). In this approach, an organism lives as long as it remains in an energetic non-equilibrium with the environment (cf. Friston & Stephan, 2007). Autonomous systems, including living organisms, carry out processes that interact with the environment to perform such work that is intended to supply energy to the system (Bickhard, 1993). 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.[16]

Autonomy understood in this way is realized in living systems by control mechanisms which are sometimes, though not always, hierarchically organized and usually form interconnected networks where mechanisms control one another within a given biological organization, creating hierarchically organized control networks (cf. Bechtel, 2019; 2020; Pattee, 1991)[17]: "Multiple control mechanisms, each operating largely independently of each other, can operate on the same production mechanism when each is responsive to different information that allows the organism to better deploy the production mechanism in maintaining itself" (Bechtel, 2020, 31).

Let us return for a moment to the determination of biological autonomy by Ruiz-Mirazo and Moreno. They maintain that a "system must be able to generate and regenerate all the constraints (…) that define it as such". What does it mean? The autonomy and dynamics of the system depend on certain boundary conditions. What distinguishes "spontaneous" dissipative structures (e.g. artificially maintained by a researcher in a laboratory) from actual autonomous systems, is the fact that, in the first case, the flow of energy and/or matter that

[16] All living autonomous organisms „must procure matter and energy from their environment and use these to construct and repair themselves" (Bechtel & Bich, 2021, 1).

[17] The heterarchical control model is "a distributed causal network that does not define an order relation or special significance to particular local causal links" (Pattee, 1991, 220). "I use the term heterarchy when there is a large-scale violation of the features I associated with hierarchy—a transitive ordering of levels, a limit of one controller for a given controlled mechanism, and fewer controllers at higher levels" (Bechtel, 2019, 625).

keeps the system out of equilibrium is not controlled by the organization of this system, while in the latter case the constraints that actually direct the flows of energy and/or matter from the environment through the constitutive processes of the system are endogenously created and maintained by this system (Ruiz-Mirano & Moreno, 2004, 238). This means that "self-organization" and "self-maintenance" are not sufficient for the system to remain in a non-equilibrium state with the environment. What is required are mechanisms that keep the system away from the thermodynamic equilibrium channel and modulate the interaction between this system and the environment so as to keep its internal dynamics running.

This fact is well illustrated by the example of the candle discussed by Bickhard (2003). According to Bickhard, burning a candle is a "self-maintenance" phenomenon because the flame helps to keep the system at non-equilibrium (assuming the right conditions: temperature and access to oxygen). The shape the candle takes is also stable given disturbances are not too strong. Nevertheless, the candle is unable to modify its own self-maintenance processes to accommodate environmental changes that may endanger the burning of the wick. This is not the case with what Bickhard calls "recursive self-maintaining systems," like a bacteria floating in a glucose gradient or a (fictional) candle capable of extracting wax from its surroundings (cf. Campbell, 1990).[18] What is characteristic of recursive self-maintaining systems is their active interaction with the environment, which in practice means that the processes that constitute the system must be directly involved in the continuous exchange of material and energy resources with the environment, and vice versa, the achievement of these resources is necessary for these processes to emerge and be maintained. For this reason, the system is co-defined by constraints that actually guide energy/matter flows from the environment. This means that "autonomous organization is only possible if it generates constraints that modulate the flows of energy so that those constraints are regenerated and contribute in this way to the recursive maintenance of the organization" (Ruiz-Mirano & Moreno, 2004, 241) or speaking in another language: "Energy is needed not just for mechanisms to perform work, but also to maintain the mechanisms themselves" (Winning & Bechtel, 2018, 11) because "Machines and biological mechanisms constrain free energy to perform work, but unless this activity can be controlled, the work will not be useful" (Bechtel, 2018, 7).[19]

[18] An analogous example concerns adaptive snowflakes with wings discussed in Friston & Stephan, 2007.

[19] We should emphasized that free energy, which Bechtel writes about, is not the same as free energy by Karl Friston and, according to Bechtel, these two concepts cannot be combined (cf. Bechtel, 2019, 634). In this paper, contrary to Bechtel, I will argue that his analyzes are also valid for Friston's free energy (cf. §8).

The above analyses clearly show that the 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, 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 imbalance with the environment. Such mechanisms are part of a heterarchical network of controllers that guarantees the biological autonomy of a given system. In this approach, mechanisms are systems of constraints that restrict the flow of free energy to perform work (Bechtel & Bich, 2021, 2).

For the above-mentioned reasons, Bechtel and Bollhagen postulate that it is necessary to account for constraints and flows of free energy when explaining cognitive mechanisms at all levels of the hierarchy because "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, 21).[20] The mechanisms are active and serve to maintain the autonomy of biological systems as a result of the constrained release of free energy. Including these kinds of constraints in the explanation of activities means breaking with the standard account of mechanistic explanation (the systems tradition).[21] If the energetic dimension is ignored, "at some point, such research typically bottoms out" and "this process leaves the active nature of activities unexplained" (Bechtel & Bollhagen, 2021, 17), because "a completely unconstrained system will have no behaviors; it would simply be disorganized motion of particles" (Winning & Bechtel, 2018, 7). The approach that takes into account the need to refer to constraints and flows of free energy will be referred to as the 'constrained mechanisms approach' and its postulate as heuristics of constrained mechanisms. It should be emphasized that this approach is not so much a break with the systems tradition, but its significant modification.

