HYBRIS nr 38 (2017)
ISSN: 1689-4286

PAWEŁ GŁADZIEJEWSKI
INSTITUTE OF PHILOSOPHY AND SOCIOLOGY
POLISH ACADEMY OF SCIENCES

JUST HOW CONSERVATIVE IS CONSERVATIVE PREDICTIVE PROCESSING?[1]

Introduction

According to the Predictive Processing (PP) framework, perception, action, and perhaps a large portion of cognition are underpinned by a mechanism of prediction error minimization. On this view, the central nervous system builds a hierarchical generative model whose job is to recapitulate the causal structure of the environment. The model generates a cascade of 'mock' predictions about incoming sensory stimuli. These predictions are matched against actual input and revised to minimize the discrepancy between the way the sensory organs are stimulated and the way they are predicted to be stimulated. What gets propagated up the hierarchy is just the prediction error signal that signifies the divergence between the two. Each level of the hierarchy minimizes the error only relative to a level directly below. The gain on the prediction error signal is mediated by precision estimations, so that, depending on the variance of the sensory signal, the processing can be modulated to rely more on the input or the internal dynamics ('prior knowledge') of the system. Perception on this view is a matter of minimizing the error by matching the internal estimates ('hypotheses') to actual environmental causes of the sensory signal. Action

[1] Work on this paper was supported by the Polish National Science Centre FUGA 3 grant (UMO-2014/12/S/ HS1/00343).

is a matter of intervening on the environment to match its state to internal estimates so that the prediction error is minimized (for reviews, see: Clark, 2013, 2016b; Hohwy, 2013; Wiese & Metzinger, 2017). And cognition can be hypothesized to result from an off-line activation of the same predictive machinery primarily involved in perception and action (see Pezzulo, 2017).

PP is surrounded by an aura of revolution, as many see in it an extremely ambitious framework that promises to provide a long-awaited (at least by some) theoretical unification for the sciences of cognition. How this supposed revolution fits into existing debates about the nature of cognition is now hotly debated. PP was initially construed in a manner that dovetails with traditional approaches in cognitive science, i.e. ones that see cognition as matter of inferential, exclusively intracranial processes involving richly structured representational states (Hohwy, 2013, in press). Following Clark (2015, 2016b), I will call this interpretation of PP 'conservative'. However, recently a number of researchers have argued that construing PP in conservative terms is mistaken. These authors opt for a 'radical' reading of PP, one that marries the framework with the idea that cognition is completely or largely non-representational as well as body- and environment-involving (Allen & Friston, 2016; Bruineberg, Kiverstein & Rietveld, 2016; Clark, 2016a, 2016b; Hutto, 2017; Orlandi, 2016, 2017). Such views situate PP firmly within the 4E approaches to understanding and studying cognition.

It seems that the literature is now shifting toward this latter, radical reading of PP. Perhaps one major reason behind this is the recognition that the PP framework finds proper theoretical home within the larger context of the Free Energy Principle (FEP; see e.g. Friston, 2010, 2013; Friston & Stephan, 2007). The FEP states that to avoid circumstances with high surprisal (i.e. ones that endanger the organism's homeostatic integrity and are unlikely given its phenotype), living creatures minimize the information-theoretic quantity of free energy. The FEP comes from a theoretical biology and applies to all, even single-cell organisms. Because, under Gaussian assumptions, long-term prediction error is equivalent to free energy, there is a tight connection between FEP and PP. PP naturally emerges as a theory of how the central nervous system, in some species, enables organisms to self-organize by avoiding surprising states. Perhaps

PP provides a sketch of a causal mechanism through which living creatures implement the FEP (Klein, 2016). The exact nature of the connection between FEP and PP is beyond the scope of the paper. I take it that although FEP puts crucial constraints on our understanding of PP (this will become apparent in the discussion to come), the two can be considered as distinct to a degree. One is a theory of life, the other is a theory of cognitive architecture tightly connected to the first. In this paper, I focus on the latter.

