Synthese (2022) 200:462
https://doi.org/10.1007/s11229-022-03891-9

ORIGINAL RESEARCH

Testable or bust: theoretical lessons for predictive processing

Marcin Miłkowski[1] · Piotr Litwin[2]

Received: 6 April 2022 / Accepted: 4 September 2022 / Published online: 2 November 2022

Abstract

The predictive processing (PP) account of action, cognition, and perception is one of the most influential approaches to unifying research in cognitive science. However, its promises of grand unification will remain unfulfilled unless the account becomes theoretically robust. In this paper, we focus on empirical commitments of PP, since they are necessary both for its theoretical status to be established and for explanations of individual phenomena to be falsifiable. First, we argue that PP is a varied research tradition, which may employ various kinds of scientific representations (from theories to frameworks and toolboxes), differing in the scope of empirical commitments they entail. Two major perspectives on PP *qua* cognitive theory may then be distinguished: generalized vs. hierarchical. The first one fails to provide empirical detail, and the latter constrains possible physical implementations. However, we show that even hierarchical PP is insufficiently restrictive to disallow incorrect models and may be adjusted to explain any neurocognitive phenomenon–including non-existent or impossible ones–through flexible adjustments. This renders PP a universal modeling tool with an unrestricted number of degrees of freedom. Therefore, in contrast with declarations of its proponents, it should not be understood as a unifying theoretical perspective, but as a computational framework, possibly informing further theory development in cognitive science.

✉ Marcin Miłkowski
mmilkows@ifispan.edu.pl

Piotr Litwin
piotr.litwin@uj.edu.pl

[1] Institute of Philosophy and Sociology, Polish Academy of Sciences, ul. Nowy Świat 72, 00-330 Warszawa, Poland

[2] Consciousness Lab, Institute of Psychology, Jagiellonian University, ul. Ingardena 6, 30-060 Kraków, Poland

1 Introduction

The predictive processing (henceforth: PP) approach to cognition (Clark, 2013) understands cognitive systems–either biological or artificial–as systems which continuously infer external causes of their sensory inputs. For this purpose, they implement prediction error minimization schemes in the form of internal hierarchical models streaming down predictions of sensory activations. Each time predictions and inputs do not overlap, the models either re-parametrize to accommodate prediction errors (which enhances their predictive accuracy in the long run) or, in a more active form of inference, prescribe action policies that make sensory data fit initial predictions (for recent primers on PP, see Hohwy 2020a; Wiese & Metzinger, 2017).

Initially a successful approach in the perception science (Clark, 2013), PP was recently extrapolated to a wide array of psychological and cognitive domains, such as interoception and emotion (Barrett, 2017; Seth, 2013), self and consciousness (Dołęga & Dewhurst, 2021; Seth & Hohwy, 2021; Woźniak, 2018), higher-order cognition (Pickering & Clark, 2014), psychopathology (Corlett et al., 2019; Sterzer et al., 2018), and many others (for a comprehensive list of philosophically-oriented research on PP, see Table S1 in Hohwy 2020a). As a result, PP became one of the most influential approaches to unifying research in cognitive science: The rapidly growing number of models classified as PP gives one the impression that it could be *the* theory in cognitive science (Clark, 2016). At the same time, the exact claims of PP remain unclear, as do the methods of testing PP empirically, and even whether that is possible at all (see, e.g., Cao 2020; Sun & Firestone, 2020).

The focus of this paper is on the empirical contents of PP. Importantly, the present inquiry does not stem from doubts about the scientific credentials of PP. We are not interested in the demarcation of science from pseudoscience, as Popper (1959) famously was. We assume that PP is scientific at the outset. Moreover, we are not worried that PP might comprise principles that are considered non-empirical, or a priori.[1] In the paper, we understand PP as a broad research tradition and discuss the whole spectrum of scientific representations under the PP label. Then we analyze them to identify those which can be assessed for falsifiability or testability. For the purpose of such assessment, we complement the Popperian understanding of falsifiability with the perspective of Taatgen (2003); that is, we expand the received notion of falsifiability centered on counterexamples to include the restrictive ability of theories that disallow incorrect models. Then, we argue why PP currently falls short on meeting requirements for a well-founded theory. In essence, the promises of theoretical unification, which are associated with PP, will remain unfulfilled unless the account provides a rich theoretical understanding. We consider falsifiability as a proxy for the ability of theories to provide an understanding of empirical phenomena.

The paper proceeds as follows. In the first section, we argue that PP is best understood in terms of a fairly large and varied research tradition, which includes several

[1] For example, some propose that the Free Energy Principle (FEP) is an a priori principle (Hohwy, 2020b). Be as it may, we do not discuss the status of the FEP, its falsifiability, contents, or implications for PP, because PP does not logically imply the acceptance of the FEP, and the discussion of complex issues surrounding it requires a separate treatment.

distinct theories and a number of distinct computational implementations. We show that scientific claims of PP are testable on various scales of abstraction, from general assumptions of this tradition to models of individual phenomena. In the subsequent section, we distinguish two major perspectives on PP *qua* cognitive theory: the generalized PP, which fails to provide crucial empirical detail, and hierarchical PP theories, which constrain possible physical implementations. Next, we argue that if hierarchical PP is tasked with providing a unified theory of cognition, it cannot be treated as a computational framework–a universal modeling tool with an unbounded number of degrees of freedom. We embrace Niels Taatgen's insight that cognitive theories may have two problems: They can generate incorrect models, which actually describe what is empirically impossible, and they can have counterexamples, or phenomena which are not predicted by a theory. The latter category is easier to address by ad hoc extensions in contrast to the first one: If a theory is so flexible that it can predict just anything, it carries no empirical information. As we show, this is currently the case with PP, which therefore remains a mere computational framework, easily confused with a unifying theory. This framework may turn out unproductive for providing a deeper understanding of phenomena, unless current modeling practices are revised.