[20] It should be added that although the concept of constraint in the sense it is used here was proposed by Pattee (1972) and later used by Marr, the same concept of constraint also appears in classical mechanics and refers to restrictions and limitations of the system's motion.

[21] Earlier, Darden (2006, 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.

5. Predictive processing: an outline

The presented approach to the analysis of active mechanisms in neuroscience proposed by the constrained mechanisms approach is of fundamental importance because it includes mechanisms that control homeostatic processes, feedback mechanisms and the circadian cycle (cf. Franklin & Wolpert, 2011). For example, the key to feedback mechanisms is the possibility that the mechanism changes its own action – that is, it acts as an agent and a patient simultaneously. Explaining these types of mechanisms is a prerequisite for explaining many life processes as well as many neural processes (Bechtel, 2008. Ch. 7) and energy conversion in the brain (Kety, 1963). "In this view, the neocortex is at the top of the hierarchy. Information procured by the senses (including proprioception) is funneled up to the neocortex where decisions for action are made and commands sent back down to lower levels of the nervous system until ultimately they are delivered to motor neurons" (Bechtel, 2020, 30-31; cf. Bechtel, 2008, Ch. 7). For this reason, it seems that the explanations formulated on the basis of cognitive sciences and neuroscience should include constraints and the energetic dimension as their constitutive component. This is due not only to the actual functioning of the mechanisms, but also to the fact that the failure to take these components into account makes it much more difficult or even impossible to determine the principled start or termination conditions. Therefore, the above analyses lead to the assumption that at least some of the findings from the latest research in the field of cognitive and systems neuroscience require rethinking. I will devote my further analyses to this issue.

The theory of mechanistic explanations is able to account for many different approaches in contemporary computational cognitive neuroscience, including the influential predictive processing (PP) framework. In this paper, I will consider how the constrained mechanisms approach can be applied to PP, where control mechanisms play an extremely important role from the level of basic neural processes to sensorimotor control and higher (Clark, 2013; 2016; cf. Franklin & Wolpert, 2011). Before moving on to this role, it is necessary to cover this research at least in a cursory manner.

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, 2020a). In this approach, 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 difference between predicted and actually sensed data). A characteristic feature of this approach 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, heuristically, 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 of the joint probability of the "observable" data E (sensory state), and H – a hypothesis about these data (trees, birds, glasses etc.). In other words, it is the product of P(H) (priors over states) and P(E/H) (likelihood of evidence probability if hypothesis is true). This means that the generative model is a statistical model of how observations are generated. 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. Thus, the generative model maps the statistical structure of a certain set of observed input data by tracing (as a result of a schematic summary – recapitulation) the causal matrix responsible for this structure. The dynamics of the coding units for such a model is used to predict the input data coming into the system. In this way, the generative model generates causes based on their effects, i.e. the content and structure of the sensory signal. The generative model continuously generates top-down predictions about the content and nature of the sensory signals generated by causes external to the model. This process is carried out at each level of the model depending on a given perceptual, cognitive or non-cognitive task. The model minimizes the so-called prediction errors, i.e. the differences between the expectations of the organism, i.e. 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. It means that updating likelihoods and priors based on prediction errors is a mechanism, and it's not the same as Bayesian inference, but ends up "looking as if" the system was doing inference, on average and over time.

Technically speaking: the generative model, according to the Bayesian rule $P(H/E)=P(E/H)P(H)/P(E)$, "as if" 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.[22] This hypothesis enables the minimization of the long-term mean prediction error (Hohwy, 2020a). 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, Friston & White, 2021, 8). Here, however, lies a fundamental difficulty.

The model inversion (mapping from causes to consequences and then inferring the causes from consequences) based on exact Bayesian inference is computationally intractable because computing the marginal likelihood P(E) (evidence) requires that the probabilities be summed up over all the states in the generative model (cf. Smith, Friston & White, 2021, 10). In practice, this means that P(E) is conditioned by all the possible hypotheses that the data can explain. For this reason, proponents of the PP framework argue that the model approximates Bayesian inference rather than computes 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 hypotheses. This way, when the model minimizes the prediction error, it also minimizes a certain *quantity* (I will come back to it later) that is always greater than or equal to 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, 294)). This means that the continuous minimization of the average prediction error by the model is possible due to the fact that the model approximates Bayesian inference. The model therefore does not directly compute true posterior distribution P(E|H), but iteratively updates the approximate posterior via gradient descent to minimize prediction errors. This way of optimizing the model allows one to find a distribution that approximates the exact posterior (an approximate posterior distribution over states makes simplifying assumptions about the nature of the true posterior distribution). This type of inference is known as variational inference. It transforms the intractable sum or integral required to perform model inversion into an optimization problem that can be solved in a computationally efficient manner (Parr, Markovic, Kiebel & Friston, 2019; cf. Smith, Friston & White, 2021, 12). The model then looks for similarities between the approximate distribution Q(E), and the generative model P(H,E), using a measure

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

called the Kullback–Leibler (KL) divergence (cf. Bogacz, 2017; Kiefer & Hohwy, 2018).[23] The model updates Q(E) until Q(E) will approximate the true posterior, while minimizing the prediction error. Let us stop at this point.

I have so far discussed the understanding of the generative model in PP and how the difficult problem of exact Bayesian inference is converted into an easy optimization problem, where the approximate posterior minimizes prediction error, under a given generative model.[24] I will now discuss the mechanistic commitments of this research framework.