The aim of the present paper is to revisit the conservative construal of PP, as it is not entirely clear what this approach to understanding the framework is committed to exactly. It is all too easy to treat the conservative approach as naively attached to an outdated, overly intellectualist and internalist view of cognition. I aim to review, clarify, and disentangle the conservative commitments of PP. I take these commitments to be distinct from each other and at least partially independent. I propose that these commitments are threefold: (1) the commitment to representationalism; (2) the commitment to the notion of inference as subserving perception and action; and (3) the commitment to internalism, where internalism means that the constitutive basis of cognition does not extend beyond the central nervous system. I want to investigate and interpret each of those commitments in a way that is both grounded in PP and charitable towards proponents of the conservative approach. The discussion to follow will show that whatever genuine conservatism can be found in PP, it is as ecumenical towards 4E approaches as conservatism gets (this amounts to a intermediate, moderate position, not unlike the one proposed in: Dolega, 2017). This paper is largely a review which aims to group ideas already scattered throughout the literature and show how they fit together.

I start (in Section 1) by addressing the role that the notion of representation plays in PP. I argue that this notion can be interpreted in a weak (pragmatist) or strong (strictly realist) way. I claim that even realistically construed, representations as postulated by PP are largely within the spirit of the 4E approaches. In Section 2, I argue that PP makes use of a liberal, and yet non-trivial notion of inference. This sort of inferentialism boils down to the claim that the transitions between representational states postulated by PP are under internal control and truth-preserving (they approximately follow a truth-preserving rule). In

Section 3, I argue that PP's pretensions to internalism are not justified by the conceptual resources of the predictive framework itself. In particular, the notion of a Markov blanket is not enough to justify the commitment to internalism. I discuss how PP relates to some other, internalism-friendly ways of delineating the boundaries of mind already present in the literature.

1. The commitment to representationalism

1.1. Weak and strong representationalism of PP

Perhaps the most obvious motivation to treat PP as committed to representational states stems from the fact that the framework conceptualizes perception in terms of Bayesian inference. Minimizing the prediction error can be treated as equivalent to maximizing the posterior probability of hypotheses about the causes of the incoming sensory signal. When looked at this way, PP is simply filled with semantic notions. The perceptual system comes up with '*hypotheses' about* distal states of environment, using *'beliefs' about* which distal causes are most likely (priors) and *about* what sort of sensory 'evidence' is to be expected given some hypothesis (prior likelihoods). These hypotheses and prior beliefs are semantically evaluable: they can go wrong in the sense of *misrepresenting* the way things are. This all should not come as a surprise, as in any Bayesian theory of perception, perceptual states are individuated by their representational relations to the environment (Rescorla 2013).

However, the mere fact that semantic notions are at use does not necessarily mean a win by default for a proponent of a conservative reading of PP. There are in fact two significantly distinct ways to understand PP's commitment to representationalism. On what we can call a weak reading, the representational notions at play merely serve as what Frances Egan calls 'intentional gloss' (Egan, 2010, 2014; for proposals that explicitly interpret PP's commitment to representationalism by invoking Egan's account of conent, see: Downey, 2017; Wiese, 2017).[2] On Egan's account of content and its role in cognitive science, to make sense of physical transactions within a

[2] Note that (Wiese 2017) does not endorse Egan's pragmatism about content and is in many respects closer to a strong reading of PP's representationalism, which will be discussed in the main text.

given (computational) system of interest, its internal structures and states are mapped onto abstract mathematical entities (like numeric values). The attribution of 'mathematical contents' enables the researchers to make sense of the computations (e.g. the operation of addition) that the system in question performs. However, this is not enough to get a full understating of the system engaged in some environment-specific cognitive task. To explain how computing some function contributes to the exercise of a cognitive capacity, 'cognitive' contents must be ascribed, i.e. contents that relate parts of the internal machinery to parts of the task-specific environment. According to weak reading of the representational commitment, this is exactly the case with the sematic notions at use in PP. These contents are ascribed to the error-minimizing computational machinery to get an understanding of how it is related to the environment, a feat that is hardly achievable with purely physical and computational description.

