Mechanistic unity of the predictive mind

Paweł Gładziejewski
pawel.gla@umk.pl

Department of Cognitive Science
Nicolaus Copernicus University
ul. Fosa Staromiejska 1a, 87-100, Toruń, Poland

Abstract: It is often recognized that cognitive science employs a diverse explanatory toolkit. It has also been argued that cognitive scientists should embrace this explanatory diversity rather than pursue search for some grand unificatory framework or theory. This pluralist stance dovetails with the mechanistic view of cognitive-scientific explanation. However, one recently proposed theory – based on an idea that the brain is a predictive engine – opposes the spirit of pluralism by unapologetically wearing unificatory ambitions on its sleeves. In this paper, my aim is to investigate those pretentions to elucidate what sort of unification is on offer. I challenge the idea that explanatory unification of cognitive science follows from the Free Energy Principle. I claim that if the predictive story is to provide an explanatory unification, it is rather by proposing that many distinct cognitive mechanisms fall under the same functional schema that pertains to prediction error minimization. Seen this way, the brain is not simply a predictive mechanism – it is a collection of predictive mechanisms. I also pursue a more general aim of investigating the value of unificatory power for mechanistic explanations. I argue that even though unification is not an absolute evaluative criterion for mechanistic explanations, it may play an epistemic role in evaluating the credibility of an explanation relative to its direct competitors.

Key-words: explanatory pluralism; explanatory unification; free energy principle; mechanistic explanation; predictive processing

1. Introduction

The strive for theoretical and explanatory unity is no longer universally regarded as a fundamental normative principle of science (Braitenbach, Choi 2017; Cartwright 1999; Dupré 1993). This at least applies to the special sciences, given that physicists have not yet forgone the search for a grand unifying theory. Cognitive science (including cognitive neuroscience) is no different in that regard. One thing to note is that, regardless of normative issues, as a matter of *fact*, there is no single overarching theory or framework under which explanations of distinct cognitive phenomena could be subsumed. Differing theories and models, based on different assumptions and concepts co-exist, corresponding to distinct research goals and domains. Some phenomena are explained representationally, while other explanations do completely without invoking semantically evaluable states; some phenomena are explained using dynamical systems theory, while others are modelled using a more old-fashioned symbolic-computational approach; some models abstract away from neuroscientific detail, while others are largely physiological, etc. Crucially, not that many cognitive scientists or philosophers still *worry* over the fragmentation of cognitive science. In fact, it has been argued that explanatory pluralism is not simply an inconvenient feature of cognitive science at this stage of inquiry, but also that the disunity is in some sense desirable (Dale 2008; Dale, Dietrich, Chemero 2009). From this perspective, the search for a theory or framework to which this diversity could be reduced looks misguided and, perhaps, futile. Plurality, not unity is the natural order of things.

The spirit of explanatory pluralism dovetails with the idea that cognitive-scientific explanations are predominantly mechanistic (Bechtel 2008; Bechtel, Richardson 1993; Craver 2007; Kaplan 2011; Piccinini, Craver 2011).[1] To explain a phenomenon mechanistically is to

[1] A clarification is in order here. Two sorts of explanatory pluralism should be distinguished: (1) pluralism regarding the explanations themselves, (2) pluralism regarding the types of explanatory strategies at use. In this paper, I appeal to (1). In the context of mechanism, this sort of pluralism means that different phenomena are explained by appeal to distinct mechanisms, which differ in terms of their functional and structural organization. Pluralism in the sense (2) pertains to the issue of whether cognitive science makes use of distinct *types* of explanation, e.g. to whether it makes use of non-mechanistic explanations (see e.g. Chirimuuta 2017; Weiskopf 2011). In the present paper, I leave this problem aside and focus my discussion on mechanistic explanations. I am indebted to an anonymous reviewer for urging me to draw this distinction.

describe an organized set of component parts and their activities that are jointly responsible for the phenomenon. Crucially for the present purposes, the quality of a mechanistic explanation can be disentangled from its unificatory potential. That is, its value is not necessarily dependent on how well it unifies distinct phenomena or on whether the explanation itself falls under some unifying principle or law. Rather, the centrally important norm that dictates the quality of a mechanistic explanation is how well it maps onto actual causal structure of a mechanism responsible for the explanandum phenomenon (Kaplan 2011). Sometimes this capacity to *track the relevant causal structure* may come at an expense of how well an explanation *unifies* phenomena. If the brain is composed of many highly heterogenous mechanisms, then this fact will be mirrored in the heterogenous, perhaps to the point of being 'monstrous', nature of the scientific models of those mechanisms (see Miłkowski 2016). These models may differ substantially, with each of them having an explanatory scope limited to a particular phenomenon. Disunity would not count against those models in such a scenario; accuracy in describing mechanisms is more important. This way, the assumption that cognitive science explains by describing mechanisms justifies dropping the search for unity as an ideal of cognitive-scientific inquiry.[2]

However, there is at least one ambitious theoretical proposal on the market which vehemently contradicts the pluralist outlook. This theory states that the brain is a prediction engine of a specific sort. Part of the attraction of this proposal is supposed to stem precisely from its unificatory power, as it is sometimes introduced as a potential 'grand unifying theory' of the brain and cognition (Friston 2009, 2010). This story is rooted in a theory of what life is, namely on the Free Energy Principle (FEP). Roughly, according to the FEP, living systems are things that maintain their own existence by minimizing the free energy of their sensory states. From the FEP's standpoint, an organism is treated as a model of the causal structure of its environment, a model which maximizes evidence of its own existence, thereby avoiding thermodynamic dispersal. To achieve this, an organism engages in actions that minimize the prediction error, i.e. the

[2] Instead of unification, mechanists tend to talk about *integrating* different strands of research by showing how they are aimed at discovering the same mechanism, or how they uncover interactions between distinct mechanisms. (See Craver 2007; see also Miłkowski 2016 for the distinction between unification and integration).

discrepancy between its expectations about its sensory states and actual states of its sensory apparatus. By way of conceptual necessity (to live *is* to minimize free energy), FEP applies to all living systems, and this large scope is where a major part of the unificatory power of the theory supposedly lies. In addition, FEP inspires a particular set of claims about cognitive architecture, usually dubbed in literature as ‘predictive processing’ (PP; see Clark 2013, 2016; Hohwy 2013, 2018; Wiese, Metzinger 2017). Again, roughly, in PP the brain is construed as storing a hierarchical generative model of the environment which sends a cascade of top-down sensory predictions to minimize the bottom-up prediction error signal, where the error signal is precision-weighted according its predicted reliability. This single computational scheme is supposed to explain perception, action and attention. Furthermore, there are attempts to use PP as a basis for explanations of more fine-grained cognitive phenomena, like aspects of social cognition, binocular rivalry, the formation of psychotic states, pain perception, conscious sense of presence, religious experience, the inability to tickle oneself, the perception of time, or even decision making while driving a car (see e.g. Brown, Adams et al. 2013; Geuter, Boll et al. 2017; Engström, Bärgman et al. 2017; Hohwy, Paton, Palmer 2015; Hohwy, Roepstorff, Friston 2008; Quadt 2017; Seth 2014; Sterzer, Adams et al. 2018; van Elk, Aleman 2017). From a pluralist standpoint, this unificatory ambition may seem preposterous, or at least suspicious.