6. Predictive processing in search of constrained mechanisms

It has been established that what proponents of PP framework are after are mechanistic explanations and that the various models of cognitive functions developed via PP are aimed at this kind of account (Friston, Fortier & Friedman, 2018; Gładziejewski, 2019). In line with this approach, it has been argued that PP provides the sketch of a mechanism (cf. Gładziejewski, 2019; Gordon et al. 2019; Harkness, 2015; Harkness & Keshava, 2017; Hohwy, 2015) i.e. an incomplete representation of target mechanism. 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, 2021b)) 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.[25]

The key to this type of practice is the recognition of cognition in the categories of mechanistic causal relations (cf. Gładziejewski, 2019, 665). Gładziejewski suggests that sketches of mechanisms provided by PP should be understood in the sense that these models

[23] The Kullback-Leibler divergence (also known as relative entropy) is a measure used in statistics and information theory to determine the discrepancy between two probability distributions. It is zero when the distributions match, and it gets larger the more dissimilar the distributions become.

[24] This is an important observation because it helps distinguish Bayesian PP from other Bayesian models present in the literature which were accused of dealing with problems that can be shown to be NP-hard, for example stating that it is impossible to check whether a set of beliefs is logically consistent (i.e., the satisfiability problem) (Oaksford & Chater, 2007; cf. Gigerenzer, Hoffrage & Goldstein, 2008).

[25] 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 & Mrsic-Flogel, 2018).

"share common core assumptions about relevant mechanisms" but not describe a single cognitive structure (mechanism).[26] 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, 666). Gładziejewski 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, 666-667). Regardless of which of the indicated heuristics is actually implemented by PP researchers (whether one or some combination of several), there is no doubt that many of PP supporters seek mechanistic explanations.[27] If this is the case, then it is legitimate to ask whether the mechanistic explanations formulated on the basis of 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 constrained mechanisms, PP should also include energy processes. This case is not obvious.

Many PP 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, 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, 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, 17).[28] It is also worth adding that the levels of bidirectional hierarchical structure are constraints for each other (Clark, 2013, 183; cf. Gordon et al. 2019). On the other hand, some have suggested that "without independent constraints on their content, there is a significant risk of post hoc model-fitting"

[26] It should be added that PP "rather explain cognition in terms of many different kinds of computation, it explains cognition by appeal to a single computation – one computational task and one computational algorithm are claimed to underlie all aspects of cognition" (Sprevak, 2021a, 1). If this is true, PP explains the phenomena on all three levels of description indicated by Marr.

[27] This is, of course, not a non-controversial view, as there are a number of researchers who will be willing to combine PP with other models of explanations (cf. Bruineberg, Kieverstein & Rietveld, 2018). The matter is certainly not unequivocal, but the discussion of this issue goes far beyond the scope of these analyses. I also ignore the doubts as to whether PP actually explains phenomena mechanistically (Litwin & Miłkowski, 2020).

[28] Venter (2021) adds body morphology to this list.

(Williams, 2020, 1753). However, it is not entirely clear in what sense these authors use this term and whether they use it in the same way.[29]

Recall: according to the constrained mechanisms approach, high level cognitive mechanisms are not autonomous, but are components of wide heterarchical networks that control the physiology and behavior of agents (Bechtel, 2019). The analysis carried out by Bechtel and colleagues convincingly shows that actual research practice, e.g. in biology, contains many examples "in which researchers (…) have not only garnered evidence for the occurrence of certain types of activities but have also offered explanations of these activities"; in describing a given activity they "identify [...] the source of Gibbs free energy that is utilized in the activity", to then "understand how that free-energy is converted into a specific activity" (Bechtel & Bollhagen, 2021, 3; Winning, 2020).[30] Is there a similar research practice in PP? To answer this question, I have to move on to the issue of the FEP, because many researchers associate PP with the active inference framework, which directly refers to the free energy category (cf. Hohwy, 2015; 2016; Millidge, 2019; Seth, 2015).

7. Variational free energy and predictive processing

In §5, I stated that the model minimizes the prediction error while minimizing a certain *quantity* that is always greater than or equal to surprisal. This quantity refers to the objective function that is known as VFE or an evidence lower bound (cf. Winn & Bishop, 2005). The introduction of VFE helps to convert exact Bayesian inference into approximate Bayesian inference. VFE was introduced by Richard Feynman to solve an intractable inference problem in quantum electrodynamics (Feynman, 1998, cf. Friston et al. 2006, 221). Minimization of a computable objective function will approximate the minimization of the evidence. This evidence is always upper bounded by the 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).[31] Thus, in

[29] One can also point to the "model" understanding of the concept of constraints 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, 2021b, 26).

[30] An example is molecular motors that convert free energy, in the form of ATP, into the exertion of force either on objects external to the cell or other components of the cell (cf. Houdusse & Sweeney, 2016).

[31] Generally speaking, it can be said that Bayesian variational methods approximate impossible integrals that occur in Bayesian inference and machine learning. Primarily, these methods serve one of two purposes: to approximate later distribution or to reduce the marginal likelihood of the observed data. This method is derived

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.