2 PP and various kinds of scientific representations

In this paper, we view PP as a research tradition. The term *research tradition* was introduced by Larry Laudan, who refined the notion of a scientific program coined by Lakatos (1970). As Lakatos (1970) observed, research programs are difficult to falsify, but they can be assessed as progressive or degenerate. However, his account of research programs requires that they be clearly divided into their immutable hard core, which defines their identity, and the "protective belt" of easily adaptable auxiliary hypotheses. But even for research programs studied by Lakatos, the existence of their immutable hard core is hardly plausible (Laudan, 1977). This is why we follow Laudan's proposal to analyze historical dynamics of research in terms of research traditions, and we see PP as a research tradition. Some bet that it is progressive (Clark, 2013, 2016; Hohwy, 2020a; Seth & Hohwy, 2021), yet there are symptoms of stagnation (Cao, 2020; Litwin & Miłkowski, 2020; Sun & Firestone, 2020).

Research traditions are characterized by three features: (a) "Every research tradition has a number of specific theories which exemplify and partially constitute it" (Laudan, 1977, p. 78); (b) they exhibit "certain metaphysical and methodological commitments which, as an ensemble, individuate the research tradition and distinguish it from others" (ibid.); (c) lastly, traditions go through a number of formulations and usually have a long history (ibid., p. 79). Let us apply this definition to PP: There are several distinct specific theories that constitute this tradition. Clark (2016) defended the view that, even though PP originated from classical representational cognitive science, it embraces crucial insights of extended and embodied approaches since cognitive control is largely realized through active inference. In contrast, Hohwy (2013, 2016) defended a fairly internalist, neurocentric, and representationalist version of PP. Finally, there is a FEP-based version, dubbed Hierarchically Mechanistic Mind, defended by Karl Friston and his collaborators as a proper model of neural architecture (Badcock et al., 2019). These three theories are quite distinct and sometimes even mutually exclusive: Hohwy's internalism may be difficult to reconcile with Clark's externalism, and is utterly incompatible with radical externalist versions of PP (Bruineberg et al., 2016; Kirchhoff & Robertson, 2018). At the same time, all these theories may be considered part of the broader research tradition, that of Bayesian accounts of cognition (Chater et al., 2010). Multiple traditions may cross in multiple ways, conspiring to create multiple variants of similar approaches.

Proponents of PP refer to it in various ways, interchangeably calling it a theory (Clark, 2013, p. 186), a framework (Clark, 2013, p. 201), but also a model (Clark, 2013, p. 181). Hence, it remains difficult to understand exactly what is meant when one talks of testability or falsifiability of PP *simpliciter*. Let us distinguish several kinds of scientific representations in computational (neuro)cognitive science to clarify this issue. First, there are computational *toolboxes* that come almost without theoretical glosses: Think of a MATLAB module. These can include various implementations of distinct predictive coding algorithms (Spratling, 2017). Second, there are computational *frameworks*, which can be used to produce particular models: Think of dedicated programming languages or generic cognitive architectures. Third, there are *theories* that can be stated in more or less abstract fashion, in terms of verbal descriptions, flowcharts, equations, axioms etc. These usually embody general principles used to systematically understand a domain of phenomena, which are then explained by producing particular *models*.

There are multiple many-to-many mappings between these entities: There could be various PP algorithms, different implementations of the same PP algorithm for particular computational platforms, and the same algorithm or its implementation could be used by researchers who subscribe to conflicting theories. Alas, this implies that the proliferation of PP models cannot be easily understood as the growing prominence of a single theory. The sheer number of combinations of these various kinds of PP representations may contribute to a combinatorial explosion of diverse approaches within a research tradition. For example, the PP algorithm known as "active inference" may underlie models developed under divergent assumptions, e.g., (1) in the internalist take on PP, active inference is just yet another part of a wider inferential process carried out by the brain (Hohwy, 2016), whereas (2) in the enactive approach to PP, active inference is all that the generative model–as a control system–ever does (Ramstead et al., 2020). Active inference is also (3) the main ingredient of a separate process theory that aims to explain neural dynamics as well as learning and behavior of living organisms (Friston et al., 2017). In addition, active inference comes in several mathematical formulations because the framework is constantly evolving (Da Costa et al., 2020).

Importantly, these various kinds of scientific representations in computational cognitive (neuro)science may have different empirical commitments. While toolboxes may carry little to no theoretical interpretation, the tool development may influence theory development (Gigerenzer, 1991, 1992). In the case of PP, there are several ways that tools impact theories: Cognitive processes are easily modeled in terms of top-down inferential processes that depend on generative models and prediction errors. Consider a modeler's idea that it might be useful to model a given mental disorder with the use of a PP algorithm. The modeler must only “fill in the blanks” (i.e., define the nodes and the ways to perturb the inferential process) to get a preliminary understanding of the cognitive phenomenon in question: The task of the researcher is to re-describe a phenomenon (e.g., a psychiatric disorder) in terms of the toolbox (e.g., a generic PP algorithm). As such, however, toolboxes need not be realistically understood and are not directly testable; only individual models can be tested. Moreover, tools may be used in a non-standard fashion and need not conform to theoretical constraints. For example, one could use original predictive coding algorithms (Cutler, 1952) to compress video files and ignore the neuroscientific uses of their later incarnations.

Frameworks carry a bit more theoretical burden. Consider that Allen Newell characterized them as “conceptual structures (often computer systems) that are reputed to be relatively content free, but that permit specific cognitive theories to be inserted in them in some fashion” (Newell, 1990, pp. 16–17). The relative lack of content implies also that they are theoretically thin. However, in contrast to tools, there are certain research exemplars one could follow to produce standard cognitive theories. For example, nosological descriptions of psychopathologies may be formalized as various forms of computational failures within the generic PP scheme (Friston et al., 2014), which creates a PP framework that may be used to generate computational models of individual mental disorders or psychopathological phenomena (e.g., Powers et al., 2017). Such a computational framework is not restricting its uses, and it’s up to the modelers’ choice to interpret parts of the model in theoretical (e.g., neurobiological or functional) predictive coding terms. Thus, what is empirically tested, again, are particular cognitive models rather than frameworks, which can be treated instrumentally and interpreted in various ways. The elbow room for such interpretations is relatively large and by itself, the PP framework may have little significance for empirical testing. However, it may be still assessed for computational theoretical constraints such as computational tractability (Kwisthout & van Rooij, 2020), even if its interpretations are ambiguous.

