Minds and Machines (2022) 32:433–459
https://doi.org/10.1007/s11023-022-09595-w

GENERAL ARTICLE

Schema-Centred Unity and Process-Centred Pluralism of the Predictive Mind

Nina Poth[1]

Received: 15 July 2021 / Accepted: 8 February 2022 / Published online: 25 February 2022

Abstract

Proponents of the predictive processing (PP) framework often claim that one of the framework's significant virtues is its unificatory power. What is supposedly unified are predictive processes in the mind, and these are explained in virtue of a common prediction error-minimisation (PEM) schema. In this paper, I argue against the claim that PP currently converges towards a unified explanation of cognitive processes. Although the notion of PEM systematically relates a set of posits such as 'efficiency' and 'hierarchical coding' into a unified conceptual schema, neither the frameworks' algorithmic specifications nor its hypotheses about their implementations in the brain are clearly unified. I propose a novel way to understand the fruitfulness of the research program in light of a set of research heuristics that are partly shared with those common to Bayesian reverse engineering. An interesting consequence of this proposal is that pluralism is at least as important as unification to promote the positive development of the predictive mind.

Keywords Predictive processing · Predictive coding · Free-energy principle · Unification · Explanation · Pluralism · PP toolbox

1 Introduction

Proponents of the predictive processing framework (PP) often claim that one of its virtues is its ability to unify many, if not all, aspects of cognition, including perception, action, learning, attention, memory, motivation, social cognition, psychopathology, language, consciousness and other phenomena (Friston, 2010; Seth, 2014; Clark, 2016; Hohwy, 2013; Kiefer & Hohwy, 2018; de Bruin & Michael, 2018; Lupyan & Clark, 2015). In each of these cases, it is assumed that agents should infer the underlying, hidden, causes (e.g., the source of a sound) of sensory inputs (e.g., the perceived sound pattern) by predicting the sensory inputs and minimising, in the

✉ Nina Poth
nina.poth@rub.de

[1] Department of Philosophy II, Ruhr-Universität Bochum, Bochum, Germany

long run, the mismatch between those predictions and the actual sensory signal at multiple levels of a hierarchical generative model (Wiese & Metzinger, 2017).

Although unification is often cited as a reason for the extraordinary excitement surrounding PP (Clark, 2013; Hohwy, 2013; Friston, 2010), there is currently no wide consensus about what the framework's unificatory status actually amounts to. One idea is that PP's unity lends explanatory grounds for choosing to work with PP in investigating cognition. For example, Clark (2013, 2016) and Hohwy (2013, 2020a) argue that prediction-error minimisation (PEM) offers a universal explanation of cognitive processes that results in greater predictive power.[1] Gładziejewski (2019, p. 670) claims that PP's unificatory credentials confer additional credibility and evidential support to specific hypotheses about predictive brain mechanisms when other criteria (e.g., empirical adequacy) fail to distinguish those hypotheses from a set of competing non-PP alternatives.[2] On the contrary, the status of the related free-energy principle (FEP) as a 'grand unifying theory' of life and mind has recently been challenged (Colombo & Wright, 2017), and Colombo et al. (2018) argue that the PP framework should be agnostic about what mechanisms are at work in characterising a given cognitive task as precision-weighted PEM, and that "credibility" and "approximation to truth" are simply the wrong terms to describe PP's unificatory value, while this negative characterisation seems to be compatible with the view that PP offers methodological benefits (Hohwy, 2020b)[3] and aesthetic appeal (Prinz, 2019).[4]

In this paper, I suggest a more nuanced perspective that balances PP's unificatory with its pluralistic aspects to positively expand its fruitfulness. In particular, I argue that PP's key concepts provide a unifying schema, but its process theories are highly pluralistic. A consequence of this result is that, contra theorists such as Clark (2013), Hohwy (2020b) and Gładziejewski (2019) and in line with Colombo and Wright (2017), PP's unity lends no strong explanatory grounds for preferring to work with it. I argue that PP's unificatory merits are not the only or major reason for choosing to work with PP on instrumental grounds, as seems to be suggested by theorists such as Hohwy (2020a). The instrumental value associated with PP bears

[1] The idea is that PP offers explanations, mainly functional ones, that are simple and cover a wide variety of phenomena, and so they have a broad scope.

[2] The key idea of the confirmation-theoretic approach to PP from Gładziejewski (2019, p. 696) is that "the fact that a given PP-model fits a recurring pattern lends it additional credibility relative to rival explanations. [... Unification] 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 the actual causal structure of the brain. But absent this sort of knowledge, unity serves as additional evidence for PP-models". A more formal account of confirmation-theoretic unification in Bayesian cognitive science is offered by Colombo and Hartmann (2017, Sect. 5.3) on the basis of a Bayesian network analysis.

[3] Hohwy (2020b) understands unification based on the FEP as a "regulatory principle, 'guiding' or 'informing' the construction of process theories".

[4] Prinz characterises PP and other approaches by Andy Clark as stories with an associated "narrative appeal" that offers "ways of thinking about minds that we can find more or less compelling because they give us new perspectives on who we are and how we relate to the world around us. Good narratives must make contact with reality, but that is not what differentiates one from another. The ones that we find compelling give us a picture of ourselves that we find rewarding in some way" (Prinz, 2019, p. 235).

many more facets than is currently reflected by researchers' polarisation towards claims about its unificatory merits. Specifically, I show that PP's fruitfulness for formulating plausible and testable hypotheses about cognitive phenomena is realised by the application of a diverse set of cognitively and pragmatically beneficial heuristics that are so far underexplored and that invite opportunities for both unification as well as for pluralism in the predictive mind. Instead of a unified account, predictive processers are free to develop heterogeneous hypotheses about how cognition works.

I begin by outlining the central tenets of the PP framework with a focus on the PEM schema and its neurocentric perspective on cognitive function in Sect. 2. In Sect. 3, I borrow Danks' (2014, Chap. 8) distinction between schema- and process-centred unifications, which I take as a standard to identify PP's unificatory status as a schema-centred unification, and I argue that PP currently fails to unify cognitive processes to a high degree since there is a plurality of PP algorithms that are difficult to integrate into a coherent set of predictions about brain function. In Sect. 4, I motivate a shift away from the idea that PP offers explanatory unifications. In Sect. 5, I argue that PP's fruitfulness for the discovery of plausible and testable hypotheses builds on the application of a diverse set of associated research heuristics, whose application is compatible with the claim that PP develops locally disconnected explanations of how the brain predicts the world. I conclude that the significance of unification should be balanced with that of pluralism in the predictive mind.

2 Central Tenets of Predictive Processing

The central assumption of the PP framework is that cognition follows a single imperative to minimise prediction error, on average and in the long run (Hohwy, 2013). As a version of the Bayesian Brain Hypothesis (Knill & Pouget, 2004), PP starts from the assumption that the brain represents information at the subpersonal level in terms of probability distributions (i.e., predictions) over possible sensory states at various levels of a hierarchical generative model. The standard view is that the model is cortically implemented by a hierarchical message-passing schema in which top-down connections of neural networks carry predictions about activities at lower levels of the cortical hierarchy and bottom-up connections carry information about errors in those predictions. Prediction-error signals encode information about the discrepancy between incoming signals and prior predictions at each level[5] (Friston, 2005; Rao & Ballard, 1999).

In perception, PEM is achieved by updating the model that generates top-down predictions at higher levels of the hierarchy in proportion to the magnitude of incoming error signals, in such a way that the resulting predictions approximate

[5] The term 'level' is taken to either refer to the different layers of an artificial neural network and its computational activity in the model or to different areas in the cortex when the model is used to characterise brain function. In the former case, levels are related functionally (the activity at one level is a function of the activity at the level below or above). In the latter case, they are related as parts to wholes, such that things at higher levels are physically composed of things at lower levels. How these two kinds of levels relate to each other is currently unclear (for review, see Sprevak, 2021b).

information from novel incoming sensory signals at lower levels. This way of updating predictions is efficient in the sense that only information about the error, as opposed to the complex incoming signal, is processed at higher levels (Rao & Ballard, 1999). This neurocentric version is complemented by an 'active inference' perspective: in action, PEM can be achieved by bringing about the predicted state with one's body (Hohwy, 2013).[6] Which option is chosen depends on the precision in proportion to which error signals are weighted. Precision is defined as the inverse of the variance of the subjective probability distribution involved in prediction. If error signals are weighted with high precision, the model is revised; if they are weighted with low precision, sensory input is adjusted so as to align with the model's predictions. Characterising perception and action in terms of precision-weighting adds to the flexibility of the framework to accommodate different aspects related to cognition, perception and action (Adams et al.'s 2013 study of the force-matching illusion provides a helpful illustration).