It is not the aim of this paper to evaluate whether the predictive mind view succeeds at unifying cognitive science (the proponents of the predictive view are understandably optimistic about this, but some authors are less convinced, see e.g. Colombo, Wright 2017). Instead, the point is to elucidate what it would even *mean* to unify cognitive science with a theory of this sort. I will argue that the unifying power of the theory does not come from FEP, as it is doubtful whether and how the principle could provide a properly explanatory unification for cognitive science. Rather, if the predictive story was to unify cognitive science, it would be by providing (in the form of PP) a functional sketch of a mechanism that turns out to *recur* throughout the brain (Danks 2014). That is, although the brain is composed of many distinct mechanisms, these mechanisms may be unified by the fact that they fall under a common blueprint in their functional organization. I also have another, more general goal, as I aim to use the PP as an instructive case study of where the value of a mechanistic explanatory unification of cognitive science may reside.

After all, the question may still be raised about whether anything is to be gained, in terms of explanatory quality, from unification. Although I agree with mechanist's denial that unification is an *absolute* evaluative norm for a mechanistic explanation, I claim that it may still serve as a *relative* norm in evaluating competing models of cognitive mechanisms.

The discussion to come is structured as follows. In section 2, I take a closer look the FEP, distinguishing it from PP and evaluating its explanatory and unificatory potential for cognitive science. In section 3, I focus on PP to put forward a different take on how it provides explanatory unification for cognitive science. In section 4, I generalize my discussion to consider the value of unification in evaluating mechanistic explanations. I close the paper with a succinct summary.

2. Free Energy Principle and the predictive mind's unificatory ambitions

In this paper, I take the 'predictive mind' view to be a combination of two ideas. On the one hand, the view is deeply rooted in the free energy principle (FEP), which is an abstract conception in theoretical biology that aims to rigorously capture what it takes to be a living agent (Friston 2012, 2013). On the other hand, the predictive view encompasses predictive processing (PP), which is a set of claims about cognitive architecture which are usually associated with FEP (see Clark 2013, 2016; Hohwy 2013, 2016; Wiese, Metzinger 2017). Given that my focus is on unifying cognitive science, I am mostly concerned with the properly cognitive part of the story, which is PP. However, PP cannot be neatly separated from FEP. In fact, the unificatory ambition of the former seems tightly connected to unificatory ambition of the latter. The 'grand unifying theory' dialectic that accompanies discussions of PP is often taken to be justified by how PP fits into a larger overarching context of FEP. I want to investigate this connection, as it is not entirely clear what sort of explanatory unification FEP is supposed to deliver and how it applies to unifying cognitive science.

FEP originates from the claim that to live is to keep oneself in far-from-thermodynamic-equilibrium steady state. According to FEP, this can be fruitfully captured in statistical terms (see Friston 2009, 2010, 2012, 2013). Each phenotype is said to 'define' or 'entail' a probability distribution over its possible internal states. After all, so long as it exists, an organism is far more likely to be in one the states that lie within the rage of states which sustain its thermodynamic viability than in a state outside of this range. Furthermore, because the internal states are dependent on the states of the external environment, an organism 'implicitly' encodes a generative model that specifies how internal states are probabilistically conditional on external states. To live, then, a system must maximize, through action ('active inference'), the evidence (posterior probability) of the model that it embodies; that is, it must act to avoid states that are surprising (i.e. are associated with large negative log probability), given this model. However, solving this problem directly is equivalent to performing optimal Bayesian inference and is computationally intractable. Another difficulty is that the organism has no 'God's point of view' from which it could directly access the true posterior distribution. According to FEP, these issues can be averted. Instead of computing posterior probabilities directly, the uses a tractable variational Bayesian method to arrive at a recognition distribution which, with adjustments made over time, starts to approximate the true distribution. The point is that the organism incrementally optimizes the recognition distribution (i.e. brings it closer to the true distribution) by minimizing the information-theoretic free energy of its own sensory states. This is possible because the free-energy of the sensory states defines an upper bound on the value of surprise; so, minimizing the former value equals minimizing the latter. Furthermore, free energy is treated as equivalent to long-term prediction error of the sensory states, i.e. the discrepancy between expectations, implicitly 'encoded' in the organism's phenotype, and the sensory feedback acquired through sampling the environment. Hence, to live a system must, over long periods of time, avoid unexpected sensory states.

To see how this general outlook promises to provide unification for *cognitive science*, two further considerations must be added. One, proponents of FEP take this principle to 'entail' PP, i.e. a set of claims about the information-processing mechanisms in the brain (Friston 2009, 2010). Neural ensembles are supposed to implement variational Bayes and the ultimate function of the brain is supposed to be minimizing the prediction error. I will return to this notion shortly. Second, FEP itself seems to bring heavy unificatory power of its own. One of the hallmarks of unified explanations is that they have large, preferably unbounded scope (Kitcher 1989; Miłkowski 2016). By conceptual necessity, FEP generalizes over all living (self-organizing, adaptive) systems. After all, once we agree on characterizing the organism as encoding a generative model of its environment, it follows that minimizing long-term prediction error of sensory states is a necessary condition for being a living system. Because FEP is both unificatory and has, according to its proponents, commitments about how cognition is realized in the brain, it is only natural to see it as holding serious unificatory promise for cognitive science.