In contrast, theories wear their commitments on their sleeves. While any theory goes beyond mere empirical content and remains empirically indeterminate, it should be at least partially falsifiable. The issue of falsifiability and testing in general is further complicated by the Duhem-Quine thesis, which states that no single theoretical proposition can be separately verified or falsified empirically. This is because theories are considered to be structures that are tested as complete wholes.

In the case of cognitive (neuro)science, the problem is even more grave because it is not at all clear what cognitive theories are and how they are represented. Only some are stated in natural language (e.g., the theory of thinking through metaphors: Lakoff & Johnson, 1980), and others are better represented by flow-charts or diagrams with verbal comments (e.g., the filter model of attention, Broadbent, 1958), by mathematical descriptions (e.g., the account of vision, Marr, 1982) or in terms of formal specification languages (for an extended analysis, see: Cooper & Guest, 2014). Devoid of clear identity, they easily mesh with background assumptions of other theories. Flowcharts or diagrams make their interpretation in simple propositional terms difficult. And what is yet more problematic is that it is virtually impossible to distinguish theoretical assumptions from implementation details. This is a notorious problem for computational models in cognitive (neuro)science (Cooper & Guest, 2014; Cooper & Shallice, 1995). Mere implementations of computer code are not cognitive theories. These implementations leave theories highly underspecified: It is unclear which parts of code should one include and which are just helper modules or ad hoc additions (Frijda, 1967). This underspecification is particularly true of predictive coding, which has several extant implementations. For example, only some of them presume that neurons encode Gaussian probability distributions. But is this a theoretical assumption or simply part of the implementation of predictive coding? Even proponents of PP differ in their treatment of some such cases (Spratling, 2013).

As we observed elsewhere (Litwin & Miłkowski, 2020), all this contributes to unclear theoretical fundamentals of PP. Not only there are many distinct versions of PP theory, but even a single version may be difficult to interpret in non-mathematical terms. For example, it remains unclear what psychological process or property is supposed to correspond to the mathematical property dubbed “precision” (ibid.). These difficulties notwithstanding, these theories do have empirical commitments (as we show below) and, in spite of all their indeterminacies and ambiguities, they feature several assumptions that are empirically falsifiable.

Finally, PP models address individual phenomena. These models are computational (even if only verbally sketched), and as such, they may play explanatory or predictive roles, which is typical for cognitive (neuro)science (Miłkowski, 2013; Piccinini, 2020). As Morrison & Morgan (1999) noted in their analysis of models in science, these could function and be constructed autonomously from theories. This is typical for PP models that cannot be derived entirely from the underspecified theory. One could even doubt whether the growth of the PP research tradition is theory-driven, and suggest that the whole tradition should be understood as an organically growing collection of particular models. This doubt could go hand in hand with the rejection of unificatory ambitions of some PP defenders: One could embrace particular PP models without assuming that these could become the unified theory of cognition any time soon.

In our view, the development of particular models and refining of theoretical principles are best intertwined, creating a creative tension that could spark further scientific progress. As is usually the case, theoretical underpinnings are difficult to establish and revise, while pragmatic considerations drive modelers towards repurposing the theoretical instruments toward their new ends. This may, however, lead to ambiguities in how the fundamental theoretical constructs are understood and what the theory actually states. Unfortunately, this may contribute to the demise of the whole theoretical edifice in the flurry of modeling work.

3 The protean nature of PP

The core issue for falsifying a given theory is to establish its empirical commitments. This is not at all easy for PP. “Predictive processing” is used to refer both to a broad class of “top-down” theories of cognition and to more specific theories based on hierarchical predictive processing posited to occur in the cortex of the brain, which is supposed to implement or approximate some form of Bayesian inference.

The first class was characterized as "generalized predictive coding" (Bastos et al., 2012). In a nutshell, "generalized predictive coding" refers to any learning scheme that updates the way it generates its output values in response to observed errors in the previously generated values. This happens if and only if expected values generated from the extant data model (usually dubbed "generative model") can be compared against some ground truth, and then drive corrections in the model (in the Appendix, we provide more mathematical details about this procedure). For example, this is the case for any kind of gradient descent learning in AI, but also for symbolic machine learning methods. Indeed, the notion of generalized predictive coding is broad: As Bastos et al. note, "It grandfathers nearly every known statistical estimation scheme, under parametric assumptions about additive noise" (Bastos et al., 2012, p. 703). As such, it covers emulation models based on Kalman filters, proposed by Grush (2003), but also any model-driven cognitive capacity modulated by feedback errors, including, for example, model-based reinforcement learning (Dayan & Berridge, 2014). In this, PP simply cannibalizes most top-down accounts of cognition, because they are predominantly model-based and require error corrections. But by itself, this version of PP also brings little, if any, new theoretical insights about cognition.

But is this version falsifiable? Let us take generalized PP to be a unified theory of cognition: This theory can be falsified by a single counterexample to the claim that cognition requires generative models that are adapted over time by applying corrections. Take animal taxes as driven by the increasing magnitude of tracked sensory stimuli, but devoid of any generative models. For example, crickets follow the sounds made by their conspecifics, but it is unlikely that they use generative models in their phonotaxis, even if their sensory processing is generally modulated by forward models (Webb, 2004). All that sensory registration requires is that the cricket is differentially sensitive to sound (Burge, 2010), which implies the cricket's sensitivity to the information about the stimulus. Under the received interpretation, bottom-up signals are stimulus-related, for example, indicating the position of a chirping male to a female cricket. As they are acquired by the female nervous system, they make some specific difference to the downstream neural firing (i.e., Shannon information carried by sensory signals must make some causal difference if the stimulus is received). Additionally, the received interpretation includes the existence of feedback top-down modulation, which is necessary, for example, for the usual perceptual phenomena, such as habituation, to occur. Thus, there must be both bottom-up and top-down signal pathways in the cricket auditory system. A defender of the PP account of this perceptual architecture might, however, switch the labels on pathways, by saying that the bottom-up pathway carries error information related to the predicted cause of the sound, i.e., the position of a male chirping cricket, and the top-down pathway provides predictions of future error input based on previously received error information. Then, the empirical evidence that allows us to ascribe cricket neurons the function to carry information about the sound is also evidence for the generalized PP reinterpretation. This move allows a PP theorist to cast their net very wide, unfortunately at the risk of trivializing the theory (Cao, 2020).