It can of course be argued that the model approximates Bayesian inference by continuous iteration updating the priors with respect to formed posteriors, which turn into priors and then a new posterior is formed in light of consecutive prediction error. However, such an approach does not reveal the fully active aspect of the process of minimizing the prediction error, which underlies the PP, and it distinguishes the models built on it from "inactive" hierarchical networks present e.g. in predictive coding (cf. Rao & Ballard, 1999).

I will now turn to the consequences of adopting free energy, as they are crucial for further analysis. At the moment, it must be stated that the normative theory gives credibility to Bayesian PP insofar as it shows predictive mechanisms based on 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). In other words: the most likely internal states, given internal model' states, parameterize a family of (normal) densities for the external states (Millidge, Seth & Buckley, 2021, 9; Parr, Da Costa & Friston, 2019, 4- 5). With this assumption in place the free energy can be viewed as the sum of the long-term average prediction error, providing the link to FEP. Heuristically speaking: in PP, by minimizing the long-term average prediction error, we postulate and optimize the statistics of an approximate variational density, which we then try to match to the desired inference distribution (Millidge, Seth & Buckley, 2021, 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. Such an approach, however, may be prone to the objection of redescription or equivocation (cf. Williams, 2021; Colombo & Palacios, 2021; cf. §9). It seems that a full answer to questions about the possibility of including flows of free energy into predictive mechanisms assumes a normative framework for PP as set by the FEP.

---

from the mathematical perturbation theory which helps find an approximate solution to a problem, by starting from the exact solution of a related, simpler problem. Originally it was used to solve intractable mathematical problems in celestial mechanics (e.g. concerning the movements of the planets).

FEP was originally introduced by Karl Friston (2009; 2010; 2012; Friston, Kilner & Harrison, 2006; Friston & Stephan, 2007; cf. Andrews, 2021) as a principle which explains how sensory cortex infers the causes of its inputs and learns causal regularities. Later, its validity was extended, inter alia, on the perception, action and organization of biological systems (from unicellular cells to social networks (Friston, 2009, 293)). According to this principle any self-organizing system that is at a nonequilibrium steady-state (NESS) with its environment must minimize its free energy.[32] 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. 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 that evolves, which means that it can be said to change over time, subject to random fluctuations. It has to be add that any self-organizing system that is at NESS, i.e. one that has an attracting set, can be described in terms of the Markov blankets (Friston, 2013; Friston, Wiese & Hobson, 2020; Wiese & Friston, 2021).[33]

NESS density means a certain probability of finding it in a particular state when the system is observed at random (Friston, Wiese & Hobson, 2020, 4). In this sense, everything that exists is characterized by properties that remain unchanged or stable enough to be measured over time. This is an essential characteristic of self-organizing systems. NESS density follows directly from the Fokker-Planck formulation of density dynamics (cf. Friston, 2019, 10-11).[34] The solution of the Fokker-Planck equation assumes that the average flow of system states has two parts: the first gradient involves surprisal gradients and the second

[32] 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.

[33] 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 the 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 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).
It should be emphasized that the 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). Without going into details, I emphasize that in these analyzes I refer to Markov blankets in Fristonian manner.

[34] It is a second-order partial differential equation derived from statistical mechanics that describes the time evolution of the probability density functions of position and velocity.

circulates on iso-probability contours (Wiese, Friston & Hobson, 2020, 5). The first counters the dispersion due to random fluctuations, such that the probability density does not change over time. Because of this, states of any system (which must converge to some attracting set, known as a pullback or random attractor[35] (Crauel & Flandoli, 1994)) must conform to a gradient flow on surprisal, i.e. the negative logarithm of the probability density at NESS (Friston & Ao, 2012; Friston, 2013). The solution of the Fokker-Planck equation shows that everything that exists, i.e. everything that can be measured, must possess the above gradient flows. 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. In other words: 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, that is, it must maximize a particular model evidence or equivalently minimise surprisal (cf. Hohwy, 2016). This approach allows us to interpret the flow of (expected) autonomous states of model as a gradient flow on something what we know as VFE,[36] 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, 2; cf. Friston, Wiese & Hobson, 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, *resp*. does not have a Markov blanket, cannot exist.[37]

8. From free energy principle to constrained predictive mechanisms

On the basis of the conducted analyses, it should 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

[35] It can be thought of either (1) as a trajectory of systemic states evolving over time, which revisit certain regions of the system's state space over and over again; or (2) as subtending a probability density over systemic states sampled at random times (Wiese & Friston, 2021, 3).

[36] 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, Wiese & Hobson, 2020, 9-11). The analysis of this issue, however, goes beyond the scope of this paper.

[37] It should be emphasized that not all existing self-organizing systems are alive. The FEP also applies to such systems, *resp*. non-biological agents, which have a certain degree of independence from the environment (Wiese & Friston, 2021, 3).

thermodynamics to NESS (Parr, Da Costa & Friston, 2019). In that sense, the FEP is true for any bounded stationary system that is far from equilibrium. This corresponds to the concept of living organisms defended by mechanists as autonomous dissipative structures (Bickhard's "recursive self-maintaining systems"), i.e. those "that [...] actually use the second law of thermodynamics to their advantage to maintain their organization" (Winning & Bechtel, 2018, 3; cf. Friston & Stephan, 2007; Kirchhoff et al. 2018; Ueltzhöffer, 2019). Broadly speaking: the findings on FEP and NESS mathematics, according to which if something exists then it must exhibit properties that may look as if it is optimizing a VFE, coincide with the heuristics of constrained 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 should be stated that the mechanistic PP *should* take into account the energetic dimension of predictive mechanisms.[38] 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, it can reasonably be considered that the FEP-based PP meets the requirements of the constrained mechanisms approach and allows us to think of predictive mechanisms as constitutive control mechanisms for autonomous systems armed with a generative model.