PP's unificatory credentials in cognitive neuroscience are often supported based on the FEP, which starts from the assumption that biological organisms need to survive and maintain homeostasis. To do this, they should seek sensory states that are highly predictable. Whether a state is predictable depends on how much information it carries for a system. Free energy is defined as "an information theory measure that bounds (is greater than) the surprise on sampling some data, given a generative model" (Friston, 2009, p. 293). This essentially narrows the range of possible physical states that the organism could be in to those that make its physical states highly probable given its surrounding environment. However, the organism itself has only indirect access to a measure of this (objective) quantity. Minimising free energy is understood to be equivalent to maximising the likelihood of incoming sensory input conditional on a representation of how that input was generated (Friston, 2005). Under the assumption that organisms represent possible sensory states in terms of subjective probability distributions, the task of minimising the discrepancy between their predictions and the actual sensory states they happen to be in (as relevant to survival) is commonly reduced to the task of minimising long-term average sensory prediction error. This task is then considered to be one of approximating Bayesian inference. What makes the FEP unifying is its simplicity and broad scope—it is considered a single principle which applies to all living organisms and, hence, to all biological forms of cognition (Friston, 2010). However, there is no wide agreement on what unification in the predictive mind actually amounts to.

3 Unification and Pluralism in the Predictive Mind

Some proponents understand PP as a "process theory of information processing" (Spratling, 2017, p. 1), and they sometimes talk as if they envision PEM to integrate processing in the brain. For instance, Hohwy (2013, p. 1) writes that PEM "is

[6] To be fair, this perspective has been developed into a framework for modelling neurocognitive processes as well (Smith et al., 2021b).

one main function of the brain, on a par with the heart's pumping of blood", and that "PEM provides a unified account of brain processes for mind, cognition, and action" (Hohwy, 2018, p. 159). Sometimes, precision-weighted PEM is presented as a canonical micro-circuit of cortical neurons (Bastos et al., 2012), as a "simple, yet comprehensive, theory of how the cerebral cortex performs Bayesian inference" (Spratling, 2016), or as "a grand unifying theory of the brain" (Clark, 2018, p. 526). Illustratively, Clark (2009, pp. 975–976) compares predictive processes to 'Escher Spaghetti' to support the claim that cognitive processing might be more integrated, and less functionally separated into distinct domains (e.g., vision, audition, etc.), than specialised researchers have been willing to admit.

These statements suggest that PP not only gives a single, accurate, description of cognitive behaviour, but that it integrates processes carrying out perception, action, and other aspects of cognition. It is not only a theoretical 'gloss' that is unified but the processes themselves are wholly predictive in nature.

3.1 A Plurality of PP Algorithms

Danks' (2014, pp. 191–204) notion of a process-centred unification can be used to put this suggestion to the test.[7] Process-centred unifications involve a single model in which the processes or operations carried out by the system are in a relevant sense *shared* across its performance in a variety of different cognitive tasks. Danks does not provide a compact definition of shared processes, but he focuses on examples of cognitive architectures such as ACT-R (Anderson, 1993). In ACT-R, each module processes production rules (if–then-statements), whose joint operation is naturally constrained by the spatio-temporal order in which the outputs of those operations are passed from one module to another. The central point is that "the same integrated system—the same processes and representations (perhaps even the same brain areas), not just the same mathematics—underlies a wide range of cognition and behavior" (Danks 2014, p. 194, emphasis added).

To make the notions of 'shared process' and 'integrated system' more precise, it is helpful to borrow from new mechanists in cognitive neuroscience, where a cognitive process is a temporally extended mechanism, and the productive continuity between its stages (these might be treated for convenience as discrete steps) is achieved by functional or causal interactions between its components and their activities (Craver, 2006; Krickel, 2018; Glennan, 1996; Bechtel & Abrahamsen, 2005). A process theory specifies the component stages and interactions that are considered relevant to a cognitive process. A process can be considered unified or shared to the extent that its component stages interact in a way that renders them invariant across different domains. The production rules in ACT-R illustrate this point: they specify the functional interactions that connect the internal operations in one module (e.g., semantic memory) to other modules (e.g., declarative memory). ACT-R is unified in the sense that the way in which these rules are compiled remains functionally

[7] These options are not exhaustive, but they seem particularly well-suited to characterise the sorts of unifications common to cognitive models.

invariant regardless of changes in the cognitive task and environmental conditions. Thus, for process-centred unifications, it is not only important that the processes carried out are all based on the same abstract specification of a function in terms of PEM, that is, there is one functional analysis for a variety of different processes, but that the way in which the components of this function are manipulated remains invariant across the system. The question is whether such a functional integration currently holds for PP as well.

The results of a review of five different PP algorithms by Spratling (2008, 2016, 2017) provide reasons to doubt that PP currently integrates predictive processes in this sense. In particular, the algorithms that currently fall under PP are **(a)** linear predictive coding in digital signal processing (Makhoul, 1975; O'Shaughnessy, 1988), **(b)** a predictive coding algorithm in the retina (Srinivasan et al. 1982), **(c)** a biased-competition version developed by Spratling (2008) to model the neural mechanisms of attention, **(d)** hierarchical predictive coding, developed by Rao and Ballard (1999) to model visual processing in the cortex, and **(e)** the free-energy version developed by Friston (2005). (c) is a non-linear extension of (d) to model biased competition in attentional selection tasks.

Mathematically, all five algorithms compute the same generic input–output function of minimising prediction error but vary in how prediction error is calculated. Specifically, they differ in details about configuration, equations, and computational activity of the associated artificial neural nets. Some algorithms align with regards to one of these features, while simultaneously adopting opposite aspects of another feature. In (a), (b) and (d), prediction error is a function of subtracting the predictive signal from the incoming signal, thereby minimising the sum of squared error. In contrast, (c) and (e) calculate prediction error as a divisive function of the input signal divided by prior predictions, minimising Kullback–Leibler divergence. However, while both (e) and (d) use both excitatory feedforward connections and inhibitory feedback connections between layers of the network, it is worth noting their differences in scope: whereas (d) is restricted to modelling early perceptual processing, (e) generalises towards other capacities associated with cognition, such as action control.

This difference is clearest in the contrast between Spratling's (c) and Rao and Ballard's (d) models, which differ with respect to their assumptions about the types of network activities that are involved in error signal coding and about whether error is computed between or within layers.[8] Crucially, these models ascribe different functions to hierarchical neural network layers and their connections in the hierarchical networks. In Rao and Ballard's model, the function of a layer is to calculate prediction error and the connections between layers of the hierarchy encode predictions.

[8] With respect to the configuration, Rao and Ballard's model assumes that connections between layers are many-to many while connections within a layer are one-to-one, but Spratling's PC/BC model assumes that connections between layers are one-to-one but connections within layers are many-to-many. Regarding computation, when Rao and Ballard's model calculates prediction error using subtraction, the PC/BC model calculates it using divisive modulation. Regarding computational activity, when error nodes in Rao and Ballard's model produce both positive and negative values, error nodes in Spratling's PC/BC model produce only positive values.

However, in Spratling's version, the weights inside a layer encode predictions, while the function of connections between layers is to compute the prediction error. Thus, when in Rao and Ballard's model, predictions are passed down to the layer below (call this stage i) and errors are passed upwards to the next layer above (call this stage ii), the opposite is the case in Spratling's model, where predictions are passed upwards (stage i*) and errors are passed downwards to the next layer below (stage ii*).

It might be objected that the two algorithms are unified because they resemble each other in several respects. For example, both assume a hierarchical network structure in which lower layers of the network are closer to the input signal and higher layers are farther removed from the input signal. In each case, a single layer consists of a set of prediction and error units connected by weighted connections. And they both assume that learning proceeds by changing the weights of the network by applying a Hebbian learning algorithm. Furthermore, each of these algorithms targets a different cognitive process, one for producing biased competition between prediction nodes and the other for producing linear coding in the visual cortex, and it is principally possible that they operate simultaneously in a cognitive system.

However, mere resemblance and parallel activity are insufficient to account for a process-centred unification. PP proponents also have to show that these processes are functionally integrated, or that they produce an explanation that is invariant across cognitive tasks. It is difficult to see how these different ways of computing PEM can be simultaneously performed by the same integrated system in a way that remains functionally invariant across vision and attentional selection tasks. It is not clear how PP algorithms functionally interact with each other, since there seems to be no productive continuity between the stages of the processes that each of them computes. In fact, stage i* does not interact with stage ii, and stage i does not interact with stage ii* either. Furthermore, the PC/PB algorithm predicts performance in biased-competition tasks and the Rao and Ballard algorithm predicts visual tasks excluding biased competition, but it is unclear how these predictions can be combined to predict performance in complex tasks that require selection of salient stimuli for visual processing, since each of these algorithms is restricted to functionally distinct subsystems, one for attention and the other for vision. Thus, in contrast to ACT-R, these algorithms do not remain invariant across tasks. Given vision Rao and Ballard's model is used, but given attentional selection, Spratling's is preferred.