But there is a second way to think of PP. To understand that it is much more constrained, one can simply supplement generalized predictive coding with a formal specification of hierarchical message passing and stipulate that PP formal constructs correspond to (roughly) Bayesian concepts such as precision and conditional expectation about (hidden) causes. Hierarchical PP cannot specify one particular algorithm, as there are many (Spratling, 2017). Importantly, the predictive coding scheme as envisioned by PP necessarily entails hierarchical organization of the inferential machinery, as nicely summarized by the defenders of the Hierarchically Mechanistic Mind (HMM) hypothesis:

> In particular, the HMM relies on the directly testable second-order hypothesis that the brain minimises prediction error via hierarchical message passing in the brain (i.e., predictive coding …), which has already been demonstrated experimentally by studies of visual processing (Badcock et al., 2019, p. 115).

While there are non-hierarchical predictive coding algorithms, as with linear predictive coding in signal processing, these are not normally considered to be part of the core PP theory of cognition (e.g., Clark 2016). Thus, the overall idea of the hierarchical PP theory is that the brain implements one of the predictive coding algorithms in a bidirectional hierarchical manner, where top-down signals represent predictions, bottom-up signals represent prediction errors, and two functionally distinct units represent predictions and prediction errors at each level of the inferential hierarchy. In other words, this definition does not include all top-down accounts of cognitive processes. There remain relevant explanatory alternatives to be considered when assessing PP explanations.

This flavor of the PP theory is more empirically informative, but it also comes at a cost. To see this, it is instructive to compare it to other theories that hypothesize propagation of error information in the brain. Initially, the proposed mathematical formalism of PP could not be mapped onto concrete neural mechanisms (Kogo & Trengove, 2015). This problem was later addressed (Keller & Mrsic-Flogel, 2018), but the proposed solution imposes additional commitments upon the architecture of canonical microcircuits (the existence of type I and type II prediction error neurons encoding, respectively, positive and negative prediction errors) while maintaining certain biologically implausible features (Spratling, 2019). Similarly, in some (but not all!) respects, hierarchical PP is less biologically plausible than other approximations of the back-propagation algorithm: Predictive coding implies implausible neuronal connectivity, such as one-to-one connections of error nodes to their corresponding value notes, which is "inconsistent with diffused patterns of neuronal connectivity in the cortex" (Whittington & Bogacz, 2019, p. 240). Finally, extant PP algorithms assume that direct communication occurs only between adjacent levels in the inferential hierarchy (Litwin & Miłkowski, 2020), and top-down connections are necessarily inhibitory (Denève & Jardri, 2016), which also seems questionable given existing knowledge on heterarchical functional anatomy of the brain (Bechtel, 2019; Pessoa, 2017, 2019).

We do not refer to the above examples to pronounce the verdict that PP is thereby disproved. What we want to stress here is that hierarchical PP is not devoid of empirical content, and, as such, may be a subject to theoretical debate, cross-model comparisons, and empirically-informed revisions generating novel predictions. For example, one may argue that top-down connections can exert excitatory influences through the suppression of inhibitory interneurons at the lower level (Kanai et al., 2015; Keller & Mrsic-Flogel, 2018). The upshot is that, while as yet we have no unequivocal neurophysiological evidence for or against hierarchical PP (Walsh et al., 2020), there are possible counterexamples that it should face, which in principle renders it a falsifiable theory.

4 Explaining just the right thing

If hierarchical PP has essential empirical commitments, it should be understood as a theory rather than a computational framework. Then, the question is whether PP has the required virtues of the theory, one of which is empirical testability. In particular, PP may be assessed for falsifiability, which we understand in the Taatgenian (2003) sense:

> The problems of a theory can be found in two categories: counterexamples, phenomena that are possible in reality but are not predicted by the theory, and incorrect models, predictions of the theory that are not possible in reality. The issue of incorrect models is especially important, because an unrestricted Turing Machine is potentially capable of predicting any conceivable cognitive phenomenon (Taatgen, 2003, p. 622).

The traditional ("Popperian") approach to falsifiability is to give counterexamples, that is, phenomena that a theory cannot easily explain. Accordingly, existing critical positions focus on individual cognitive phenomena problematic for PP, such as motivation a.k.a. "the dark room problem" (Klein, 2018; Sun & Firestone, 2020), thought (Williams, 2020) or delusional beliefs (Williams, 2018). However, even if PP is incorrect in particular cases, such incorrectness is not exactly telltale. After all, no other cognitive theory can explain everything–defenders of PP can readily admit that this also applies to PP (see, e.g., Seth et al., 2020), and respond that the coverage of phenomena explained grows dramatically over time. While we argued elsewhere that the growth in question is only apparent (Litwin & Miłkowski, 2020), we do not claim that there is absolutely no growth in coverage; it is only much slower than usually touted.