At this point, however, a legitimate doubt arises: is the Fristonian VFE the free energy that mechanists such as Bechtel write about? Bechtel himself excludes such a link: "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" (Bechtel, 2019, 634; cf. Bich & Bechtel, 2021, 52). This approach seems to exclude the idea of using VFE as a constraint for a mechanistic PP at least in the sense that Bechtel and colleagues propose. It is doubtful, however, whether Bechtel rightly excludes the 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, 127) and as such it is distinguished from the thermodynamic free energy referred to by Bechtel (cf. Moreno & Mossio, 2014). However, in newer approaches, Friston and colleagues show, based on the mathematical relationships between non-equilibrium dynamics, variational inference and stochastic thermodynamics, that

[38] It is worth adding here that the determination of a Markov blanket, e.g. for the brain, may give a formal basis for partitioning the brain into functional units and then to specific control mechanisms (cf. Hipólito et al. 2021). These remarks are important for the possible "use" of Markov blankets in a mechanistic decomposition (according to epistemic interpretation of mechanisms (Bechtel, 2008)).

VFE is the same as thermodynamic free energy: "that any interesting ensemble (that has measurable characteristics) must have a random dynamical attractor to which it converges. On this view, the thermodynamic free energy decreases as the ensemble approaches nonequilibrium steady-state (i.e., its random dynamical attractor)". Heuristically „is consistent with the notion of free energy as the thermodynamic energy available to do work when an ensemble is far from equilibrium" (Frison, 2019, 66-67; Parr, Da Costa & Friston, 2019)[39]. If it is so that the free energy flows constitutive of the active mechanisms can be described in terms 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.[40]

The approach defended here reveals a new dimension of FEP's normativity. According to the approach defended by Friston, FEP is a (normative) state theory i.e. a normative principle that things may or may not conform to, and PP is a process theory – a hypothesis on how that principle is realized (Friston, Fortier & Friedman, 2018, 21).[41] The proposed mechanistic integration of PP with FEP reveals that FEP is a normative theory for PP in the sense that it sets a norm that *should* be met by mechanistically non-trivial PP models, assuming the implementation of the constrained mechanisms approach and its heuristics. According to this norm, PP models *should* have an energetic component if they are to be mechanistic. This approach can be treated as a voice in the discussion on the status of PP and its relation to the FEP. In this approach, FEP not only constraints the space of possible algorithms for PP (Spratling, 2017), but also indicates energetic constraints for the causal organization of all autonomous systems, including those that are armed with 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)

[39] It should be emphasized that 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.

[40] This approach can be accused of the "consistency fallacy" (cf. Mole & Klein, 2010; Litwin & Miłkowski, 2020), which concerns that data that are consistent with a given theory cannot be offered, because of this consistency, as evidence in support of that theory. Something more is needed. According to this, the claim that the notion of free energy is consistent with the notion of thermodynamic energy is not sufficient to conclude that the notion of VFE can be considered as constraint in Bechtel's meaning. The objection can be answered that the said this "consistency" is not a consistency between different "data" (which is an example of consistency fallacy described in the literature), but a consistency in a formal sense, i.e. based on mathematics. Therefore, it is difficult to treat both terms of free energy as data supporting any theory.

[41] In sense that „The free energy minimising 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, Badcock & Friston, 2017, 6).

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.[42]

9. Discussion

The presented analyses are only a rough framework of how to integrate PP with FEP using constraints and free energy flows. If the approach proposed here is valid, it has certain consequences for a number of discussions among PP and FEP researchers. First of all, it allows for a new way to approaching the PP-FEP relationship. If FEP refers to self-organizing adaptive systems, as described in the dynamical system theory (DST), 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, 1459).[43] An example of this type of practice can be found, among others in Badcock et al.: "mechanisms involve a dynamic, bidirectional relationship between specialised 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)" (Badcock, Friston & Ramstead, 2019, 105).[44] In this approach, FEP provides specific constraints for a PP's scheme of mechanism. Therefore, it is a stronger commitment than that suggested by Gładziejewski, stating that "FEP only puts extremely general constraints on the causal organization of organisms, perhaps to the point of lacking any non-trivial commitments about it" (Gładziejewski, 2019, 664) or Harkness: "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, 2). I also argue that the normative dimension of the FEP may

[42] According to Clark: „our basic evolved structure (gross neuroanatomy, bodily morphology, etc.) may itself be regarded as a particularly concrete set of inbuilt (embodied) biases that form part of our overall 'model' of the world" (Clark, 2016, 175).

[43] „Cognitive modelers begin by designing a computational model that hypothesizes a mechanism with specified parts, operations, and organization, and then try to show that such a model can mimic the cognitive phenomena of interest. If successful, they offer it as a model of the actual mechanism. Some researchers go beyond basic mechanistic modeling to include dynamic phenomena among the explanatory targets" (Bechtel & Abrahamsen, 2010, 332).