As this analysis shows, we currently have a variety of distinct PP algorithms which, as a collective, model processes that are not integrated in the relevant sense of a process-centred unification. Some of these versions of PP rely on shared mathematical principles (e.g., Kullback–Leibler divergence), but lack shared assumptions about the configuration of parts (e.g., the hierarchical organisation of the network) and types of activities (e.g., whether bottom-up error signals carry positive or negative values). Many aspects in how prediction error is minimised in each model are not shared with other models. Thus, even if each algorithm computes a function that minimises prediction error, the different ways of doing so are currently not unified to a high degree. In effect, it is difficult to see how the distinct predictions that these algorithms generate (e.g., for attention and vision) can be combined into a single

coherent process theory. There could be functional integration in case there was only one algorithm that was responsible to carry out all cognitive tasks with a single predictive process. However, none of these algorithms currently stands out as the best supported. Thus, although it is correct that PP offers process-level models, these models do not jointly identify a unified cognitive process. This conflicts with proponents' assumption that cognitive processing itself is unified. Some PP theorists seem to implicitly acknowledge this by characterising the framework as a 'toolbox' with a stock of algorithms (Clark, 2013; Hohwy, 2020a; Litwin & Miłkowski, 2020). However, they fail to emphasise the lack of causal or functional integration across these tools.

3.2 Schema-Centred Unity in the Predictive Mind

On the other hand, there is a certain kind of unity to the plurality of PP algorithms, since, even if they differ regarding their specific assumptions about how error is computed etc., they all in a sense share some basic structure or "explanatory motifs" (Aitchison & Lengyel, 2017). A more suitable way to characterise this is using Danks' (2014) notion of a *schema-centred* unification.[9] This describes an abstract structure that is shared by a variety of distinct cognitive theories and models that are instantiating that structure. Schema-centred unifications combine only a few ingredients and do not identify with specific models, which typically involve additional posits that may differ significantly from each other. Prime examples are Bayesianism,[10] and connectionism[11] both of which unify a variety of different Bayesian or connectionist models of different cognitive phenomena[12] by subsuming them under a common classificatory schema.

[9] Gładziejewski (2019) uses this label to characterise PP's unificatory credentials, but does not distinguish it from process-centred unifications.

[10] Bayesian models describe cognitive tasks using three ingredients: a hypothesis space, a prior probability distribution across hypotheses, and a likelihood function that relates the evidence to hypotheses (representing the agent's expectation to observe the evidence given that the hypothesis was true). Bayes' theorem combines these ingredients in a single formal scheme: $pr(h|e) \propto pr(e|h)pr(h)$. Bayesian learners choose a hypothesis as a function of the loss expected from choosing incorrectly, or they choose the hypothesis that obtains the maximum a posteriori probability, or they average across hypotheses (Griffiths et al., 2008).

[11] In connectionism, the structure resembles a network and concerns the algorithm by which the system operates. Connectionist networks have an input layer of nodes associated with activation values, a number of intermediate layers that propagate that activation and an output layer of nodes associated with activation values indicating performance. The most common learning algorithm is backpropagation, which relies on gradient descent learning (McClelland et al., 1986).

[12] For example, specific Bayesian models may differ in the type of learning rule they use to update old priors into novel posteriors. For example, some models follow strict conditionalisation, where learners are completely certain about the evidence, and others follow Jeffrey conditionalisation, where learners' observations do not always lead to certainty about the associated evidential statement (Huber, 2016). In connectionism, we can have one model that postulates two distinct causal mechanisms for the acquisition of verb syntax in children, and another model that postulates a single mechanism (Abrahamsen & Bechtel, 2006, p. 166).

One of the main advantages to schema-centred unifications is that they are simple and tidy. However, this can be understood in various ways. For example, mathematical models are often simple not because they demand a few entities to exist in the world but because they use only a few variables or a single formula to express a complex idea. This elegance is a structural property, measured by the syntactic complexity of the models' formalisms, regardless of whether these formalisms describe real entities. This contrasts with the orthodox understanding of parsimony, according to which a simple theory demands few things to exist (Quine, 1948). For example, due to their ontological commitments, mechanistic models in cognitive science would deliver parsimonious explanations to the extent that they would postulate only a small number of entities and activities in parts of a neural system. Mechanism schemas are simple in this sense, as they leave out many components of a mechanism (Craver, 2007), while remaining committed to ontological claims.

PP unifies cognition similar to both Bayesianism and connectionism.[13] PP covers a wide variety of cognitive phenomena in terms of a relatively elegant description of PEM. PEM is elegant because we only have to accept that cognition is precision-weighted PEM, and this implies no commitment to any specific algorithm specifying how PEM is carried out. Furthermore, although PP's diverse algorithmic posits lack shared operations, PEM classifies each algorithm under a common conceptual schema. For example, although each algorithm codes information in a different way, each way counts as 'efficient'.[14] The PEM schema simplifies algorithmic specifications and identifies shared mathematical features between them. PP thereby satisfies the typical features of schema-centred unifications: we have a reduction of the total number of algorithmic specifications to a single PEM schema that combines a small set of salient mathematical features.

3.3 Shifting Focus Away from Explanatory Unification

PP's status as a schema-centred unification does not necessarily make it more explanatory. Some proponents associate PP's unificatory merits with reasons to believe that it offers better explanations than available alternatives (e.g., Gładziejewski, 2019; Clark, 2016; Hohwy, 2013). However, it is unclear what justifies this association. Perhaps schema-centred unifications could be associated with law-like explanations because they are simple and abstract. Indeed, classical explanatory unifications

[13] PEM can be seen as one way to update hypotheses in approximation to Bayesian inference. Both PEM and other Bayesian learning rules minimise the Kullback–Leibler divergence between the prior and the posterior distributions (Kwisthout et al., 2017; Sprenger & Hartmann, 2019, Chap. 1). In this sense, learning is considered to be optimal to the extent that prior information is updated conservatively or to the extent that information is integrated in a statistically optimal way. PEM might be seen as a special case of Jeffrey conditionalisation, since it is assumed that the incoming signal is noisy and uncertain. PP shares with connectionism a commitment to algorithmic specifications and assumes artificial neural networks, as illustrated in the previous section.

[14] The relevant features might not be shared by all algorithms. For example, linear predictive coding for signal detection (a) does not imply hierarchical coding. So the feature 'hierarchical' might only apply partly to the PP toolbox.

like Newtonian dynamics have been analysed in these terms. They unify by allowing the generalisation of a few or stringent patterns of argument or elegant predictions to a wide range of observations and broader regularities (Friedman, 1974; Kitcher, 1989). The FEP could fall into this category. As an idealised mathematical representation of what it means to live, it is simple and has a broad scope.[15] However, a major problem is that the simplicity associated with a schema-centred unification does not necessarily confer greater truth or empirical adequacy to it. This problem is reflected in PP's research practice. Especially in the domain of perception, much of the evidence compatible with the predictions generated by PP is simultaneously compatible with alternative approaches that suggest bottom-up processing (Walsh et al., 2020). It is also compatible with some of PP's predictions that a phenomenon in question could have been brought about by a top-down non-PP algorithm, such as pure direct variable coding (Aitchison & Lengyel, 2017). The claim that cognition corresponds to PEM is difficult to confirm or disconfirm, since this is underdetermined by the empirical data, and its mere consistency does not show that PP is the best explanation of these findings. A legitimate concern is that the PEM schema might be nothing more than scientists' preferred information-theoretic terminological 'gloss' applied to the data (Cao, 2020).

Two recent observations support the claim that PP lacks empirical adequacy. The first is that key terms constituting the PEM schema obtain no clear interpretation. For example, 'precision' is sometimes used across psychology and neuroscience to mean 'salience', 'high confidence or trust' or simply 'dopaminergic gain' (Litwin & Miłkowski, 2020, pp. 22–24) and 'prediction' is often used interchangeably with 'anticipation' and 'expectation' (Ficco et al., 2021, p. 11). As Litwin and Miłkowski (2020) suggest, without adequate means to translate between these terms, their co-identification makes it difficult to pin down which psychological entity is uniquely being referred to across studies, effectively preventing convergence to shared interpretations and accumulation of evidential support. The second observation is that the mapping between PP's algorithmic specifications and the brain's neural architecture is currently too imprecise to be uniquely testable, since additional assumptions concerning neurophysiological details are needed to implement them (Sprevak, 2021c).[16] However, since these assumptions are not direct consequences of PP's theoretical assumptions, any confirmation associated with a test of algorithmic implementations might carry only spurious support for the PEM schema (cf. Cooper & Guest, 2014).