One major doubt about whether FEP could deliver successful explanatory unification for cognitive science lies in the question about the explanatory status of the principle itself. To provide an explanatory unification for cognitive science, the FEP needs to be in some sense *explanatory*. Furthermore, even if shown to be explanatory, FEP needs to provide us with an explanation of an appropriate *sort*, namely of the sort that cognitive scientists strive for. And on the view employed in the present paper, what cognitive scientists ultimately seek is to latch onto the causal nexus of the world to uncover the *causal* basis of cognition (but see note 1). This can be done by either characterizing the causal-etiological antecedents of cognitive phenomena or by uncovering their constitutive dependency on a lower-level mechanism, comprised of a set of organized, active components of the cognitive system (Craver 2007).

However, FEP does not seem like a causal-etiological or causal-mechanistic explanation at all (see also Colombo, Wright, 2018; Klein 2018). It is an abstract, formally expressed principle that characterizes an imperative rule regarding what an organism necessarily needs to do to persevere. This principle is also descriptive insofar as the behavior of any system that resists structural disintegration can be characterized in terms of maximizing evidence for a generative model that the system in question ‘embodies’. Necessarily, all living systems obey the FEP. This means that the behavior of any such system can be represented as a trajectory in a state-space which is jointly determined by the prior beliefs ‘embodied’ in the system and its current sensory states. But understood this way, FEP stands as an ingenious technical *redescription* of what adaptive or self-organizing behavior is, rather than an *explanation* of it. Of course, the way we describe phenomena guides our explanatory practices, and so unifying phenomena by subsuming them under one description might invite explanatory unification. Still, descriptive unification is not yet explanatory unification (Danks 2014).

To counter this criticism, one might note that FEP allows one not only to describe, but also to *predict* how the system's state-space trajectory will evolve over time, and how it would evolve under a range of counterfactual scenarios. This way, FEP could be seen as a basis for covering-law explanations, with the principle itself serving as a biological law or law-like regularity, which allows (given antecedent conditions, i.e. the model and sensory state) predictions about actual and possible behavior. Although this is a promising way of interpreting the explanatory role of FEP, doubts about its usefulness for cognitive science remain. The idea that cognitive phenomena can be properly explained in a nomological manner has been contested (Bechtel 2008; Craver 2007; Cummins 2000; Glennan 2017). It has been argued that law-like regularities act as mere descriptions of phenomena; that covering-law 'explanations' confound prediction with explanation, as it is possible to predict phenomena without knowing their causes or underlying mechanisms; that the covering-law view of explanation does not adequately characterize explanatory practices of cognitive scientists. Arguably, those well-known issues could be raised with regards to FEP. The principle can be said to describe adaptive behavior and allow for its prediction, but only in a highly idealized way that abstracts from the behavior's underlying causes. So, under the covering-law reading, FEP is at least potentially explanatory, just not in the exact sense of providing causal/mechanistic explanations which cognitive scientists are interested in. [3]

But perhaps the discussion so far gets things fundamentally wrong. Perhaps it is a mistake to seek explanatory unification in the FEP itself. Rather, FEP plays a unificatory role only through its relation to PP. While FEP serves as an abstract principle or law, PP provides a sketch of a cognitive mechanism that realizes the free-energy minimization. It is PP that acts as proper explanation of cognitive phenomena in this story. And it is PP where some sort of explanatory unification is to be found. The intuition behind this is that FEP renders PP as something more than yet another empirical hypothesis regarding the nature of cognitive mechanisms. The PP is supposed to not simply turn out, as a matter of fact, to provide a successful explanatory

[3] Another possibility might be that FEP is explanatory in still some other, non-mechanistic (non-causal) sense of 'explanation'. For example, it might be argued that FEP explains in virtue of showing how certain features of mechanisms that realize cognition are necessitated by mathematical facts (see Chirimuuta 2017; Lange 2013). Discussing this possibility is, however, beyond the scope of the present paper.

unification of cognitive phenomena. Rather, its unificatory power is purportedly derived – as a matter of principle, not (just) fact – from its connection to FEP.

Although I think that there is a relatively weak reading of the FEP-PP relation that goes some way to justifying this general intuition (I will turn to this in the next section), a far stronger view is sometimes associated with the FEP literature (see e.g. Colombro, Wright 2018). On this view, FEP *a priori* necessitates the truth of PP. That is, FEP entails facts about cognitive architecture, down to the neural level. As noted by Colombo and Wright (2018, p. 18), FEP theorists sometimes proceed *more geometrico*, by deducing, from axioms and formulae, seemingly contingent facts like the hierarchical organization of cortical layers or the existence of neural adaptation and repetition suppression. Once one adopts this strategy of theorizing, the unificatory status of PP is clear. FEP is, by conceptual necessity, true of any living (adaptive, self-organizing) agent. FEP entails PP as an account of its realizing mechanism. Since FEP applies universally, so does PP.

Several worries about the legitimacy of this move emerge. What immediately invites caution is the suspect epistemic status of the reasoning behind this kind of defense of PP's unificatory role. After all, we are led to believe that relatively fine-grained details of cognitive architecture can be derived *a priori* from FEP. And FEP, as its proponents themselves agree (see e.g. Friston, Thornton, Clark 2012), is ultimately a mathematically refined expression of a tautological-sounding statement that to live is to actively avoid thermodynamic death. It seems like too much is deduced from too little, giving the argument a worryingly 'Hegelian' flavor (Chemero 2011).

Another point is that FEP is too general in scope to provide a proper sort of unification for cognitive science. If FEP entails constraints on the causal organization of free-energy-minimizing systems, these constraints should be taken to apply to all systems that fall under the principle. However, the latter category encompasses single-cell organisms, multicellular organisms that lack a nervous system and cognitively sophisticated animals like octopuses, whose nervous system differs significantly in its organization from, say, a human nervous system. From FEP's vantage point, a *Paramecium* or a sponge minimizes the free energy of its sensory states in the same sense as a chimp. The class of systems that fall under FEP then includes seemingly non-cognitive systems, systems that count as barely or minimally cognitive and full-blown cognitive systems that differ substantially among each other (like cephalopods and primates). It is doubtful that there is a core cognitive mechanism such that *all* these systems fall under the FEP in virtue of being equipped with this mechanism. It seems more likely that what unifies those systems and makes them all fall under the FEP is that they have a dispositional, system-level property of *acting adaptively*. This makes them 'appear as if' they were sampling the environment to find themselves in unsurprising sensory states. *This* fact allows us to construe them all as free-energy-minimizers.