[44] In this sense, FEP could be considered "formally expressible and neurobiologically plausible physics of the mind" (Badcock, Friston & Ramstead, 2019, 109).

actually be explanatory, and not just "a regulatory principle" "guiding" or "informing" the construction of process theories, as Hohwy (2020b, 39) suggests.[45]

My considerations also shed some light on a number of critical works concerning either the FEP itself or its relationship with the PP. There is a place here to mention only the most important recent papers. I will limit myself to the two recently published papers, emphasizing certain issues. In the paper *Non-equilibrium thermodynamics and the free energy principle in biology,* Colombo and Palacios 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, 2). 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 (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 the DST, as sketches of mechanisms.[46] The key to this position is the identification of Fristonian free energy with thermodynamic free energy (cf. §8). From this perspective, FEP is a realization of the heuristics of constrained mechanisms, which dictates that constitutive mechanisms for autonomous systems, constraints and free energy should be included in the explanation.

Colombo and Wright (2021) take into account that the analyses carried out by FEP supporters can be treated as sketches of mechanisms in the sense of Piccinini and Craver (2011), however, they treat them only as weak explanatory idealizations. It seems, however, that in the light of the constrained mechanisms approach, sketches of free energy flow mechanisms can be used in the formulation of schemes of mechanisms with specific explanatory powers. This paper suggests such an approach that uses PP's sketches of mechanisms and emphases their explanatory power (cf. Gładziejewski, 2019).[47]

In his paper *Is the brain an organ for free energy minimization?* Daniel Williams claims that "The FEP can be interpreted in two ways: as a claim about how it is possible to redescribe the existence of self-organising systems (the Descriptive FEP), and as a claim

[45] The full justification of this claim, however, goes beyond the analyzes presented here.
[46] It should also be added that FEP as an extension of DST (Aguilera et al. 2021) is descriptive and also explanatory because dynamical models can be treated as mechanistic descriptions (cf. Zednik, 2008).
[47] This approach allows to defend the view that the FEP complies with the mechanistic approach, which to some extent is contradicted by e.g. Beni (2018).

about how such systems maintain their existence (the Explanatory FEP). Although the Descriptive FEP plausibly does identify a condition of the possibility of existence for self-organising systems, it has no important implications for our understanding of how the brain works" (Williams, 2021, 2). I argue that the arguments I have presented allow us to conclude that what he describes as "the Descriptive FEP" and what I believe to be a description of the dynamics of autonomous systems expressed in the language of the DST has important consequences for the explanation of how the brain works because it imposes energy constraints on the system, which are key to explaining how the brain processes information (cf. Laughlin et al. 1998). In practice, this means that DST descriptions are mechanistic. If so, then the distinction proposed by Williams into "the Explanatory FEP" and "the Descriptive FEP" is unjustified, because "the Explanatory FEP" *de facto* implies "the Descriptive FEP".

The analyzes presented here also allow us to refer to the belief shared by Williams (2021) and Colombo with Wright (2021) that FEP is based on a transcendental argument: "imperative to minimise free energy can be derived from transcendental reflection on conditions of the possibility of existence for self-organizing systems "(Williams, 2021, 3). According to this analyzes, it is difficult to agree with this view, because (1) the FEP defines conditions that are sufficient but not necessary for the existence of for self-organizing systems (e.g. due to a number of additional assumptions about the nature of life and adaptation); (2) FEP allows you to think of self-organizing systems *as if* they minimize VFE in the sense that it indicates one (but not necessarily the only) constraint to their explanation; (3) FEP as an extension of DST is primarily a formal tool for describing the dynamics of self-organizing systems and does not offer any metaphysical theses and in this sense is "almost completely free of empirical content" (Williams, 2021, 10). It is therefore difficult to agree with the statement that "the Descriptive FEP is the only interpretation of the FEP licensed by the transcendental argument" (Williams, 2021, 13), because "the Descriptive FEP" is *only* justified by specific Bayesian research practice and DST language. In this sense, the very transcendental argument referred to by Williams, Colombo and Wright can (*only*) be treated as a form of abductive inference and FEP as one of the possible and permissible (in this sense sufficient but not necessary) models explaining life.[48]

[48] It should be said that an exhaustive discussion with these approaches goes far beyond these analyses and requires in-depth research and the formulation of exhaustive arguments. Here I am only emphasizing some issues.

Conclusion

In the present paper I wondered to what extent the postulate of some mechanists about the need to include in the explanations such constitutive components as constraints for mechanisms and free energy flows can be applied to a mechanistically interpreted PP framework. I found that the position defined by me as the constrained mechanisms approach is justified in relation to PP, because the actual research practice in this framework corresponds to the heuristics of constrained mechanisms and is related to those approaches that assume the FEP as a normative framework for the process theory realized by PP. According to the approach I defend, non-trivial PP models should include an energetic component, if they are to be mechanistic. The discussion presented here may be of great importance for considering the relationship between PP, FEP and active inference. It 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. This approach additionally reveals the groundlessness of many critics of the FEP and the FEP-based PP.

References

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*, 4, (pp. 1–14). Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573055.

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

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. *Adapt. Behav*. 14, 171-185. https://doi.org/10.1177/105971230601400208.

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

Bechtel, W. (2018). The Importance of Constraints and Control in Biological Mechanisms: Insights from Cancer Research. *Philosophy of Science*, 85(4), 573-593.

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. (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) <http://philsci-archive.pitt.edu/view/confandvol/confandvol2020PSA.html>.