[15] Friston (2013, pp. 112–113) excludes PP from causal explanations but this leaves it open whether they function as law-like explanations.

[16] Standard assumptions are that the cortical hierarchy in the brain and neural connections between its areas implement the hierarchical structure and connections between network layers, that cortical areas closer to the sensory surface implement lower layers of the network, and that ascending (descending) cortical pathways implement feedforward (feedback) connections in the network. Additional assumptions are often made. For example, it is often assumed that changes in the precision associated with prediction error signals are encoded in changes in the long-term post-synaptic gain of superficial pyramidal cells (Adams et al., 2013). However, this assumption is rather ad hoc (Sprevak, 2021c, pp. 25–26).

Some proponents try to circumvent these problems by separating PP's conceptual, unified aspects from the program's associated empirical investigations. For example, Hohwy (2020a) claims that the FEP is a "mathematically enshrined conceptual analysis, and therefore not something in need of empirical evidence". This limits its broad scope exclusively to the domain of analytic conclusions. However, this characterisation is at odds with scientific practice in PP research. As he himself (2020a, p. 220) observes, "there is now concerted effort in cognitive neuroscience to generate and test distinctive predictions of PP [...]." Insofar as PP counts as an approach to brain function, it targets neurocognitive phenomena and proponents clearly do attempt to test it empirically as well. Appeal to enshrined conceptual analysis evokes more questions than answers, since it remains unclear how these two features, the elegant conceptual analysis and the diverse set of hypotheses for empirical test, jointly fit to a unified explanation.

At this point, it is helpful to note that the PP framework is standardly analysed along the three levels of Marr's (1982) approach to vision as an information-processing system, which has been discussed in detail in many other places but his distinction among the three levels of analysis is worth recapitulating since it provides a scaffold for making my point. At Marr's computational level of analysis, researchers specify the task faced by the cognitive system, why it is appropriate and the logic of its potential solution. The PEM schema in PP operates perfectly at this level. Marr's level of representation and algorithm identifies the representations (e.g., zeros and ones, in the case of a cash register) and how these are manipulated by the system to solve the problem (e.g., adding two numbers). The PP toolbox corresponds to analyses at this level. Finally, level of implementation identifies how this solution is physically realised (e.g., in the hardware of the cash register). PP's claims at this level remain most speculative. It is typically suggested that PP's explanatory merits should be assessed in terms of all three levels (see Sprevak, 2021a, for a review).

This complicates the relationship between unification and explanation in the predictive mind. On the one hand, the PEM schema is highly elegant (it combines a small set of salient mathematical features to describe the system's task), and so it might count as using a highly stringent argument pattern. However, it is apparent from the above considerations that the PEM schema itself does not derive any concrete predictions about real-world phenomena, and this reduces the breadth of its scope to the domain of conceptual analysis, as opposed to empirical prediction. In these terms, it is problematic to say that the PEM schema provides an acceptable explanation. On the other hand, the PP toolbox comes closer to the domain of empirical prediction, of which it develops a relatively diverse range based on its algorithmic specifications (e.g., the prediction that error signals are passed upwards, which follows from Rao and Ballard's model and that signals are passed downwards, which follows from Spratling's biased-competition model; predictions about changes in dopaminergic-gain that surround common Friston models likewise result from adopting the free-energy version). However, as is argued in Sect. 3.1, the PP toolbox (alone) does not unify these potential predictions in a sense that corresponds to a very stringent or invariant argument pattern. In these terms, it is problematic to say that PP's predictions provide a unification. Together, these trends evoke a tension that does not fit well into the unificationist account of explanation along the

lines of Kitcher (1989), according to which a unifying explanation derives the predictions about many diverse phenomena *while* using only a few or stringent argument patterns to do so. That is, PP's unifying element is not identical to an acceptable explanation of cognition, since only in conjunction with additional algorithmic and implementational theory specifications are concrete predictions being produced. However, these additions are diverse and do not obviously correspond to a stringent argument pattern. Thus, insofar as PP's explanatory merits should be assessed across the whole range of Marr's levels of analysis, one can currently not speak of a highly unified explanation.

Given these difficulties with PP's current explanatory status, I suggest to step back and first consider the question in what sense the FEP and the PEM schema can contribute to the positive development of the program, if only in virtue of their conceptual features. Subsequently, questions concerning the adequacy of the resulting scientific explanations can be asked. That is, instead of focusing on how well PP' unificatory merits already provide explanations of cognition, I suggest focusing instead on the ways in which these and other merits can be employed to develop such explanations. Thus, in the remainder of this paper, I suggest what I deem is a more suitable way to analyse PP's credentials by focusing on their cognitive and heuristic value, as opposed to whether PP is true or better confirmed than available alternatives.

4 How to Develop PP into a Fruitful Research Program

While critics such as Cao (2020) and Litwin and Miłkowski (2020) admit that PP models do provide a novel "interpretative gloss" and that this can be somehow "fruitful", they rarely explain how its "suggestive heuristic effects" could positively contribute to the "innovative and productive" status of neuroscientific research based on PP. Ivani (2019, p. 3) rightly warns that "fruitfulness [...] can be easily ascribed to many programs because its definition is loose and no clear strategy for assessing it is provided." This makes it all the more important to explain which principles PP's fruitfulness relies on. Building on Ivani's approach, a starting point to assess PP's fruitfulness is to study the research heuristics that it uses to qualitatively "extend and improve" (Ivani, 2019, p. 5) its content. It is not clear that PP explains cognition due to its unifying element, but it is still possible to expand the program in ways that are scientifically relevant. In the following, I show that PP offers several (non-exhaustive) research heuristics and discuss their contribution to the positive development of the program.

4.1 Research Heuristics Characterising PP

4.1.1 The Push-down Heuristic

Firstly, in developing specific algorithmic models of brain function, predictive processers might endorse what Zednik and Jäkel (2016, p. 3967) call the "push down

heuristic". They adopt Marr's (1982) levels-framework to spell out this heuristic to show how Bayesian reverse-engineering practices can move beyond the computational level, but their approach to heuristic strategies can be transferred to PP as well. Accordingly, this heuristic can be used to "push down" the mathematical characterisation of ideal agents (according to the FEP those maintaining homeostasis by minimising long-term average prediction error) at the computational level to characterise the varying processes at the algorithmic level of brain function that produce the behaviour. In other words, the PEM schema is pushed down to describe the activity of artificial neural network models and the interactions between deep pyramidal cells, which are said to "carry predictions", and superficial pyramidal cells, which are said to "carry prediction errors", in the brain. Zednik and Jäkel argue that the push-down heuristic is often used to choose among a set of candidate algorithmic models to work with (although this implies nothing about whether this model is more true than any of the others). We see the result of this selection in the PP toolbox: although we have no clear winning algorithm, there are relatively few algorithms that are currently considered worthwhile investigating under the assumption of the PEM schema.

4.1.2 The Tools-to-Theory Heuristic

Another heuristic borrowed from Zednik and Jäkel (2016, p. 3970) is the "tools-to-theory heuristic", which "encourage[s reverse-engineers] to introduce algorithms from completely different domains of inquiry". Examples are Gibbs sampling and particle filtering, which are special cases of Monte Carlo algorithms for approximating Bayesian inference. These algorithms were initially created for use in machine learning and statistics, but the tools-to-theories heuristic allows predictive processers to use them as adequate characterisations of psychological processes. For example, these algorithms can be used to build in limitations of memory and processing capacity by making each consecutive step in the process dependent on only the information processed in the previous step. Clark (2016, p. 61) characterises PP in a similar way:

> Action-oriented predictive processing models come tantalizing close to overcoming some of the major obstacles blocking previous attempts to ground a unified science of mind, brain, and action. They take familiar elements from existing, well-understood, computational approaches (such as unsupervised and self-supervised forms of learning using recurrent neural network architectures, and the use of probabilistic generative models for perception and action) and relate them on the one hand to a priori constraints on rational response (the Bayesian dimension) and, on the other hand, to plausible and (increasingly) testable accounts of neural implementation.

Most of the simulations of neuronal responses predicted by PP models involve the use of algorithmic tools that are borrowed from neighbouring disciplines. Examples are artificial neural networks and variational inference from machine learning (Neal & Hinton, 1998).

4.1.3 The Plausible Algorithms Heuristic

Furthermore, PP theorists seem to be using what Zednik and Jäkel (2016, pp. 3971–3972) call the "plausible algorithms heuristic". This heuristic uses established principles in psychology and neuroscience to guide selection of algorithmic models from a candidate space. PP starts from the assumption that brains face the inverse problem (they have only indirect access to the outside world and so must rely on sensory information to infer its hidden states) and employs the assumption that neural systems deal with uncertainty since sensory channels are noisy. This already restricts plausible candidates to the domain of probabilistic algorithms that can capture these statistical regularities in sensory inputs.