A related point is that even if we allow that all living systems realize the FEP by implementing PP, this comes at a cost of PP being too liberal an account of cognitive mechanisms. Take for example the fact any free-energy-minimizer is described as a generative model 'embodying' or 'encoding' prior beliefs about the causal structure of its environment. As other authors have already noted (Bruinenberg, Kiverstein, Rietveld 2018; Clark 2017), the sense of 'belief' at use here is extremely loose. FEP is liberal about how those priors are realized in the system, as any morphological feature of an organism in virtue of which the organism 'fits' an aspect of its environment can be said as 'encoding' a prior 'belief' about this aspect. Even single-cell organisms count as prior-belief-holders on this rendering. This might mean that the contents thus ascribed to an agent are not observer-independent, semantic properties of the agent's internal states, causally shaping its behavior. Because FEP is so liberal about how prior beliefs are realized, ascriptions of prior beliefs may have merely fictional, 'as if' status (see Downey 2018). Alternatively, the intentional commitments of FEP might be construed realistically, assuming a realist view that is relaxed with respect to commitments about internal mechanisms (see e.g. Dennett 1991, Schwitzgebel 2002). In any case, the point is that intentional ascriptions in FEP are simply meant to capture the adaptive value of agent's features, rather than provide a story about mechanisms underlying the agent's behavior.

Relatedly, consider how FEP parcels any living system (see e.g. Friston 2009, 2012, 2013) into internal states, sensory states, active states (which determine the system's actions), and distinguishes those from the external states. Internal, sensory and active states are characterized functionally at a very coarse level of grain. For example, sensory states are defined as part of a

Markov blanket that separates the system from its surroundings (Friston 2013). All this means is that, given knowledge about the current sensory state, the internal states of the system are conditionally independent from the external states. But to have sensory states in this technical sense, all that is required is for the system to have a boundary – which is to say that *it is a system* distinguishable from its environment. On this construal, not only retinal or tactile input to the human brain, but also states of a plasma membrane shielding single cell's organelle from external environment count as 'sensory'. This shows, again, how FEP only puts extremely general constraints on the causal organization of organisms, perhaps to the point of lacking any non-trivial commitments about it.

Although probably not conclusive, those points cast doubt over the possibility of FEP unifying cognitive science. The elegant picture of a simple principle with an explanatory scope that encompasses all living things and from which facts about causal mechanisms of cognition can be deduced may appear appealing. Under closer scrutiny, it is far from clear whether the principle in question is explanatory and whether any sort of sufficiently detailed causal story is entailed by it. Some of the unificatory allure of the predictive mind is lost.

3. Unifying cognitive science with predictive mechanisms

In this section I propose a different way of looking at the predictive mind's unificatory credentials. Roughly, the idea is that while unification cannot be derived from first principles, it may be achieved if the account of cognitive architecture that the predictive view puts forward proves to have wide explanatory scope. This puts PP, rather than FEP, center stage. I will start out by outlining PP and a different, more relaxed perspective on how it relates to FEP. Then I will combine the mechanistic view of explanation with Danks' (2014) notion of schema-centered unification to present a different interpretation of the predictive mind's unificatory role.

While FEP belongs to theoretical biology, PP constitutes the properly *cognitive* part of the 'predictive mind' view. As I take it, PP is an account of architecture which goes beyond the assumptions present in FEP (for detailed expositions, see Clark 2013, 2016; Hohwy 2013; Wiese, Metzinger 2017). It takes the neural structures to encode an internal statistical model of the causal layout of the environment, a model that has been argued to function as an action-oriented structural representation (Gładziejewski 2016; Kiefer, Hohwy 2017; Williams 2017). This model is updated to provide estimates of the most likely causes of incoming sensory signals, in a way that approximates Bayesian inference. This is achieved by sending top-down predictions aimed at minimizing the prediction error, which is the discrepancy between the predicted and incoming sensory signals. The model is hierarchical, with each level exclusively sending prediction signals to, and receiving error signals from, the level directly below it in the processing hierarchy. What is propagated up the hierarchy are just the error signals. These signals are precision-weighted according to their predicted precision, so the relative contribution to processing of top-down and bottom-up factors is flexibly regulated on the fly. This scheme can subserve perceptual processes and attention, with attention explained in terms of precision weighing. But it can also account for motor control assuming the error signal is minimized by changing the environment through action rather than by changing the internal estimates. Assuming that the brain's statistical model of the environment can be employed off-line and stores representations that substantially abstract from the sensory periphery, PP could also scale up to explain cognition classically understood (Gładziejewski 2016; Clark 2013; Hohwy 2013; but see Williams, 2018).

Because of the reservations mentioned in the previous section, I take it that the relationship between the account of cognition just outlined and the FEP is not one of entailment. Still, those two are closely related. Given that the prediction error can be treated as equivalent to the free-energy of the sensory states, PP provides a plausible account of how *some* types of organisms may realize the FEP. However, rather than assuming that there is a relation of *a priori* necessitation between the two, it seems more reasonable to treat FEP as a powerful heuristic guide for the development of PP (see Zednik, Jäkel 2016). Perhaps FEP gives rise to PP only in combination with other evolutionary or design considerations. What some organisms, like single cells or sponges, achieve through direct interactions with the environment, others can only do by intracranially predicting their own future sensory states. This way, FEP, when combined with other considerations, makes PP architecture natural to be expected as a solution to the problem of how to minimize the free energy. There is a reason why this sort of scheme would evolve. But even if PP gains some pragmatic leverage thanks to the FEP, it functions as another account of cognitive architecture on the market. It is not necessitated by first principles.

As with other proposals regarding cognitive architecture, on this view PP can only succeed insofar as it turns out fruitful in providing detailed explanatory models of cognitive phenomena, ones that are rich in empirical predictions and can survive experimental scrutiny. And assuming mechanism about cognitive-scientific explanation, these need to be models of *mechanisms*.

Here I follow authors who already opted for treating PP as a mechanism sketch (Harkness 2015; Hohwy 2018). By providing a mechanistic sketch, PP represents the relevant mechanism in terms of the functional roles played by its components, leaving out details regarding the neural structures that realize these functions (Piccinini, Craver 2011). As such, it does not stand as an explanation on its own, but constitutes an explanation-to-be, waiting to be filled out with structural and organizational details. It can only be touted as true or accurate mechanistic explanation if the relevant functional sketch is shown to correspond to the organized components of the brain which are responsible for the phenomena being explained. For example, the precision weighting may be realized by dopaminergic gating, and perhaps distinct efferent and afferent neural pathways can be ascribed the role of transmitting top-down predictions and bottom-up error signals, respectively. This is not only a rational reconstruction of what PP *should* strive for to play an explanatory role. I take it that this view is also implicitly present in the explanatory practice of the proponents of PP, who make attempts to find the neural realizers for the prediction error minimization (see e.g. Bastos, Usrey et al. 2012; Friston, FitZgerald et al. 2017; Kanai, Komura et al. 2015). Hence, based both on assumptions about the nature of explanation and the scientific practice, a crucial condition on PP's explanatory success is that it cuts cognition at its mechanistic causal joints.