Bechtel, W. & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, 421–441.

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 Part A*, 41(3), 321-333.

Bechtel, W. & Bich, L. (2021) Grounding cognition: heterarchical control mechanisms in biology. *Phil. Trans. R. Soc. 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*, 1-25. https://doi.org/10.1007/s11229-021-03350-x.

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.

Beni M. D. (2019) Conjuring Cognitive Structures: Towards a Unified Model of Cognition. In: Nepomuceno-Fernández Á., Magnani L., Salguero-Lamillar F., Barés-Gómez C., Fontaine M. (eds) *Model-Based Reasoning in Science and Technology. MBR 2018. Studies in Applied Philosophy, Epistemology and Rational Ethics*, vol 49, (pp. 153-172). Springer, Cham. https://doi.org/10.1007/978-3-030-32722-4_10.

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

Bickhard, M. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence*, 5, 285–333.

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. https://doi.org/10.1007/978-94-007-1044-3_6.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, 76, 198–211. https://doi.org/10.1016/j.jmp.2015.11.003.

Bruineberg, J., Dolega, K., Dewhurst, J. & Baltieri, M. (2021) The Emperor's New Markov Blankets. Preprint.

Bruineberg, J., Kiverstein, J. & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 195, 2417-2444. https://doi.org/10.1007/s11229-016-1239-1.

Campbell, D. T. (1990). Levels of Organization, Downward Causation, and the Selection-Theory Approach to Evolutionary Epistemology. In Greenberg, G., & Tobach, E. (Eds.) *Theories of the Evolution of Knowing*. (pp. 1-17). Hillsdale, NJ: Erlbaum.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *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: Oxford University Press.

Colombo, M. & Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biol Philos*, 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.

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

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

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. *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 Ecol Evol*, 31(8), 622-632. https://doi.org/10.1016/j.tree.2016.04.009.

Cummins, R. (1975). Functional Analysis. *Journal of Philosophy*, 72, 741–764.

Darden, L. (2006). *Reasoning in Biological Discoveries*. Cambridge: Cambridge University Press.

Darden, L. (2008). Thinking Again about Biological Mechanisms. *Philosophy of Science*, 75, 958–969.

Davies, P. C. W. (2019). *The Demon in the Machine: How Hidden Webs of Information Are Solving the Mystery of Life*. Chicago: The University of Chicago Press.

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

Feynman, R. P. (1998). *Statistical Mechanics: A Set Of Lectures*. Avalon Publishing.

Fodor, J. A. (1968). *Psychological Explanation*. New York: Random House.

Franklin, D.W. & Wolpert, D. M. (2011). Computational mechanisms of sensorimotor control. *Neuron*, 72(3), 425-42. doi: 10.1016/j.neuron.2011.10.006.

Friston, K. J., Kilner, J. & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology*, 100(1–3), 70–87.

Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13(7), 293–301.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Friston, K. J. (2011). What is optimal about motor control? *Neuron*, 72(3), 488-498.

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. & Ao, P. (2012). Free energy, value, and attractors. *Comput. Math. Methods Med*. 2012, 937860.

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

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., Mattout, J., Trujillo-Barreto, N., Ashburner, J. & Penny, W. (2007). Variational free energy and the Laplace approximation. *Neuroimage*, 34, 220–234.

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

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.

Gigerenzer, G., Hoffrage, U. & Goldstein, D. G. (2008). Fast and frugal heuristics are plausible models of cognition: Reply to Dougherty, Franco-Watkins, and Thomas. *Psychological Review*, 115(1), 230–239

Glennan, S. & Illari, P. (eds.) (2018). *The Rutledge handbook of mechanisms and mechanical philosophy*. London and New York: Routledge.

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

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 Biol*, 17(4): e3000233. https://doi.org/10.1371/journal.pbio.3000233.

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). Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570153.

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). Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573178.

Hasselmo, M. E. (2012). *How we remember: Brain mechanisms of episodic memory*. Cambridge, MA: MIT Press.

Hempel, C. G. & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, 15, 135–175.

Hipólito, I., Ramstead, M. J. D., Convertino, L., Bhat, A., Friston, K. J. & Parr, T. (2021). Markov Blankets in the Brain. *Neuroscience & Biobehavioral Reviews*, 125, 88-97. https://doi.org/10.1016/j.neubiorev.2021.02.003.

Hohwy, J. (2013). *The predictive mind*. Oxford: 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). Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570016.

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, 50(2), 259–285.

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.

Houdusse, A., & Sweeney, H. L. (2016). How myosin generates force on actin filaments. *Trends in Biochemical Sciences*, 41(12), 989–997. https://doi.org/10.1016/j.tibs.2016.09.006.

Illari, P. M. & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, 2, 119–135. https://doi.org/10.1007/s13194-011-0038-2.

Kaan, E., & Swaab, T. Y. (2002). The brain circuitry of syntactic comprehension. *Trends in Cognitive Science*, 6(8), 350–356.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, 183, 339–373. https://doi.org/10.1007/s11229-011-9970-0.

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 of Science*, 78, 601–627.

Keller, G. B. & Mrsci-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, 2(100), 24, 424-435 https://doi.org/10.1016/j.neuron.2018.10.003.

Kety, S. S. (1963). The Circulation and Energy Metabolism of the Brain. *Neurosurgery*, 1(9), 56–66. https://doi.org/10.1093/neurosurgery/9.CN_suppl_1.56.