PP adds to this the principle of efficient coding, according to which neuronal populations encode incoming information in proportion to their channel capacity or response range, which accounts for the reliability of signal processing despite the presence of noise.[17] Two further examples are Hebbian learning and hierarchical information processing across multiple layers of the cortex. None of these principles originates from PP (these principles are basic to computational neuroscience and psychology) but each principle inspires the formulation and implicitly acts as a plausibility criterion in determining which kinds of algorithms can be added to the PP toolbox.

4.1.4 The Unification Heuristic

Furthermore, PP's unifying aspect, the PEM schema, can be seen as a conceptual tool that facilitates the search of an elegant interpretation of the available empirical findings from psychological and neuroscience. This is exemplified by the division of labour between the PP toolbox and the PEM schema, which serve both as tools that PP researchers have at their disposal. One is a practical tool accumulated via methods acquired from neighbouring disciplines for the purpose of making local predictions about brain function (when conjoined with the relevant neuroanatomical and physiological auxiliaries); the other is a cognitive tool for synthesising and systematising these findings in a relatively simple manner to allow researchers to make sense of the brain's purpose for all the neuronal activity that can be recorded. A similar idea is mentioned in Hohwy (2020b), who argues that unification based on the FEP functions as a "regulatory principle, 'guiding' or 'informing' the *construction* of process theories".

[17] This proposal has been famously spelled out in Barlow's redundancy-reduction hypothesis, according to which neurons maximise the ratio of the information that a neuron's response rate y and a stimulus x carry about each other to the neuron's channel capacity (Barlow, 2001). However, it is disputed in how far this conception of neural coding captures the neuronal representation of information, since, among other things, this conception fails to capture the aboutness of mental states (the mutual information measure is symmetric and hence insufficient to capture this directionality) and it lacks the aspect of subjectivity (see Sprevak, 2020; Isaac, 2019; Figdor, 2020, for critical discussion).

One might wonder whether unification offers any serious benefit to the fruitfulness of PP, insofar as it does not seem to be really explanatory (Sect. 3.3).[18] To explain this, let me add that unification proposes strong theoretical or conceptual constraints on what are viable algorithmic-level specifications (i.e., only those that fit the PEM schema). As is often noted by proponents of reverse-engineering, the minutiae of detail in the brain is overwhelming and researchers are rational to avail themselves to additional conceptual aids that function as guides to navigate through massive amounts of data from neuropsychological studies, hence their preference to begin at Marr's (1982) computational level of analysis. Following Marr (1982), the benefit of top-down analyses is that they provide a clear conception of the mind that contributes most to understanding a cognitive capacity (e.g., seeing) because it guides what to look for when studying the vast details of the brain. In PP, this role is fulfilled by the PEM schema, which, in its elegance, functions as a regulative ideal for the interpretation and synthesis of diverse sets of ideas, assumptions and data, often borrowed from neighbouring disciplines. Dennett (1994) adds that reverse-engineering is "the interpretation of an already existing artefact by an analysis of the design considerations that must have governed its creation" (ibid., p. 683), and it is to "prove, through building, that you have figured out how the human mechanism works" (ibid. p. 684). Here, having available starting assumptions of optimal design are a precondition for building machines that eventually serve as toy models to investigate the human mechanism.

PP researchers follow the same strategy when inferring from the observed behaviour (e.g., of neurons or persons) and the rationality assumption that the goal is to minimise prediction error the most plausible candidate cognitive processes (i.e., algorithmic implementations) that have generated it. It is in the spirit of reverse engineers such as Marr and Dennett that only gathering empirical data without an initial description of the cognitive task in mind is much more likely to lead to fruitless research. One would have much data but not know how to understand it. Their point was not to deny that descriptions of the underlying cognitive processes and states as well as the activities of cells that implement these processes are also important, but these descriptions are insufficient on their own to offer a complete understanding of a cognitive capacity. The role of the unification heuristic (but also inter-field collaboration and other heuristics following below) is to make this inference easier and faster, in this case, by providing conceptual clarity and choice criteria for further analyses at other levels. In this sense, unification is indeed a regulative device that constrains the theoretical inference process even if it might not by itself offer any substantial explanations or unique predictions concerning cognitive mechanisms. In other words, the PEM schema carries cognitive and pragmatic benefit for proponents of the program, and in this sense, its associated schema-centred unification is part of what makes PP theoretically fruitful.

There are interesting parallels to recent discussions on the role of theory for psychological science. van Rooij et al. (2018) argue that rational analyses are in some sense "as if" and contain constructive elements that might belong to the realm of the

[18] I thank an anonymous reviewer for asking me to clarify this point.

instrumental or fictional, as opposed to the actual.[19] However, it is important to note that abstraction from real-world detail does not mean regress. Even if rational analyses and unifying characterisations are often empirically underdetermined, they can still contribute to the development of plausible explanations, since abstraction and idealisation do not imply that the corresponding theory is unrealistic or inconsistent. For instance, Poirazi and Papoutsi (2020) argue that modelling facilitates data summaries and "the synthesis of existing data into concrete theories" and thereby "increases tractability" (ibid., p. 312). They likewise point to the cognitive benefits associated with unification, claiming that "[r]esearchers can make better inferences" because "refined models will unify fragmented data" (bid., p. 318). This resonates with the recent debates on the positive role of computational modelling in constraining theory building in cognitive science between van Rooij and Baggio (2021) and Guest and Martin (2021), who investigate the conditions for deriving high-verisimilitude theories before empirical test. These authors suggest that science, by which they mean the construction of experimental effects, should not be dissected from good theorising, by which they mean the development of plausible explanations of real-world cognitive capacities. This resonates well with the reverse engineering perspective, where the PEM schema does not itself explain the mind and brain, but if figures within a set of methodological tools for developing such explanations (Zednik & Jäkel, 2016, p. 397). The idea that unification is a crucial step into this direction is not novel; already in in the Critique of Pure Reason, Kant (1998) claims that unity is a precondition for for scientific inquiry and the acquisition of scientific knowledge. Insofar as "discovery [i]s part of the way science works" (Milkowski, 2014, p. 12), unification is scientifically useful, and so the elegance associated with the PEM schema is useful to the positive development of the PP program.

4.1.5 The Pushing-down Complexity Heuristic

PP also seems to use heuristics that go beyond those discussed by Zednik and Jäkel (2016). In an excellent review, Sprevak (2021c) characterises PP algorithms and their relation to implementational assumptions in a way that suggests yet another heuristic, which is that "[t]o a first approximation, the predictive coding research programme tends to 'push down' complexity and variation between cognitive processes and tasks into complexity and variation at the level of physical implementation" (Sprevak, 2021c, p. 5). According to Sprevak, PP aims for a maximally simple or elegant account of cognition and behaviour at the computational and algorithmic levels, and it accommodates the complexity of actual brains and behaviour by allowing for complex and diverse accounts at the implementational level. That is, PP does not explain complex behaviour by assuming that brains should solve a variety of different computational problems, nor does it assume that brains use many different algorithms for solving those problems. Instead, PP assumes that their physical implementations are extremely complex and diverse. Pushing down complexity is different from Zednik and Jäkel's push-down heuristic since it contributes to the

[19] I thank an anonymous reviewer for highlighting this.

open-ended character of PP (there are no constraints concerning which accounts are permissible at Marr's implementational level) and it does not make PP more testable, and what is pushed down is not mathematical structure but questions about metaphysical details. It can nevertheless be treated as another research heuristic within the PP programme. Proponents may sometimes trade unity and simplicity at the computational and algorithmic levels to include suitable aspects of diversity, for instance it might be assumed that several PP algorithms should be used to solve a given PEM problem (e.g., to accommodate complex combinations of visual and attentional selection tasks). The pushing-down complexity heuristic characterises predictive processers' tendency to tackle issues of complexity that are observed in the phenomena by adding assumptions at the level of implementation.