I propose that this view of PP as a mechanism sketch should be nuanced in the following way. Sometimes PP is introduced using sweeping notions, like the claim that prediction error minimization is 'all the brain ever does' (see e.g. Hohwy 2013, p. 7). Although potentially true at some level of abstraction, such claims seem limited in their explanatory power. It would be uninformative to say that the *brain as such* is one big prediction-error-minimizing mechanism that gives rise to a variety of cognitive phenomena. Furthermore, this sort of 'holistic' dialectic is at odds with assumptions that mechanism makes about explanation. Mechanistic explanation is piecemeal, in that distinct cognitive phenomena are usually explained by appealing to functionally and causally distinct mechanisms. In fact, mechanisms are partially individuated based on phenomena they explain; they are always mechanisms *of* phenomena (Bechtel 2008; Craver 2007).

It is hard to see how this should not apply to PP as well. Note that there are multiple distinct models based on PP put forward as explanations of distinct phenomena. It seems unlikely that they all appeal to a *single mechanism*. PP should not be committed to the claim that, say, low-level visual edge detection, folk physics and the disruption of social cognition in autism share a common neural mechanism. In principle, it is plausible that the brain harbors a number of causally and functionally distinct mechanisms that fall under the PP scheme. There may be multiple prediction-error-minimizing hierarchies responsible for distinct phenomena. In addition, distinct levels within a single such hierarchy could count as distinct mechanisms. In other words, there may be many distinct, at least partially independent mechanisms responsible for distinct phenomena, with each of them consisting of a hierarchical model (or a single level within such model) minimizing the prediction error. We may call them 'predictive mechanisms' or 'PP-mechanisms'. This way, PP captures a pattern of functional organization that recurs throughout those mechanisms. The brain is not simply a predictive mechanism – it is a collection of predictive mechanisms.[4]

If this interpretation of PP's explanatory commitments is right, the unificatory ambitions of PP emerge as a species of what Danks calls a 'schema-centered' unification. Schema-centered unifications arise

> 'when we have a collection of distinct cognitive theories and models that are nonetheless all instantiations of the same type of structure (in some

[4] This multiple-predictive-mechanisms interpretation of PP is defended here based on methodological considerations regarding how mechanistic explanation works in general. This is not enough to completely rule out the possibility that single-mechanism view of PP is true. Ultimately, this is an empirical matter. I thank an anonymous reviewer for pointing this out.

> sense). In other words, schema-centered accounts argue for cognitive "unification" in virtue of some common template that is shared by all the individual cognitive models, rather than through shared cognitive elements (…) across those models.' (Danks 2014, p. 176)

Similarly, what we call 'PP' divides into many distinct PP-models, aimed at representing mechanisms of distinct phenomena. These models are unified not by describing a single cognitive structure (mechanism), but because they share common core assumptions about relevant mechanisms.

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. These distinct explanatory strategies can be easily discerned in existing literature, but it may be useful to list them here explicitly.

First, there may be distinct neural mechanisms which fall under the same predictive scheme. In particular, distinct phenomena could be explained by appeal to distinct prediction-error-minimizing hierarchies. For example, different sensory modalities could be underpinned by distinct, largely independent such hierarchies, each aiming to minimize the prediction error in a way that is confined to a given sensory channel. It is also well established that there is a functional specialization within modalities, e.g. with distinct cortical mechanisms responsible for extracting different visual features, like color or motion (Zeki, Watson et al. 1991). Again, from PP's unificatory standpoint, each such mechanism could be regarded as preforming the same sort of approximate Bayesian inference, with types of visual features as 'hypotheses' that best explain distinct statistical regularities in visual input.

Second, there is a possibility that distinct *levels* within a *single* hierarchy could explain distinct cognitive phenomena. Drayson (2017) argues that the causal dependency between different layers in a predictive processing hierarchy is intransitive. If level M+1 causally influences level M, and level M causally influences Level M–1, then it is not the case that Level M+1 is causally influencing Level M–1. This makes non-adjacent levels causally independent enough to be considered distinct modules, at least on quite relaxed criteria of modularity (Drayson 2017). This opens the possibility that distinct levels within a single hierarchy could serve a mechanisms of distinct phenomena. One obvious division of explanatory labor of this kind would be between perception and cognition. According to PP, different layers of the hierarchical model track causal patterns that appear at different spatiotemporal scales, with levels high in the hierarchy tracking regularities which abstract away from rapid changes of the current sensory input (Hohwy 2013). As such, it might be argued that these higher levels are well-poised to explain 'thinking' or 'higher' cognitive phenomena (however, see Williams, 2018).

A third possible strategy consists in pointing to distinct *aspects* of PP-mechanism as explanatory. That is, given a particular mechanism, certain aspects of its functioning could account for specific phenomena. For example, the estimated-precision-based regulation of gain on the prediction error signal has been put forward by the proponents of PP as an explanation of attention (Clark 2013; Hohwy 2013). By analogy, disruptions of certain aspects of the functioning of PP-mechanisms may explain cognitive disfunctions. For illustration, aberrant weighting of the error signal relative to prior beliefs has been argued to explain hallucinations and delusions that accompany mental illness (Fletcher, Firth 2009; Sterzer, Adams et al. 2018).

Fourth, the ways in which distinct PP-mechanisms become integrated may play explanatory roles. Although the present approach suggests the existence of many distinct PP-mechanisms, these do not have to be completely causally disconnected from each other. In fact, PP presents us with straightforward ways of understanding of how these mechanisms could be integrated, at least from a computational point of view. The most obvious possibility is how correlations between distinct signals (associated with distinct inferential hierarchies) can be integrated into a representation of a common cause at a higher inferential level. This is how PP accounts for multimodal integration or feature binding (Hohwy 2013; Wiese 2017). Another possibility is to treat interactions within a single inferential hierarchy as explanatory. For example, it might be argued that PP accounts for mental imagery as a sort of off-line simulation, whereby imagining results from endogenous sensory sampling (Clark 2013). This process would originate at relatively high levels of the hierarchy, generating a cascade of top-down 'mock' sensory signals, activating lower levels.[5]

4. The value of unification for mechanistic cognitive science

Underlying the discussion so far was the assumption that explanatory unification *matters*, and so PP gains some additional value due to its unificatory credentials. My aim now is to put this assumption under scrutiny. Does the promise of unification that it brings give additional credibility to PP? Does the unificatory potential confer additional explanatory value on PP? Although I do not have definite answers on offer, I will tentatively sketch out what I take to be a promising way of understanding the value of schema-centered unification. Before I proceed with this positive view, we need to be clear about the roles that explanatory unification probably *cannot* play.