Incorrect models, even though they have been out of the spotlight in the context of falsifiability, are much more important than counterexamples, because they show that a theory is insufficiently *restrictive* to disallow false models. Such an overly expressive theory shares the fate of an unrestricted universal Turing machine, since it can produce any model – also for cognitively impossible phenomena. There is also a danger in overfitting already obtained data: A theory which merely accommodates any possible result, and morphs as soon as we gather new information, has little-to-no predictive value. Consider the following example from the literature:

> The statistical independence between the identity and location of objects in the visual world suggests an anatomical dissociation between models or representations of the "what" and "where" attributes of (hidden) causes of visual

> input (i.e., knowing what an object is does not tell you where it is). This is precisely what we see in the distinction between the ventral ("what") and dorsal ("where") streams in the cortical hierarchy (Badcock, Friston, Ramstead, et al. 2019, p. 1330)

The dissociation between ventral and dorsal streams is frequently cited as an example of adaptive functional segregation that mirrors the causal structure of the environment, as envisioned by PP (Friston, 2013; Friston & Buzsáki, 2016). It is also argued that this neuroanatomical solution improves the computational efficiency of the inferential hierarchy. The explanation has an immediate intuitive charm; however, to mitigate the risk of its being incorrect, PP should also impose in advance clear theoretical constraints on the number, reducibility, and degree of independence of factors present in the sensory input. The number of specialized sensory pathways or areas is possibly smaller than the number of orthogonal factors in the sensory stream. If theoretical claims are tailored to the evidence post hoc, a remarkable fit between what the theory *would expect* and data can easily disappear as soon as new empirical evidence arrives.

In their recent paper, Pitcher & Ungerleider (2021) provide neuroanatomical evidence that the two-pathways model of visual processing is outdated, and they call for its revision. They describe another anatomically segregated pathway whose functional properties distinctively pertain to dynamic aspects of social perception, such as the recognition of bodily movement and facial expressions. This is obviously problematic for the original PP claim. Hindsight is 20/20, and, certainly, the existence of three visual pathways, dedicated separately to statistically independent "what," "where," and "[social] who," is perfectly reasonable. However, even though the list of independent factors in the visual input may always be supplemented by new additions, this process reveals incorrect models along the way and exposes the unidirectional revision dynamics (post-hoc fitting without new predictions or *a priori* constraints) of PP.

There are also reasons to believe that the "statistical independence" hypothesis is in general incorrect. The third visual pathway shares input properties (representations) with the ventral ("what") pathway as it also processes *bodies, faces*, and *motion*. And it does so independently: Pitcher & Ungerleider (2021) provide an example of prosopagnosic patients whose superior temporal sulci (STS) exhibited regular face-selective responses despite lesions in face-selective areas in the ventral pathway. The existence of two decoupled yet co-selective areas in distinct visual pathways breaks the link between statistical independence and functional specialization. It also questions whether simple factorization aptly captures functional nuances:

> It is clear that while 'what', 'where', and 'how' can describe the many facets of visual object recognition, these terms are wholly inadequate when it comes to describing the complexity and nuances of even basic social interactions. There is no simple one-word description that can encompass the functions of the third visual pathway. Rather, it appears that the visual input into the STS is integrated with other sensory inputs to enable primates to understand and interpret the actions of others (Pitcher and Ungerleider 2021, p. 9)

Finally, the identity and location of an object are statistically independent irrespective of the way they are detected, but this functional segregation does not seem to be true of all sensory modalities. Two existing olfactory pathways are not those of "what" and "where," but of sensing the environment and of finding a receptive mate (Firestein, 2001). Three distinct processing streams have been also proposed in the somatosensory domain, although their functions remain to be defined (Saadon-Grosman et al., 2020).

The two visual pathways "hypothesis" leads astray irrespective of the angle we scrutinize it from, and illustrates a more general PP propensity for the virtually unrestricted generation of incorrect models. This propensity is not limited to the domain of neural implementation: While PP specifies a universal algorithm of unsupervised learning, applicable to a wide variety of cognitive systems, it does not put meaningful constraints on the behavioral and cognitive patterns that can evolve. According to the *complete class* theorem, any decision can be cast as Bayesian given a proper loss function and a set of prior beliefs, at least for finite probability spaces (Robert, 2007). The complete class theorem was repeatedly argued (e.g., Friston et al., 2016; Friston, 2017; Parr et al., 2018) to be a foundation for countless PP insights into abnormal behaviors as "idiosyncratic rationalities":

> (...) there is always a set of prior preferences that renders any behaviour (approximately) Bayes optimal. At first glance, this may seem disappointing; however, turning the argument on its head, the complete class theorem means that we can always characterise behaviour in terms of prior preferences (Friston et al. 2016, p. 876)[2]

However, by turning the argument on its head, we may render it invalid, which seems to be the case here. The complete class theorem is a recipe for an unrestricted modeling tool rather than informative theoretical contributions (Jones & Love, 2011; Bowers & Davis, 2012). Deriving models from individual datasets may result in non-generalizable models, biased estimates, and erroneous conclusions, especially if there can exist multiple distinct prior-cost function pairs yielding equally optimal fits (Friston, 2011). Flagging all (possible) behaviors as rational under very particular circumstances opens plenty of flexible re-parameterization options, which weakens the theory (cf. Taatgen, 2003), especially if we also consider that prediction error minimization is detached from temporal constraints of the system's performance. PP operates over *appropriately* long-term timeframes (cf. Hohwy, 2020a; Seth et al., 2020), which means that any locally suboptimal behavior can appear optimal from an *appropriately* extended perspective.

[2] This interpretation is not necessarily correct. The complete class theorem states that for any decision problem, there exists an admissible Bayesian estimator; that is, an estimator using a Bayesian decision rule is never dominated by (worse than) an estimator using some other decision rule. The fact that the Bayesian decision rule yields the best estimator for any decision problem does not entail that an admissible Bayesian estimator will always be accurate or that any behavior can be plausibly cast as Bayes optimal (we would like to thank Matteo Colombo for the clarification on this issue). However, Friston's interpretation is much more relevant to our argument, given that it is assumed by PP modelers and directly motivates their modeling practices.

By means of complete class theorem, PP may even necessitate needless reconceptualizations of relatively well-understood phenomena (Jurjako, 2022). Let us consider self-deception, which is a motivationally biased process of forming beliefs, in which they evolve in the direction opposite to what is provided by evidence. PP takes such cases as instances of idiosyncratic rationalities (given a unique history of internal model development), doing away with a palpable phenomenal distinction between beliefs and desires just because it does not differentiate between them at the algorithmic level (Jurjako, 2022; Williams, 2021). A variety of forms of motivated and unmotivated reasoning is aggregated under a homogenous category of optimality. Such a broad category lacks explanatory force, does not offer valuable insights (for a related point, see Sun & Firestone 2020), and is actually counterproductive, as it trivializes the explanations that we already have (for a comprehensive review of many different shades of self-deception, see Butterworth et al., 2022).