As an example of this heuristic, consider Sprevak's (2021b, Sect. 7) discussion of the Müller-Lyer illusion. This illusion occurs when we observe two straight lines, one with arrows pointing inwards, the other with arrows pointing outward, as being of a different length, even if they are in fact of exactly the same length. We could use a ruler and measure their length, thereby revising our belief that the lines have different length. However, there seems to be absolutely no way in which we can integrate this information with our perceptual experience. We continue to perceive the lines as being of different lengths, no matter what. Sprevak uses this example to illustrate an obstacle for the claim that cognisers (like us) always follow the goal of minimising prediction error (in the long run). The point of the illusion example is to show that we are sometimes unable to revise our predictions about the world. This, of course, is not a good sign for the PEM account of cognition. The pushing-down complexity heuristic endows predictive processers with a standard strategy to respond to such problems. They can maintain that the system attempts to solve the task of long-term PEM, but appeal to differences in its algorithmic and hardware implementations to explain away such "anomalies". In the case of the Müller-Lyer illusion, Sprevak appeals to the addition of implementational limitations such as a limit on how rapidly and how quickly physical resources can change while carrying out a specific task and algorithm to explain why a cogniser might have difficulty to adjust parameters of the internal generative model. We might fail to revise our perceptual model of the two lines because our hardware for doing so is simply too slow or unresponsive to accommodate the relevant change. This is a case of pushing-down complexity because the explanation at the computational level remains very simple, and the divergence from initial predictions is accommodated by adding parameters at the implementational level. By using the pushing-down complexity along Marr's levels, researchers obtain additional room to leave their proposals at higher levels of analysis intact while adding hypotheses to accommodate anomalies in behaviour at lower levels.

4.1.6 The Interfield-Collaborations Heuristic

Finally, the FEP may work as a discovery heuristic because it might facilitate interfield collaborations within the life sciences. Previous work by Colombo and Wright (2021) attributes a potential epistemic and pragmatic role to the FEP as a 'first principle' in the life sciences alongside organicist and mechanistic views. They suggest

that the principle can "afford a common intellectual framework for researchers from different communities to work together to answers questions of common concern" (Colombo & Wright, 2021, p. 3485). However, they do not elaborate what this means. They only argue that "[t]he diversity of expertise involved in understanding brains and organisms, and the fragmentation in present-day neuroscience and biology, highlights the need for principles that could afford a common intellectual framework for researchers from different communities to work together to answers questions of common concern. FEP is an impressive candidate for one such first principle [...]" (Colombo & Wright, 2021, p. 3485). The FEP might be best understood as a cognitive tool to facilitate collabrations for answering research questions that are shared across fields, as opposed to a theory that competes for truth. This, in essence, seems to render it as an instrument for establishing inter-field collaborations (Darden & Maull, 1977).

A typical feature of interfield-collaborations is that they allow researchers from different fields to share problems by transforming proper terms from one field to another. For example, Darden and Maull (1977) discuss the transformation of 'mutation' from genetics, where it initially meant the heritable alteration in the genotype of an organism, to biochemistry, where it is understood as the heritable alteration in base sequence. Darden and Maull emphasise that the meaning of 'mutation' is essentially *shared* by the two fields, in the sense that "... heritable alteration in the genotype was heritable alteration in the base sequence of DNA..." (Darden & Maull, 1977, p. 151); there was no replacement of one by the other term in the sense that "claims about mutation from genetics were retained and biochemical claims added" (ibid., p. 152). The important aspect of such meaning transformations for interfield collaborations is that they retain the knowledge associated with the term's original use and add knowledge in light of its novel use in the neighbouring field or theory. On the basis of such shared concepts, it is possible for researchers working with different theories to solve problems jointly. In particular, meaning trasnformations are accompanied by problem shifts such that the employment of the term in both disciplines addresses a shared problem. As a consequence of the shared theoretical vocabulary, a problem that arises in one field or theoretical context can be shifted to another field or theoretical context that might contribute with novel tools and ideas for solving it. The relevant solution is inter-theoretic because the problem is shared by the two fields or theories, and so is its solution.

Several instances of PP research similarly tend to display researchers' sharing of terms across areas in cognitive science, and thereby nourish interfield collaborations. PP transforms terms associated with the notion of prediction from information theory, physics and the philosophy of science, where it initially referred to the inference from past to future events (e.g., the sun has always risen, so how probable is it to rise tomorrow?) to the philosophy of mind, psychology, AI and neuroscience, where it refers to the probabilistic representation of certain states (e.g., how probable is it that an object is moving, given certain changes associated with its shadow?, Kersten et al., 1996) at the subpersonal level of the brain. While in physics, the aim is to predict physical happenings in the world (e.g., bodies falling, gas expanding), in PP, the aim is to describe the brain's predictions of the next sensory input by processing probabilistic representations of the world and updating these in

light of prediction errors. PP researchers borrow these terms to establish a novel understanding of the problem of prediction that builds on shared key concepts, such as "Shannon information", which initially refers to a mathematical function of the physical probability distribution over a set of outcomes (e.g., the sun is going to rise, the sun is not going to rise). PP researchers transform this concept to elucidate the contents of the brain's predictions, which are now understood as a function of the subjective probability distribution over a set of mentally-represented outcomes (e.g., the object is moving, the object is not moving) (see Sprevak, 2019; Isaac, 2019; Figdor, 2020, for philosophical discussion). Relatedly, PP researchers borrow the term "entropy", which in statistical mechanics can be understood as referring to the uncertainty associated with the precise microscopic arrangement of the components of a system, given certain macroscopic parameters like pressure, temperature and volume. A higher entropy macrostate implies a greater uncertainty regarding which particular microstate the system is in (where the macrostate is the specification of the macroscopic parameters of the system, and the microstate is the specification of the microscopic parameters of the system's components). In PP, "entropy" refers to the uncertainty associated with the internal predictions of a model of the world; if the model is associated with low entropy, this means that the states sampled from it are highly predictable. Despite their transformations, PP researchers commonly take these meanings to be shared (at least to some extent) across disciplines since their mathematical characterisation remains the same, and this correspondence allows researchers to outsource the tools and ideas available from information theory and apply them to issues concerning brain and mind. Another example is the transformation of "learning" from psychology, where it is understood as the revision of expectations about future states of the world in light of a mismatch with the actual experience (which is the prediction error) and its associated strength (Recorla & Wagner, 1972),[20] to "learning" in psychiatry and neuroscience, where it also means the revision of expectations about future states of the world in light of a mismatch with the actual experience. However, this meaning is now framed in statistical terms, where prediction is understood as the weighted mean of a random variable, its mismatch to the value that is observed is the prediction error, and the updating of predictions takes place at multiple levels of a hierarchical system (Corlett et al., 2020). Associative strength is framed in terms of precision weights onto prediction signals and error signals; where more 'precise' error signals require changes in prior belief, and more precise predictions persist information from error signals. On this basis, PP creates a narrative that carries the initial conception of learning significantly further towards other contexts, thereby offering novel perspectives to develop explanations of complex cognitive phenomena, such as, for instance, psychopathological symptoms. It thereby adds to its previous understanding, "leveraging" and "influencing" recent advances in reinforcement learning (Corlett et al., 2020). Thus, in this application of interfield collaboration, leveraging novel findings towards other ideas is a consequence of a shared conception of learning.

[20] I thank an anonymous reviewer for pointing me to this aspect.

Generally, none of these heuristics should be understood as individually necessary or sufficient for deriving explanations; they jointly partake in the development of explanations in cognitive science. Furthermore, the application of one heuristic can to some extent constrain or afford the application of another heuristic. For example, pushing-down complexity can act in the service of unification, since it contributes to keeping the description at the computational level clean and tidy. Furthermore, the elegance associated with the PEM schema can contribute to the improvement of the sharing of concepts and collaborating on solutions to problems across theoretical contexts; its simple structure can be easily understood and effectively used to systematise findings from other fields, thereby making collaboration more efficient. Finally, unification as a regulative ideal also influences the choice of criteria for what counts as a plausible algorithm from the perspective of PEM, and constrain the selection of tools to be added from neighbouring disciplines with the tools-to-theories heuristic to develop the program further.

5 Benefits and Risks Associated with PP Heuristics

The use of these discovery heuristics is not limited to establishing empirical discoveries. Research heuristics are often involved in the equally significant attempt to find appropriate formulations of testable hypotheses and to draw connections between findings that are already known in order to discover novel solutions to existing problems. The goal is not to directly choose among a set of competing hypotheses that one which is true, but "to facilitate the formulation of such hypotheses, and to thereby make possible their eventual (dis)confirmation through subsequent psychological or neuroscientific research" (Zednik &a Jäkel, 2016, p. 3971). Discovery heuristics can be considered as a means for "*developing* explanations" (Zednik & Jäkel, 2016, p. 3985, original emphasis).