Consider the relation between unification and explanation in the context of cognitive science. It is doubtful that unification can be treated as *constitutive* of an explanation (which goes contrary to an intuition that guides unificationist accounts of explanation, see Friedman 1974; Kitcher 1989). After all, there may be non-explanatory unifications. It has been argued that at least some purported unifying dynamical explanations are in fact instances where a single mathematical formalism merely *describes* multiple phenomena (see Zednik 2011). That is, they describe *what* the system is doing, rather than explain *how* it is doing it. On one reading, outlined in section 2, FEP provides a merely descriptive unification, in that it allows us to describe diverse systems in terms of generative models maximizing their own evidence.

It could be argued that while unification is not constitutive of an explanation, it is a normative criterion of its quality or its proximity to truth. Although explaining is distinct from

[5] It is important to avoid potential confusions regarding the distinct notions of 'level' at use when we talk about integrating PP-mechanisms. When we speak of interactions between 'higher' and 'lower' levels within a given inferential hierarchy, we mean processing levels, which correspond to causally related stages in a computational process. The levels are not (or do not have to be) hierarchical in the sense of being componential. The PP-mechanisms being integrated appear at the same c*omponential* or *mechanistic* level (see Craver 2007).

unifying, once you have an explanation of a phenomenon, it better be unificatory. That is, in virtue of being unificatory – and presumably along with meeting some other criteria – an explanation is a *good* explanation, or one that can be regarded as (approximately) true. However, if we assume the mechanist view of cognitive-scientific explanation, the quality of an explanation should be disentangled from whether it successfully unifies distinct phenomena. A theoretically heterogenous, disjoint, non-unified bunch of mechanistic models, fragmentized in the sense of each being directed at explaining a distinct phenomenon, can be regarded perfectly good and truth-approximating as long as those models map onto causal structure of relevant mechanisms (see Miłkowski 2016).

Still, it seems that there is a more modest role that unification (of the schema-centered variety) *could* play. Consider a following highly simplified scenario. Imagine that we are interested in providing mechanistic explanations for a set of cognitive phenomena $P_1, P_2, P_3,\dots P_n$. For each phenomenon, there are distinct, competing mechanistic models aimed at explaining it. Imagine that we are *not* in an epistemic situation where any of those models is empirically confirmed enough to emerge as a clear winner. Of course, this does not mean that we are unable to judge the quality of those proposed models, based on criteria pertaining to, say, how biologically realistic they are, how well they fit existing data, the range and level of detail of new empirical predictions they afford, etc. The point is simply that for any phenomenon, there exists no single model whose confirmatory status is high enough for us to rationally treat this model as successfully explaining the phenomenon.

Imagine, further, that the competing models of mechanisms can be divided into two categories. On the one hand, each of $P_1, P_2, P_3\dots P_n$ has a *PP-model* that aims to explain it in terms of a hierarchical generative model minimizing the prediction error signal. Each of those models describes a *distinct mechanism*, but they all belong to a single 'family' in virtue of falling under the same scheme. On the other hand, each of $P_1, P_2, P_3\dots P_n$ also has a number of alternative mechanistic models that have close to nothing in common. That is, while PP-models are unified under a common schema, other models constitute a theoretically pluralistic hodge-podge. They are based on different guiding ideas, differ in paradigmatic assumptions, pertain to different computational schemes or drop computationalism altogether. Some of those models postulate rule-based operations over symbolic representations, others explain by appeal to brain-encoded Bayesian networks, some assume rich innate knowledge, some eschew innateneness as much as possible, some appeal to representation-free, direct coupling, etc.

Importantly, in this scenario, for any explanandum phenomenon, we are not able to judge whether a PP-model or any of the other diverse models is a better explanation. Given criteria like those mentioned earlier, it may be that some of the explananda have a PP-model that is up be there among with the highest-valued competitors, while for others a PP-model may fall short of some of its rivals. On average, however, the family of PP-models does not fare significantly better or worse than the models belonging to the pluralistic bunch.

The purpose of this scenario is to present a case in which we have direct competition between a schema-based unification and pure pluralism. That is, unity is the only difference-maker between competing models here. We are presented with an epistemic situation in which we are unable to decide between competing explanantia and everything we have to guide our choice is that some of the competitors belong to a recurring explanatory pattern, while others are more like isolated islands (Miłkowski 2016). There is *nothing* going for PP-models other than the fact that each of them belongs to larger family.

Here is an intuition that I want to pump: the fact that a given PP-model fits a recurring pattern lends it additional credibility relative to rival explanations. Put differently, there is something distinctly *ad hoc* about other, fragmented explanations, and this works to PP's advantage. It does not, by itself, make PP-models unconditionally good or true. If they are to succeed *qua* explanations, it is still necessary to show that PP-models map onto actual causal structure of the brain. But absent this sort of knowledge, unity serves as additional evidence for PP-models. Other things being equal, we are rational in ascribing more credibility to the PP-model. This way, PP's promise to provide a scheme-centered unification offers a reason to care about PP, or to have additional hope that it approximates truth.

Of course, we should never be satisfied with intuition pumping alone and should rather strive to justify the intuitive judgment in a more explicit, rigorous manner. Importantly, there is an argument, due to Sober (2003; see also Foster, Sober 1994), that seems tailor-cut for the present purposes. Sober's claim is akin to the one defended here: that unification plays an evidential role, by conferring more credibility on an explanation compared to its less unified alternatives. To justify this claim, Sober makes use of Akaike's solution to the problem of statistical model selection. Roughly, the point is that unification makes explanatory models (construed in Sober's discussion as statistical models of data) simpler. When unifying, we trade a variety of distinct explanations of distinct sets of data for a single explanation of those sets of data. This way, more unified explanations (models) account for data using *fewer adjustable parameters* – they are simpler. And simplicity makes those models less susceptible to bias by noise in data. By avoiding overfitting, unified explanations turn out better at predicting new data. This argument aims to show that something of a very concrete value – namely, predictive accuracy – is gained from unification after all.