Summing up, PP can potentially generate a vast set of incorrect models. Its enormous flexibility results in a pitfall of "predicting" phenomena as envisioned by the current scientific consensus, including those that will turn out (or have already turned out) to be incorrect. Moreover, any behavior may be rendered optimal given a proper temporal horizon or prior-cost function coupling. Thus, hierarchical PP lacks the tools for the prior assessment of model plausibility and appears to be a computational framework under the guise of a theory. It shares the fate of other computational frameworks that are not empirically informative in themselves, and even a good fit to evidence does not necessarily make a framework valuable (Roberts & Pashler, 2000). Frameworks may be productive if they inspire novel predictions and new lines of research, or they may be unproductive if they generate models in constant need of ad hoc patching and offering few scientific insights (Griffiths et al., 2012). It is for PP modelers only to decide their ways.

5 Conclusion

Although relatively recently spawned, predictive processing is already a vast research tradition, expressed in distinct scientific representations–spanning from particular models, through (more or less general) theories to computational frameworks and toolboxes. To determine its theoretical status, we focused on the testability and empirical commitments of two main approaches to PP *qua* theory: its generalized and hierarchical versions. Generalized PP is broad enough to encompass any model-based accounts, and as such does not yield many informative insights. Hierarchical PP, on the other hand, does provide some preliminary constraints on possible implementations, which makes it a better candidate for an overarching theory in cognitive science, as frequently advertised.

As we have shown, in contrast with what its proponents declare, hierarchical PP is currently used as a mere framework: a computationally universal yet theoretically agnostic modeling tool. It can hardly be empirically disproved as it may be flexibly adapted to provide any explanation, also for phenomena which actually do not (or cannot) exist. This is because hierarchical PP lacks essential detail regarding the functioning of cognitive systems, and particular PP models can be easily customized through ad hoc assumptions and parameter adjustments. While this is common malpractice in computational modeling in cognitive (neuro)science, we urge PP defenders to fill in the important details of their grand theory. Otherwise, it may turn out to be just a series of uninformative platitudes, which, even if falsifiable, would not offer satisfactory new understanding in cognitive science.

Focusing on testability allows us also to point out that more attention should be paid to the refinement of the hierarchical message passing theory, because this is what makes PP experimentally contentful. Moreover, models should explicitly specify the properties of explananda in a theoretically-motivated fashion, and provide comparisons with alternative models. These standards may seem difficult to fulfill in the modeling practice, but failure to do so may be detrimental for the whole research tradition. In short, where PP needs more work is not exactly in providing more mathematical finesse, but in developing its theoretical underpinnings in a robust fashion.

6 Appendix

The underlying idea of generalized PP may be summarized by sketching its algorithm in neural network terms, adapted from (Whittington & Bogacz, 2017, p. 1243). Bars above variables denote vectors (e.g., $\bar{x}$). The output layer of the artificial neural network is 0, and the input layer l_{max}. The training sample is denoted by pairs of training vectors $\bar{\boldsymbol{s}}^{\boldsymbol{in}}$ and $\bar{\boldsymbol{s}}^{\boldsymbol{out}}$, which are iteratively presented to the network.

for all Data **do**

$$\bar{\boldsymbol{x}}^{(0)} \leftarrow \bar{\boldsymbol{s}}^{\boldsymbol{out}}$$

$$\bar{\boldsymbol{x}}^{(\boldsymbol{l_{max}})} \leftarrow \bar{\boldsymbol{s}}^{\boldsymbol{in}}$$

repeat.

Inference: Eqs. 1, 2.

until convergence.

Update weights: Eq. 3.

Equation 1 specifies the error $\epsilon_i^{(l)}$ depending on each value of network node activity $x_i^{(l)}$, mean $\mu_i^{(l)}$ predicted by a higher level (layer) and scaling value $\Sigma_i^{(l)}$:

$$\epsilon_i^{(l)} = \frac{x_i^{(l)} - \mu_i^{(l)}}{\Sigma_i^{(l)}}$$

(1)

Equation 2 specifies the rule for changes in $x_b^{(a)}$ over time:

$$\dot{x}_b^{(a)} = -\epsilon_b^{(a)} + \sum\nolimits_{i=1}^{n(a-1)} \epsilon_i^{(a-1)} \theta_{i,b}^{(a)} f\prime \left(x_b^{(a)} \right),$$

(2)

where $\theta_{i,b}^{(a)}$are weights of synaptic connections and *f'* is a derivative of a certain non-linear node activation function *f*. The change in synaptic weight is given in 3; variables in steady state, i.e., after convergence is achieved, are denoted with an asterisk (e.g., F*):

$$\frac{\partial F*}{\partial \theta_{b,c}^{(a)}} = \epsilon_b^{*(a-1)} f\left(x_c^{*(a)}\right)$$

(3)

While the pseudocode and all three equations were put forward in the neural network context (a proposal for a predictive coding approximation of the backpropagation algorithm with local learning rules), they can be easily understood more broadly for any machine learning scheme based on gradient descent. Generally, any array data structure will do as well, and the core of the setup consists of chosen functions for updating variable values in an array over time (e.g., even the ordinary least squares method for linear regression can be implemented this way). An artificial neural network operating in accordance with this description learns a generative model of the data, conditioned on the input and encoded in the connection weights between the layers.

Acknowledgements The authors would like to thank Matteo Colombo, Paweł Gładziejewski, Casper Hesp, Tomasz Korbak, Michał Piekarski, Maxwell Ramstead, Wiktor Rorot, as well as participants of the Neural Mechanism Online seminar and two anonymous referees of this journal for their comments on earlier drafts of this paper.

Funding statement The work in this paper was funded by a National Science Centre (Poland) research grant under the decision 2014/14/E/HS1/00803 (PI: Marcin Miłkowski). The funding source was not involved in the writing of the manuscript or decision to submit it for publication.

Declarations

Conflict of interest None.