There is promise that PP heuristics can contribute to progress at this front. For example, Harkness and Keshava (2017, p. 8) characterise the relationship between abstract Bayesian ideas and PP such that "by taking evidence from both the computational level (provided by Bayesian models) and implementational level (provided by neurophysiological findings) into consideration, one may, albeit provocatively, conclude that the algorithmic level (predictive processing) can be regarded as the best candidate to form the bridge between behavior and the brain." The idea is that aside from using Bayesian ideas to deal with problems of uncertainty, PP's additional algorithmic posits about precision-weighting, PEM and hierarchical message passing more readily constrain the set of available auxiliary assumptions that are needed to identify a specific set of neurocognitive phenomena (e.g., ascending and descending pathways in the brain). While Bayesian models can arrive at reliable predictions about neural behaviour on the basis of ideal observer models (e.g., Aitchison et al., 2021), this is typically only in conjunction with both additional algorithmic and neurophysiological assumptions whose specific choice is relatively unconstrained by Bayesian formalisms. The heuristics framework suggests that albeit both Bayesian models and the PP toolbox can be applied to neuroscientific

results, PP's additional algorithmic specifications more readily afford formulation of hypotheses about neurocognitive phenomena.

The major benefit of discovery heuristics is that they can improve the quality of a research program by making research more efficient and easy, given the limited time and resources that researchers face. When a research program draws attention to certain research questions and possible answers over others, thereby excluding other kinds of research questions and answers, this can contribute to its positive cognitive effects (i.e., effects on researchers' cognitive performance). For example, such effects can be thinking through problems and their solutions more effectively or finding novel answers to problems more quickly. By guiding researchers' attention to 'relevant' sets of research questions (given the core assumptions of the program) and excluding others that are deemed 'irrelevant', heuristic strategies such as the unification and pushing-down complexity heuristics can help researchers to make the identification of basic cognitive problems and their solutions manageable. This is illustrated by PP's focus on questions concerning prediction. A possible risk associated with this is ignorance. For example, there might be a risk that PP's overt focus on concepts like 'prediction' might impede the program's ability to include issues surrounding belief-desire psychology and the nature of thought (Dewhurst, 2017; Williams, 2018).[21]

Furthermore, because they are fallible, heuristic strategies make the navigation through Marr's cascade more efficient, since, "[i]f a particular heuristic leads to the formulation of many false hypotheses, it is likely to do more harm than good, because it will lead to the disproportionate consumption of time and scientific resources" (Zednik & Jäkel, 2016, p. 3985). However, this also highlights that the inferences that result from the application of PP heuristics must be handled with care. Discovery heuristics are often applied in a rather unprincipled way (their application often depends on personal interest) and, although they have practical and cognitive utility, they do not guarantee convergence to truth or explanatory advancements. Admittedly, under certain conditions, heuristic strategies *could* even serve to track the truth. In suggesting this, Zednik and Jäkel (2016, p. 3985) point to the importance of *systematic biases* when selecting solutions to targeted cognitive problems.

> Most heuristics do not highlight solutions at random, but systematically, by selecting only those solutions that exhibit a particular set of characteristics. The extent to which a heuristic strategy is an efficient guide to truth may depend on the nature of its bias, i.e. the kinds of considerations that are invoked to select individual solutions. [...] Insofar as [the principles in which these considerations are rooted] are at least approximately true, [some of these heuristics] can be viewed as reasonable guides to truth; their potential to lead

[21] However, see Colombo and Fabry (2021), for an account that desires can be accommodated in a predictive framework of self-deception, and recent work on active inference in decision-making suggest no clash between PP and folk psychology (Smith et al., 2021a). Williams (2020) outlines a set of challenges against PP's treatment of the notions of belief and thought that have so far not been met.

> researchers astray is no worse than the fallibility of [the principles exploited]. (Zednik & Jäkel, 2016, p. 3985)

PP employs a systematic bias towards using predictive thinking to explain intentional behaviour in virtue of probabilistic content that is guided by the assumption that the brain deals with uncertainty and noise reduction in probabilistic terms. This bias roots in principles from computational neuroscience such as efficient and hierarchical coding. As Zednik and Jäkel suggest, the important question is whether the basic principles in which this bias is rooted is also approximately true. Downey (2018) suggests that predictive thinking can indeed help researchers to arrive at reliable inferences about neural happenings by helping them to track the same causal changes as those in neurophysiological explanations.[22] This could be the case if the algorithms governing researchers' simulations abide to (approximately) the same rules as those governing neurophysiology and thus stipulating that neuronal populations compute these algorithms to generate activity patterns could allow researchers to track different sorts of activity patterns and changes in distinct functional areas in the cortex. However, whether this is the case has to be awaited. Reverse-engineering in cognitive science takes time, and progress in this direction can be made.[23]

A first step to realising the opportunities for PP's positive development is to pay more attention to the possible significance of pluralism in the predictive mind. Insofar as PP's diverse algorithmic posits could be grounded in neurophysiological activity, this suggests a metaphysical pluralism. That is, given their lack of invariance, PP algorithms are likely to only hold in local domains.[24] Since each algorithm targets a slightly different domain of cognition, the scope of each corresponding set of physical implementations is likely to be confined to local domains of brain function. If a given PP algorithm was correct, it would only correctly apply to a small portion of brain function (e.g., the portion responsible for biased competition), and not everywhere in the brain. This invites a nuanced perspective on unity and pluralism in the predictive mind. Eventually, unification in the development of the program might trade with pluralism as researchers move through Marr's cascade to reverse engineer the mind. Even if researchers strongly agree on the assumption that the task of cognition is PEM (i.e., they adhere to a schema-centred unity), they might end up disagreeing strongly on their specific perspectives about how

[22] He originally developed this idea under a fictionalist interpretation of 'make-believe games'. However, it seems to be readily applicable to the PP toolbox as well.

[23] One possible point of departure would be to add the constraint that PP's discovery heuristics should be used to formulate not only hypotheses that are novel and consistent with the evidence but, more importantly, hypotheses that have a high informative content and that, upon being tested, will yield informative answers to the questions about cognition that they address. Informative hypotheses have long been considered as strong because their low initial probability makes them extra informationally relevant on acquired empirical evidence for or against them (Popper, 1954; Bar-Hillel, 1955).

[24] The idea is inspired by Cartwright's (1994) approach to nomological metaphysical pluralism. She opposes fundamentalism, the claim that the laws of physics are "universal - that [they hold] everywhere and governs in all domains" by holding that the laws of physics apply only under specific conditions. For example, Newton's equations characterising falling bodies do not apply to bank notes falling from towers; they only hold in a given domain in which they are specified (e.g., not when it is windy).

cognition works. Admittedly, unification can play a guiding role in the choice of modelling approaches and algorithms from other areas (i.e., the choice is guided by the aim to choose only algorithms that suit an interpretation of the cognitive task in terms of PEM). In this regard, unification can limit, at least to some extent, the variety of distinct modelling approaches that might be considered. However, it also needs to be acknowledged that the impact of unification is itself limited, since the PEM schema still leaves much room for a diverse range of algorithmic specifications and implementations from Bayesian statistics and computational neuroscience, among which researchers might apply only their favourite tools, and not all researchers might decide to use unification as a guiding constraint. Researchers who focus on developing PP algorithms and implementations might be inclined to rely more on the tools-to-theories heuristic and find many different tools from neighbouring disciplines acceptable to model and theorise about cognition. Furthermore, the application of the pushing-down complexity heuristic invites many non-elegant, complex hypotheses to reappear at the level of implementation. Together, different PP researchers are likely to eventually diverge towards heterogeneous and potentially narrow-ranged explanations about how mind and brain work. In this sense, there remains an insufficiently acknowledged but seemingly fruitful tension between both unity and pluralism in the predictive mind.

6 Conclusion

Proponents of PP often claim that one of its greatest virtues is its unificatory power. This claim is rarely properly explained. I have argued that PP offers a schema-centred unification by virtue of the PEM framework but it fails to deliver a process-centred unification, due to its employment of a plurality of distinct algorithms in the 'PP toolbox'. Furthermore, in focusing mainly on the unificatory aspects of the predictive mind, proponents of PP have paid too little attention to a variety of other aspects that indicate its fruitfulness to making the development of explanations of neurocognitive phenomena easier and faster. In outlining an account of heuristics to be employed for the positive development of the program, I have argued that both, aspects of unity and of pluralism, find their place in the predictive mind.

Acknowledgements I am indebted to Adrian Downey, Wanja Wiese, Krzystof Dołega, Marco Faccin, Matthew Sims, Nicholas Rebol, Bartosz Radomski, Elmarie Venter, Paola Gega, Tobias Schlicht and Tobias Starzak for very constructive comments and discussion on earlier drafts of this paper.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was funded by a Grant from the Volkswagen Stiftung for the Project 'Situated Cognition. Perceiving the World and Understanding other minds' in the Institute of Philosophy II at Ruhr-Universität Bochum.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest The author declares no conflicts of interest.

References

Abrahamsen, A., & Bechtel, W. (2006). Phenomena and mechanisms: Putting the symbolic, connectionist, and dynamical systems debate in broader perspective. In *Contemporary debates in cognitive science*. Blackwell.