Sober did not develop his proposal with the mechanistic view of explanation in mind. The question now is whether his line of thinking applies for mechanistic explanations. A potential problem lies in that the argument just outlined equates unity with simplicity, in a way that does not translate straightforwardly to mechanism. The key is to note how mechanism individuates explanations. For Sober's argument to work, what we need is a set of phenomena $P_1, P_2, P_3 \ldots P_n$, and a competition between (1) a single explanation E whose scope encompasses all those phenomena, and (2) a set of distinct explanations $E_1, E_2, E_3 \ldots E_n$, each aimed at distinct, single phenomenon. Only then it can be argued that (1) is simpler than (2), hence offering more predictive power. But this scenario does not necessarily apply to schema-centered unification, as understood in mechanistic terms. For mechanists, what counts as distinct explanation is a separate mechanism (or a model of such mechanism). And in our imagined scenario, each of $P_1, P_2, P_3 \ldots P_n$ is explained in terms of a causally/functionally *distinct* predictive mechanism. So, in this case, when comparing PP-models with their alternatives, we are *not* trading a variety of distinct explanations for a single explanation (or more explanations for fewer explanations) of all phenomena. Hence, the direct move from unification to simplicity, crucial for Sober's argument, is prohibited. Furthermore, it is not given that for any of the explananda, a PP-model is the simplest among all the alternative explanations of this explanandum. Among the diverse competitors, there may be ones that have the advantage of being simpler (regardless even of the criterion of simplicity we may want to use). Therefore, more simplicity is not guaranteed even relative to a particular explanandum phenomenon. To sum this up, it is hard to see how schema-centered unification guarantees simplicity, and hence enhances the predictive power of an explanation. And Sober's argument for the value of unity could only work under the assumption that it does.

I think that there are at least two ways in which this argument could be countered. One, and less promising, way is to claim that Sober's reasoning can be salvaged in the present context, if we assume a different way of individuating explanations. It might be argued that although PP-models describe distinct worldly mechanisms, they form a 'single' explanation in virtue of sharing the same basic assumptions about causal/functional organization of relevant mechanisms. That is, there may at be least fewer type-individuated explanatory posits (like 'error signal', 'precision weighting', 'sensory predictions') across PP-models than across their competing alternatives. Under this construal, PP arguably *does* offer a simpler 'explanation' than the alternatives. The problem with this answer is that we would need a further reason to think that simplicity of an explanation translates to additional predictive power.

What I think is much more promising option is to drop the search for a way of defending PP-based schema-centered unification on conceptual grounds. Instead, we should rather focus on more empirical considerations to see whether we are likely to find a recurring pattern of functional organization in neural mechanisms of cognition. If this is the case, then an explanatory model that fits a larger pattern would gain additional credibility. For example, some have argued that brain circuits are redeployed across the evolutionary and ontogenetic timeline. On this view, 'it is quite common for neural circuits established for one purpose to be exapted (…) during evolution or normal development, and be put to different uses, often without losing their original functions' (Anderson 2010, p. 245). We may speculate that FEP makes a predictive mechanism likely to emerge at some point in the evolution of nervous systems. And once this kind of mechanistic organization emerges, it is then continuously redeployed for other purposes (see also Pezzulo 2017). This would make it likely that there is a recurring pattern in neural organization, such that different cognitive functions make use of mechanisms based on the same PP-based organizational scheme. Thus, assuming neural redeployment, schema-centered unification may be regarded as more likely true than rampant pluralism.

5. Conclusions

Predictive Processing (PP), the view that the brain is a predictive machine striving to minimize precision-weighted prediction error signals, has its roots in the Free Energy Principle (FEP), a biological principle meant to capture the nature of entropy-avoiding systems. In this paper, I argued that although much of PP's unificatory ambition stems from its connection to FEP, it is hard to see how the appeal to FEP could warrant those pretensions. FEP fails to unify cognitive science directly or by entailing PP. The reason for this is that FEP does not equip us with a detailed story about the *mechanisms* of cognition. Even if we agreed that FEP entails commitments about mechanisms, it would be only at the expense of them being too general, far from what cognitive scientists strive for. If the predictive story about how the brain works is to unify cognitive science, this will probably not be achieved by deducing or deriving the truth of PP from first principles. FEP serves not as an axiomatic cornerstone for cognitive science, but rather as fertile heuristic guide for developing hypotheses about how cognition works. Successful unification through PP can only be established by developing detailed PP-based mechanistic models of phenomena, verifying those models empirically and finding if they have explanatory advantages over competing models. I argued that what can be achieved this way is what Danks (2014) calls a 'schema-centered' unification. The idea is that distinct phenomena are presumably underpinned by *distinct mechanisms*, i.e. concrete, spatiotemporally bound collections of active component parts of the central nervous system. Those mechanisms interact, sometimes partially overlap, but often they are functionally or causally independent from each other. The point is that on the PP view of things these mechanisms, while separate, instantiate the same schema. I outlined different ways is which a collection of distinct PP-mechanisms could explain a variety of cognitive phenomena, unifying them under the same core schema. There is no tension between that sort of unity and the mechanistic view of cognitive-scientific explanation.

All this opens the question of why we should care about unified explanations in cognitive science at all. The mechanistic view not only claims that explanation is distinct from unification, but also that unificatory power is not a normative criterion on which a given explanation should be evaluated. To address this issue, I argued that unification could play a role in deciding between competing models of a given phenomenon, where none of those models emerges as clear winner according to other criteria of explanatory value. The fact that a cognitive model fits into a recurring pattern could be taken as lending additional credibility to this model relative to its competitors. Unity is not a universal, unconditional measure of explanatory quality – yet sometimes it could have a role to play in guiding rational choices between competing explanations.

Acknowledgments

I wish to thank two anonymous reviewers for this journal for their insightful comments. I have also benefitted from discussing a previous version of this paper at an advanced seminar at the Department of Cognitive Science at Nicolaus Copernicus University, as well as at the *Philosophy of Cognitive Science* seminar at the Institute of Philosophy and Sociology of the Polish Academy of Sciences.

Literature

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33, 245–313.

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

Bechtel, W. (2008). *Mental mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bechtel, W., Richardson, R. (1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Cambridge (MA): The MIT Press.

Breitenbach, A., Choi, Y. (2017). Pluralism and the unity of science. *The Monist*, 100, 391–405.

Brown, H., Adams, R.A., Parees, I., Edwards, M., Friston, K.J. (2013). Active inference, sensory attenuation, and illusions. *Cognitive Processing*, 14, 411–427.