References

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

Badcock, P. B., Friston, K. J., Ramstead, M. J. D., Ploeger, A., & Hohwy, J. (2019). The hierarchically mechanistic mind: an evolutionary systems theory of the human brain, cognition, and behavior. Cognitive Affective & Behavioral Neuroscience, 19(6), 1319–1351. https://doi.org/10.3758/s13415-019-00721-3

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1–23. https://doi.org/10.1093/scan/nsw154

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron*, *76*(4), 695–711. https://doi.org/10.1016/j.neuron.2012.10.038

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

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414. https://doi.org/10.1037/a0026450

Broadbent, D. E. (1958). *Perception and communication*. Oxford: Pergamon press

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2016). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, 1–28. https://doi.org/10.1007/s11229-016-1239-1

Burge, T. (2010). *Origins of objectivity*. Oxford: Oxford University Press

Butterworth, J., Trivers, R., & von Hippel, W. (2022). The better to fool you with: Deception and self-deception. *Current Opinion in Psychology*, *47*, 101385. https://doi.org/10.1016/j.copsyc.2022.101385

Cao, R. (2020). New Labels for Old Ideas: Predictive Processing and the Interpretation of Neural Signals. *Review of Philosophy and Psychology*, *11*(3), 517–546. https://doi.org/10.1007/s13164-020-00481-x

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811–823. https://doi.org/10.1002/wcs.79

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

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

Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research*, *27*, 42–49. https://doi.org/10.1016/j.cogsys.2013.05.001

Cooper, R. P., & Shallice, T. (1995). Soar and the case for unified theories of cognition. *Cognition*, *55*(2), 115–149. https://doi.org/10.1016/0010-0277(94)00644-Z

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R. (2019). Hallucinations and Strong Priors. *Trends in Cognitive Sciences*, *23*(2), 114–127. https://doi.org/10.1016/j.tics.2018.12.001

Cutler, C. C. (1952). Differential quantization of communication signals. https://patents.google.com/patent/US2605361A/en. Accessed 9 March 2022

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. J. (2020). Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology*, *99*, 102447. https://doi.org/10.1016/j.jmp.2020.102447

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive Affective & Behavioral neuroscience*, *14*, 473–492. https://doi.org/10.3758/s13415-014-0277-8

Denève, S., & Jardri, R. (2016). Circular inference: mistaken belief, misplaced trust. *Current Opinion in Behavioral Sciences*, *11*, 40–48. https://doi.org/10.1016/j.cobeha.2016.04.001

Dołęga, K., & Dewhurst, J. E. (2021). Fame in the predictive brain: a deflationary approach to explaining consciousness in the prediction error minimization framework. *Synthese*, *198*(8), 7781–7806. https://doi.org/10.1007/s11229-020-02548-9

Firestein, S. (2001). How the olfactory system makes sense of scents. *Nature*, *413*, 211–218. https://doi.org/10.1038/35093026

Frijda, N. H. (1967). Problems of computer simulation. *Behavioral Science*, *12*(1), 59–67. https://doi.org/10.1002/bs.3830120109

Friston, K. J. (2011). What Is Optimal about Motor Control? *Neuron*, *72*(3), 488–498. https://doi.org/10.1016/j.neuron.2011.10.018

Friston, K. J. (2013). Active inference and free energy. *Behavioral and Brain Sciences*, *36*(3), 212–213. https://doi.org/10.1017/S0140525X12002142

Friston, K. J. (2017). Precision Psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *2*(8), 640–643. https://doi.org/10.1016/j.bpsc.2017.08.007

Friston, K. J., & Buzsáki, G. (2016). The Functional Anatomy of Time: What and When in the Brain. *Trends in Cognitive Sciences*, *20*(7), 500–511. https://doi.org/10.1016/j.tics.2016.05.001

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

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

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148-158. https://doi.org/10.1016/S2215-0366(14)70275-5

Gigerenzer, G. (1991). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychological Review*, *98*(2), 254–267. https://doi.org/10.1037/0033-295X.98.2.254

Gigerenzer, G. (1992). Discovery in Cognitive Psychology: New Tools Inspire New Theories. *Science in Context*, *5*(2), 329–350. https://doi.org/10.1017/S0269889700001216

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): comment on Bowers and Davis (2012). *Psychological Bulletin*, *138*(3), 415–422. https://doi.org/10.1037/a0026884

Grush, R. (2003). In Defense of Some 'Cartesian' Assumptions Concerning the Brain and Its Operation. *Biology and Philosophy*, *18*, 53–93

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

Hohwy, J. (2016). The self-evidencing brain. *Nous*, *50*(2), 259–285. https://doi.org/10.1111/nous.12062

Hohwy, J. (2020a). New directions in predictive processing. *Mind & Language*, *35*(2), https://doi.org/10.1111/mila.12281

Hohwy, J. (2020b). Self-supervision, normativity and the free energy principle. *Synthese*. https://doi.org/10.1007/s11229-020-02622-2

Jones, M., & Love, B. C. (2011). Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*(04), 169–188. https://doi.org/10.1017/S0140525X10003134

Jurjako, M. (2022). Can predictive processing explain self-deception? *Synthese*, *200*(4), 303. https://doi.org/10.1007/s11229-022-03797-6

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: Biological Sciences*, *370*(1668), 20140169. https://doi.org/10.1098/rstb.2014.0169

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron*, *100*(2), 424–435. https://doi.org/10.1016/j.neuron.2018.10.003

Kirchhoff, M. D., & Robertson, I. (2018). Enactivism and predictive processing: a non-representational view. *Philosophical Explorations*, *21*(2), 264–281. https://doi.org/10.1080/13869795.2018.1477983

Klein, C. (2018). What do predictive coders want? *Synthese*, (195), 2541–2557. https://doi.org/10.1007/s11229-016-1250-6

Kogo, N., & Trengove, C. (2015). Is predictive coding theory articulated enough to be testable? *Frontiers in Computational Neuroscience*, *9*. https://doi.org/10.3389/fncom.2015.00111