Kiefer, A. & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese*, 195, 2387–2415. https://doi.org/10.1007/s11229-017-1435-7.

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.

Laughlin, S., de Ruyter van Steveninck, R. & Anderson, J. (1998). The metabolic cost of neural information. *Nat Neurosci*, 1, 36–41. https://doi.org/10.1038/236.

Litwin, P. & Miłkowski, M. (2020). Unification by fiat: arrested development of predictive processing. *Cognitive Science*, 7(44), 12867. https://doi.org/10.1111/cogs.12867.

Machamer, P. (2004). Activities and Causation: The Metaphysics and Epistemology of Mechanisms. *International Studies in the Philosophy of Science*, 18, 27–39.

Machamer, P. K., Darden, L. & Craver, C. F. (2000). Thinking about Mechanisms, *Philosophy of Science*, 57, 1–25.

Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman & Co.

McCulloch, W. S. (1945). A heterarchy of values determined by the topology of nervous nets. *Bull. Math. Biophys*, 7, 89–93. https://doi.org/10.1007/BF02478457.

Millidge, B. (2019). Implementing Predictive Processing and Active Inference: Preliminary Steps and Results. PsyArXiv. https://doi.org/10.31234/osf.io/4hb58.

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.

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.

Mole, C., & Klein, C. (2010). Confirmation, refutation, and the evidence of fMRI. In S. J. Hanson & M. Bunzl (Eds.), *Foundational issues in human brain mapping*, (pp. 99–112). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262014021.003.0010.

Moreno, A., & Mossio, M. (2014). *Biological autonomy: A philosophical and theoretical inquiry*. Dordrecht, the Netherlands: Springer.

Mumford, S. (2004). *Laws in Nature*, London: Routledge.

Nowakowski, P. (2017). Bodily Processing: The Role of Morphological Computation. *Entropy*, 19, 295; doi:10.3390/e19070295.

Oaksford, M. & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford: Oxford University Press.

Parr, T., Da Costa, L., & Friston, K. J. (2020). Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A*, 378(2164), 20190159.

Parr, T., Markovic, D., Kiebel, S., & Friston, K. J. (2019). Neuronal message passing using Mean-field, Bethe, and Marginal approximations. Scientific Reports, 9, 1889.

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: 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*. San Francisco: Morgan Kaufmann Publishers.

Piccinini, G. & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183(3), 283–311.

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

Raja, V., Anderson, M. L. (2021). Behavior Considered as an Enabling Constraint. In F. Calzavarini & M. Viola (Eds.), *Neural Mechanisms, Studies in Brain and Mind 17*, (pp. 209-232). https://doi.org/10.1007/978-3-030-54092-0_10

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.

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.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420.

Rosen, R. (1991). *Life itself: A comprehensive inquiry into the nature, origin and fabrication of life*. New York: Columbia University Press.

Ruiz-Mirazo, K., & Moreno, A. (2004). Basic autonomy as a fundamental step in the synthesis of life. *Artificial Life*, 10, 235–259.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

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). Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570108.

Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of Science*, 77(4), 477–500.

Silberstein, M. & Chemero, A. (2013). Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences. *Philosophy of Science*, 5(80), 958-970. https://doi.org/10.1086/674533.

Simon, H. (1969). *The Sciences of the Artificial*. Cambridge, MA:MIT Press.

Smith, R., Friston, K. J. & Whyte, C. (2021). A Step-by-step Tutorial on Active Inference and Its Application to Empirical Data. PsyArXiv. January 2. https://doi.org/10.31234/osf.io/b4jm6.

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.

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.

Sprevak, M. (2021a). Predictive coding I: Introduction. [Preprint]. http://philsci-archive.pitt.edu/id/eprint/19365.

Sprevak, M. (2021b) Predictive coding IV: The implementation level. [Preprint]. http://philsci-archive.pitt.edu/id/eprint/19669.

Ueltzhöffer, K. (2019). https://kaiu.me/2019/10/09/life-and-the-second-law/ (access 27.11.2021)

Umerez, J. & Mossio, M. (2013). Constraint. In W. Dubitzky, O. Wolkenhauer, K. H. Cho & H. Yokota (Eds.), *Encyclopedia of systems biology*, (pp. 490–493). Berlin: Springer. https:// doi.org/10.1007/978-1-4419-9863-7.

Weiskopf, D. A. (2016). Integrative modeling and the role of neural constraints. *Philosophy of 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.

Wiese, W. & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. W: T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*, 1, (pp. 1–18). Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573024.

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? *Philos Stud*, 1-22. https://doi.org/10.1007/s11098-021-01722-0.

Wimsatt, W. (1974). Complexity and Organization. In K. F. Schaffner and R. S. Cohen (Eds.), *Boston Studies in the Philosophy of Science*, vol. 2, (pp. 67-86). Dordrecht, Holland: Reidel.

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.

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21, 615– 628.

Varela, F. (1979). *Principles of biological autonomy*. New York: Elsevier.

Venter E (2021) Toward an Embodied, Embedded Predictive Processing Account. *Front. Psychol*. 12:543076. https://doi.org/10.3389/fpsyg.2021.543076,

Zednik, C. (2008). Dynamical models and mechanistic explanations. In B. C. Love, K. McRae, V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, (pp. 1454–1459). Austin: Cognitive Science Society.