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.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Chemero, A. (2011). *Radical Embodied Cognitive Science*. Cambridge (MA): The MIT Press.

Chirimuuta, M. (2017). Explanation in computational neuroscience: causal and no-causal. *The British Journal for Philosophy of Science*, 69, 849–880.

Craver, C. F. (2009). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neurosicence*. Oxford: Oxford University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 181–204.

Clark, A. (2016). *Surfing Uncertainty. Prediction, Action, and the Embodied Mind*. Oxford: Oxford University Press.

Clark, A. (2017). How to knit your own Markov blanket: Resisting the second law with metamorphic minds. In: T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*. MIND Group.

Colombo, M., Wright, C. (2017). Explanatory pluralism: An unrewarding prediction error for free energy theorists. *Brain and Cognition*, 112, 3–12.

Colombo, M., Wright, C. (2018). First principles in the life sciences: the free-energy principle, organicism, and mechanism. *Synthese*, DOI: https://doi.org/10.1007/s11229-018-01932-w.

Cummins, R. (2000). 'How does it work?' versus 'What are the laws?': Two conceptions of psychological explanation. In: F. Keil, R. A. Wilson (eds.). *Explanation and Cognition* (pp. 117–145). Cambridge (MA): The MIT Press.

Dale, R. (2008). The possibility of a pluralist cognitive science. *Journal of Experimental & Theoretical Artificial Intelligence*, 20, 155–179.

Dale, R., Dietrich, E., Chemero, A. (2009). Explanatory pluralism in cognitive science. *Cognitive Science*, 33, 739– 42.

Danks, D. (2014). *Unifying the Mind: Cognitive Representations as Graphical Models*. Cambridge (MA): The MIT Press.

Dennett, D. (1991). Real patterns. *The Journal of Philosophy*, 87, 27–51.

Downey, A. (2018). Predictive processing and the representation wars: a victory for the eliminativist (via fictionalism). *Synthese*, 195, 5115–5139.

Drayson, Z. (2017). Modularity and the predictive mind. In T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*. MIND Group.

Dupré, J. (1993). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge (MA): Harvard University Press.

Engström, J., Bärgman, J., Nilsson, D., Seppelt, B., Markkula, G.M., Bianchi Piccinini, G.F., Victor, T. (2018). Great expectations: A predictive processing account of automobile driving. *Theoretical Issues in Ergonomics Science*, 19, 156–194.

Fletcher, P.C., Frith, C.D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature reviews. Neuroscience*, 10, 48–58.

Forster, M., Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science*, 45, 1–36.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, 71, 5–19.

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

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

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

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

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

Friston, K., Thornton C., Clark, A. (2012). Free-energy minimization and the dark-room problem. *Frontiers in Psychology*, 3, 1–7.

Geuter, S., Boll, S., Eippert, F., Büchel, F. (2017). Functional dissociation of stimulus intensity encoding and predictive coding of pain in the insula. *eLife*, e24770.

Glennan, S. (2017). *The New Mechanical Philosophy*. Oxford: Oxford University Press.

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

Harkness, D. (2015). From explanatory ambition to explanatory power. In: T. Metzinger, J. M. Windt (eds.). *Open MIND*. Frankfurt am Main: MIND Group.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Hohwy, J. (2018). The predictive processing hypothesis. In A. Newen, L. Bruin, S. Gallagher (eds). *The Oxford Handbook of 4E Cognition*. Oxford: Oxford University Press.

Hohwy, J., Paton, B., Palmer, C. (2015). Distrusting the present. *Phenomenology and the Cognitive Sciences*, 15, 315–335.

Hohwy, J., Roepstorff, A., Friston, K. J. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition*, 108, 687–701.

Kanai, R., Komura, Y., Shipp, S., Friston, K. J. (2015). Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B*, 370, 20140169.

Kaplan. D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, 183, 339–373.

Kiefer, A., Hohwy, J. (2017). Content and misrepresentation in hierarchical generative models. *Synthese*, 195, 2387–2415.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In: P. Kitcher, W. C. Salmon (Eds.). *Scientific Explanation* (pp. 410–505). Minneapolis: University of Minnesota Press.

Klein, C. (2018). What do predictive coders want? *Synthese*, 195, 2541–2557.

Lange, M. (2013). What makes a scientific explanation distinctively mathematical? *The British Journal for the Philosophy of Science*, 64, 485–511.

Miłkowski, M. (2016). Unification strategies in cognitive science. *Studies in Logic, Grammar and Rhetoric*, 48, 13–33.

Pezzulo, G. (2017). Tracing the roots of cognition in predictive processing. In T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group.

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

Quadt, L. (2017). Action-oriented predictive processing and social cognition. In T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*. MIND Group.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs*, 36, 249–275.

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

Sober, E. (2003). Two uses of unification. In: F. Stadler (ed.). *Institute Vienna Circle Yearbook 2002, 2003* (pp. 205–215). Dordrecht: Kluwer.

Sterzer, P., Adams, R.A., Fletcher, P., Frith, C., Lawrie, S.M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., Corlett, P.R. (2018). The predictive coding account of psychosis. *Biological Psychiatry*, S0006-3223, 31532–4.

van Elk, M, Aleman, A. (2017). Brain mechanisms in religion and spirituality: An integrative predictive processing framework. *Neuroscience & Biobehavioral Reviews*, 73, 359-378

Weiskopf, D. (2011). Models and mechanisms in psychological explanation. *Synthese*, 183, 313–338.

Wiese, W. (2017). *Experienced Wholeness: Integrating Insights from Gestalt Theory, Cognitive Neuroscience and Predictive Processing*. Cambridge (MA): The MIT Press.

Wiese, W., Metzinger, T. (2017). Vanilla predictive processing for philosophers: A primer on predictive processing. In T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*. MIND Group.

Williams, D. (2017). Predictive processing and the representation wars. *Minds and Machines*, 28, 141–172.

Williams, D. (2018). Predictive coding and thought. *Synthese*, DOI: https://doi.org/10.1007/s11229-018-1768-x.

Zednik, C. (2011). The nature of dynamical explanation. *Philosophy of Science*, 78, 238-263.

Zednik, C., Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, 193, 3951–3985.

Zeki, S., Watson, J.D.G., Lueck, C.J., Friston, K., Kennard, C., Frackowiak, R.S.J. (1991). A direct demonstration of functional specialization in human visual cortex. *The Journal of Neuroscience*, 11, 641–649.