Kwisthout, J., & van Rooij, I. (2020). Computational Resource Demands of a Predictive Bayesian Brain. *Computational Brain & Behavior*, *3*, 174–188. https://doi.org/10.1007/s42113-019-00032-3

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965. Vol. 4 Vol. 4* (pp. 91–195). Cambridge: Cambridge University Press

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago, Ill.; London: University of Chicago Press

Laudan, L. (1977). *Progress and Its Problem: Towards a Theory of Scientific Growth*. Berkeley, Calif: University of California Press

Litwin, P., & Miłkowski, M. (2020). Unification by Fiat: Arrested Development of Predictive Processing. *Cognitive Science*, *44*(7), e12867. https://doi.org/10.1111/cogs.12867

Marr, D. (1982). *Vision*. New York: W. H. Freeman and Company

Miłkowski, M. (2013). *Explaining the Computational Mind*. Cambridge, Mass.: MIT Press

Morrison, M., & Morgan, M. S. (1999). Models as mediating instruments. In M. S. Morgan, & M. Morrison (Eds.), *Models as Mediators* (pp. 10–37). Cambridge: Cambridge University Press

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, Mass. and London: Harvard University Press

Parr, T., Rees, G., & Friston, K. J. (2018). Computational Neuropsychology and Bayesian Inference. *Frontiers in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00061

Pessoa, L. (2017). A Network Model of the Emotional Brain. *Trends in Cognitive Sciences*, *21*(5), 357–371. https://doi.org/10.1016/j.tics.2017.03.002

Pessoa, L. (2019). Neural dynamics of emotion and cognition: From trajectories to underlying neural geometry. *Neural Networks*, *120*, 158–166. https://doi.org/10.1016/j.neunet.2019.08.007

Piccinini, G. (2020). *Neurocognitive Mechanisms: explaining biological cognition*. Oxford: Oxford University Press

Pickering, M. J., & Clark, A. (2014). Getting ahead: forward models and their place in cognitive architecture. *Trends in Cognitive Sciences*, 1–6. https://doi.org/10.1016/j.tics.2014.05.006

Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a Third Visual Pathway Specialized for Social Perception. *Trends in Cognitive Sciences*, *25*(2), 100–110. https://doi.org/10.1016/j.tics.2020.11.006

Popper, K. R. (1959). *The Logic of Scientific Discovery*. Oxford: Routledge

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596–600. https://doi.org/10.1126/science.aan3458

Ramstead, M. J., Kirchhoff, M. D., & Friston, K. J. (2020). A tale of two densities: active inference is enactive inference. *Adaptive Behavior*, *28*(4), 225–239. https://doi.org/10.1177/1059712319862774

Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation* (2nd ed.). New York: Springer

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–358

Saadon-Grosman, N., Arzy, S., & Loewenstein, Y. (2020). Hierarchical cortical gradients in somatosensory processing. *NeuroImage*, *222*, 117257. https://doi.org/10.1016/j.neuroimage.2020.117257

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573. https://doi.org/10.1016/j.tics.2013.09.007

Seth, A. K., & Hohwy, J. (2021). Predictive processing as an empirical theory for consciousness science. *Cognitive Neuroscience*, *12*(2), 89–90. https://doi.org/10.1080/17588928.2020.1838467

Seth, A. K., Millidge, B., Buckley, C. L., & Tschantz, A. (2020). Curious Inferences: Reply to Sun and Firestone on the Dark Room Problem. *Trends in Cognitive Sciences*, *24*(9), 681–683. https://doi.org/10.1016/j.tics.2020.05.011

Spratling, M. W. (2019). Fitting predictive coding to the neurophysiological data. *Brain Research*, *1720*, 146313. https://doi.org/10.1016/j.brainres.2019.146313

Spratling, M. W. (2013). Distinguishing theory from implementation in predictive coding accounts of brain function. *Behavioral and Brain Sciences*, *36*(3), 231–232. https://doi.org/10.1017/S0140525X12002178

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

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., et al. (2018). The Predictive Coding Account of Psychosis. *Biological Psychiatry*, *84*(9), 634–643. https://doi.org/10.1016/j.biopsych.2018.05.015

Sun, Z., & Firestone, C. (2020). The Dark Room Problem. *Trends in Cognitive Sciences*, *S1364661320300589*, https://doi.org/10.1016/j.tics.2020.02.006

Taatgen, N. A. (2003). Poppering the Newell Test. *Behavioral and Brain Sciences*, *26*(5), 621–622. https://doi.org/10.1017/S0140525X03390132

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242–268. https://doi.org/10.1111/nyas.14321

Webb, B. (2004). Neural mechanisms for prediction: do insects have forward models? *Trends in Neurosciences*, *27*(5), 278–282. https://doi.org/10.1016/j.tins.2004.03.004

Whittington, J. C. R., & Bogacz, R. (2017). An Approximation of the Error Backpropagation Algorithm in a Predictive Coding Network with Local Hebbian Synaptic Plasticity. *Neural Computation*, *29*(5), 1229–1262. https://doi.org/10.1162/NECO_a_00949

Whittington, J. C. R., & Bogacz, R. (2019). Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences*, *23*(3), 235–250. https://doi.org/10.1016/j.tics.2018.12.005

Wiese, W., & Metzinger, T. K. (2017). Vanilla PP for Philosophers: A Primer on Predictive Processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and Predictive Processing*. Frankfurt am Main: MIND Group. http://www.predictive-mind.net/DOI?isbn=9783958573024. Accessed 24 July 2022

Williams, D. (2018). Hierarchical Bayesian models of delusion. *Consciousness and Cognition*, *61*, 129–147. https://doi.org/10.1016/j.concog.2018.03.003

Williams, D. (2020). Predictive coding and thought. *Synthese*, (197), 1749–1775. https://doi.org/10.1007/s11229-018-1768-x

Williams, D. (2021). Epistemic Irrationality in the Bayesian Brain. *The British Journal for the Philosophy of Science*, *72*(4), 913–938. https://doi.org/10.1093/bjps/axz044

Woźniak, M. (2018). "I" and "Me": The Self in the Context of Consciousness. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01656

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.