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry, 4*, 47.

Aitchison, L., Jegminat, J., Menendez, J. A., Pfister, J.-P., Pouget, A., & Latham, P. E. (2021). Synaptic plasticity as Bayesian inference. *Nature Neuroscience, 24*(4), 565–571.

Aitchison, L., & Lengyel, M. (2017). With or without you: Predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology, 46*, 219–227.

Anderson, J. R. (1993). *Rules of the mind*. Lawrence Erlbaum Associates, Inc.

Bar-Hillel, Y. (1955). Comments on 'degree of confirmation' by Professor K. R. Popper. *The British Journal for the Philosophy of Science, 6*(22), 155–157.

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems, 12*(3), 241.

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.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(2), 421–441.

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.

Cartwright, N. (1994). Fundamentalism vs. the patchwork of laws. *Proceedings of the Aristotelian Society, 94*, 279–292.

Clark, A. (2009). Spreading the joy? Why the machinery of consciousness is (probably) still in the head. *Mind, 118*(472), 963–993.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

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

Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences, 17*(3), 521–534.

Colombo, M., Elkin, L., & Hartmann, S. (2018). Being realist about Bayes, and the predictive processing theory of mind. *The British Journal for the Philosophy of Science, 72*(1), 1–39.

Colombo, M., & Fabry, R. E. (2021). Underlying delusion: Predictive processing, looping effects, and the personal/sub-personal distinction. *Philosophical Psychology, 34*(6), 829–855.

Colombo, M., & Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. *The British Journal for the Philosophy of Science*, *68*(2), 451–484.

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. (2021). First principles in the life sciences: The free-energy principle, organicism, and mechanism. *Synthese, 198*(14), 3463–3488.
Cooper, R. P., & Guest, O. (2014). Implementations are not specifications: Specification, replication and experimentation in computational cognitive modeling. *Cognitive Systems Research, 27*, 42–49.
Corlett, P. R., Mohanty, A., & MacDonald, A. W., III. (2020). What we think about when we think about predictive processing. *Journal of Abnormal Psychology, 129*(6), 529.
Craver, C. F. (2006). When mechanistic models explain. *Synthese, 153*(3), 355–376.
Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.
Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. MIT Press.
Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science, 44*(1), 43–64.
de Bruin, L., & Michael, J. (2018). Prediction error minimization as a framework for social cognition research. *Erkenntnis, 86*, 1–20.
Dennett, D. (1994). Cognitive science as reverse engineering: Several meanings of "Top-down" and "Bottom-up". In *International congress of logic, methodology and philosophy of science*, Dordrecht.
Dewhurst, J. (2017). Folk psychology and the Bayesian brain. In *Philosophy and predictive processing* (Vol. 9). MIND Group.
Downey, A. (2018). Predictive processing and the representation wars: A victory for the eliminativist (via fictionalism). *Synthese, 195*(12), 5115–5139.
Ficco, L., Mancuso, L., Manuello, J., Teneggi, A., Liloia, D., Duca, S., Costa, T., Kovacs, G. Z., & Cauda, F. (2021). Disentangling predictive processing in the brain: A meta-analytic study in favour of a predictive network. *Scientific Reports, 11*, 16258.
Figdor, C. (2020). Shannon + Friston = content: Intentionality in predictive signaling systems. *Synthese, 199*, 1–24.
Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy, 71*(1), 5–19.
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological sciences, 360*(1456), 815–836.
Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*(7), 293–301.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.
Friston, K. (2013). Active inference and free energy. *Behavioral and Brain Sciences, 36*(3), 212–213.
Gładziejewski, P. (2019). Mechanistic unity of the predictive mind. *Theory and Psychology, 29*(5), 657–675.
Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis, 44*(1), 49–71.
Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In *The Cambridge handbook of computational psychology* (pp. 59–100). Cambridge University Press.
Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620970585
Harkness, D. L., & Keshava, A. (2017). Moving from the what to the how and where: Bayesian models and predictive processing. In W. Wiese & T. Metzinger (Eds.), *Philosophy and predictive processing*. MIND Group.
Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
Hohwy, J. (2018). Prediction error minimization in the brain. In *The Routledge handbook of the computational mind*. Routledge.
Hohwy, J. (2020a). New directions in predictive processing. *Mind and Language, 35*(2), 209–223.
Hohwy, J. (2020b). Self-supervision, normativity and the free energy principle. *Synthese, 199*, 1–25.
Huber, F. (2016). Formal representations of belief. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, spring 2016 edition. Metaphysics Research Lab, Stanford University.
Isaac, A. M. (2019). The semantics latent in Shannon information. *The British Journal for the Philosophy of Science, 70*(1), 103–125.
Ivani, S. (2019). What we (should) talk about when we talk about fruitfulness. *European Journal for Philosophy of Science, 9*(1), 1–18.
Kant, I. (1781/1998). *Critique of Pure Reason* (P. Guyer and A.W. Wood, Trans.). Cambridge: Cambridge University Press.
Kersten, D., Knill, D. C., Mamassian, P., & Bülthoff, I. (1996). Illusory motion from shadows. *Nature, 379*, 31.
Kiefer, A., & Hohwy, J. (2018). Content and misrepresentation in hierarchical generative models. *Synthese, 195*(6), 2387–2415.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation.* University of Minnesota Press.
Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719.
Krickel, B. (2018). *The mechanical world*. Studies in brain and mind (Vol. 13). Springer.
Kwisthout, J., Bekkering, H., & Van Rooij, I. (2017). To be precise, the details don't matter: On predictive processing, precision, and level of detail of predictions. *Brain and Cognition, 112*, 84–91.
Litwin, P., & Miłkowski, M. (2020). Unification by Fiat: Arrested development of predictive processing. *Cognitive Science, 44*(7), e12867.
Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language–perception–cognition interface. *Current Directions in Psychological Science, 24*(4), 279–284.
Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE, 63*(4), 561–580.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Freeman.
McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. (1986). *Parallel distributed processing* (Vol. 1). MIT Press.
Milkowski, M. (2014). Reverse engineering in cognitive science. In *Regarding the mind, naturally: Naturalist approaches to the sciences of the mental* (pp. 12–29). Cambridge Scholars Publishing.
Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models* (pp. 355–368). Springer.
O'Shaughnessy, D. (1988). Linear predictive coding. *IEEE Potentials, 7*(1), 29–32.
Poirazi, P., & Papoutsi, A. (2020). Illuminating dendritic function with computational models. *Nature Reviews Neuroscience, 21*(6), 303–321.
Popper, K. R. (1954). Degree of confirmation. *The British Journal for the Philosophy of Science, 5*(18), 143–149.
Prinz, J. (2019). Ways of mindmaking. In *Andy Clark and his critics* (pp. 222–237). Oxford University Press.
Quine, W. V. O. (1948). On what there is. *Review of Metaphysics, 2*(5), 21–38.
Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.
Recorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning: Current research and theory*. Appleton-Century-Crofts.
Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience, 5*(2), 97–118.
Smith, R., Friston, K., & Whyte, C. (2021). A step-by-step tutorial on active inference and its application to empirical data. *Journal of Mathematical Psychology*. https://doi.org/10.31234/osf.io/b4jm6
Smith, R., Ramstead, M. J., & Kiefer, A. (2021). Active inference models do not contradict folk psychology. *Scinapse*. https://doi.org/10.31234/osf.io/kr5xf
Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research, 48*(12), 1391–1408.
Spratling, M. W. (2016). A neural implementation of Bayesian inference based on predictive coding. *Connection Science, 28*(4), 346–383.
Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition, 112*, 92–97.
Sprenger, J., & Hartmann, S. (2019). *Bayesian philosophy of science*. Oxford University Press.
Sprevak, M. (2019). Two kinds of information processing in cognition. *Review of Philosophy and Psychology, 11*, 591–611.
Sprevak, M. (2020). Two kinds of information processing in cognition. *Review of Philosophy and Psychology, 11*(3), 591–611.
Sprevak, M. (2021a). Predictive coding I: Introduction (preprint).
Sprevak, M. (2021b). Predictive coding III: The algorithmic level (preprint).
Sprevak, M. (2021c). Predictive coding IV: The implementation level (preprint).
Srinivasan, M. V., Laughlin, S. B., & Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 216*(1205), 427–459.
van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620970604
van Rooij, I., Wright, C. D., Kwisthout, J., & Wareham, T. (2018). Rational analysis, intractability, and the prospects of 'as if'-explanations. *Synthese, 195*(2), 491–510.

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.

Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 1.* MIND Group.

Williams, D. (2018). Predictive coding and thought. *Synthese, 197*, 1–27.

Williams, D. (2020). Is the brain an organ for prediction error minimization? (preprint).

Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese, 193*(12), 3951–3985.