Now, the important thing to take from this is that under this weak interpretation, any content to be found in perceptual states postulated by PP is of derived nature. Intentional properties (cognitive contents) are ascribed to the internal machinery for purely pragmatic reasons. That is, the internal states do not have cognitive contents intrinsically or essentially, but purely in virtue of interpretative acts on part of the researchers engaged in explaining cognitive functions. Thus construed, content is not a causally efficacious property of 'hypotheses' or 'prior beliefs', but may be rather seen as nothing more than a useful fiction (Downey, 2017; for a discussion of fictionalism about representation, see Sprevak, 2013). Overall, this sort of view renders PP representational in such a minimal sense that not many proponents of the 4E approaches would presumably be moved by it. After all, on this weak reading, what we are dealing with is simply a representational gloss on a non-representational mechanism. The representational vocabulary may be of crucial heuristic value, but cognition as such turns out contentless.

Still, there is a far stronger way to interpret PP's commitment to representationalism. On this reading, PP postulates a rich set of states with real, causally efficacious representational content. The justification for such a view comes from a close inspection of the role played in PP's overall computational machinery by the generative model. The generative model is

supposed to 'recapitulate' the causal structure of the environment and send a top-down stream of multi-level, cascading sensory predictions. There are strong reasons to regard the generative model as contentful and engaged in a nontrivially representational role (for more detailed and closely related discussions, see: Gładziejewski, 2016; Kiefer & Hohwy, 2017; Wiese, 2017; Williams, 2017). First, it generates, in perceptual inference, estimates of the environment which guide cognitive system's practical engagements with the environment. It is action-guiding. Second, the model's ability to play this function is dependent on how well the functional relationships between encoded variables resemble the causal structure of the environment. The degree of structural match between the model and the environment is causally relevant to a degree in which the model is effective at enabling adaptive, self-maintaining actions (see Gładziejewski & Miłkowski, 2017). This way, content becomes the fuel of practical success, not just a matter of passively mirroring the environment. Third, the model performs a largely endogenously-controlled, predictive simulation. It exhibits at least some degree of detachment or independence from current sensory stimulation. It could be argued that the simulations in question can be run purely off-line, i.e. outside of any direct engagements with the environment (Pezzulo, 2017). Fourth, insofar as the model undergoes correction in light of the prediction error, it can be said to be capable of detecting cases when its representations are inaccurate. More precisely, the Kullbach–Leibler divergence between true posterior and recognition (model-based) probability distributions can be understood as a sort of measure of misrepresentation (Kiefer & Hohwy, 2017). The lesson, then, is that the generative model constitutes an action-guiding, detachable structural representation, capable of detectable representational error. This is a robust and metaphysically realist incarnation of representationalism, arguably immune to recent trivializing arguments against representation (see Gładziejewski, 2015, 2016; Gładziejewski & Miłkowski, 2017).

1.2. Strong representationalism about PP: how conservative?

Let us focus further on PP's strong representationalism, as this is what proponents of 4E approaches would presumably take issue with. It could be suggested that by invoking the concept of an internal model, conservative

rendering of PP construes representations involved in perception as action-neutral, disembodied inner replicas or reconstructions of the world (Clark, 2015, 2016b). On closer inspection, this sort of assessment turns out unfair towards conservatism. In fact, as far as robust and metaphysically realist representationalism goes, the (strong) notion of representation in PP is very much compatible with the spirit of 4E approaches.[3] There are four reasons to see PP's commitment to strong representationalism as not-so-conservative after all.

First, note that PP postulates a complex processing architecture subserving the process of minimizing the prediction error. The generative model is just a part, albeit important, of this larger architecture. It is entirely possible that this scheme includes both representational and nonrepresentational aspects or parts. Even strong commitment to representationalism in PP does not have to entail a view on which *all* there is to cognitive processing is representation-munching. In addition to the generative model, PP comes with at least three other posits: (1) the sensory signal which results from the world affecting the sensory apparatus of an organism, (2) the prediction error signal which is propagated bottom-up, and (3) precision estimators which regulate the gain on the prediction error signal. For each of those posits, we may ask whether its functioning is representational in nature. Although a case could be made that precision estimators are representational (Wiese, 2016), the same may not apply to the sensory signal. The latter acts as a mere causal mediator incapable of representational error (Gładziejewski, 2017). And there is still a further question of whether the bottom-up error signals earn a representational reading (Orlandi, 2016 can be read as providing a negative verdict here). The point is that PP does not come with wholesale representationalism; there may be purely non-representational structures and processes involved in perception and action control.

Second, even on the strong reading of PP's representationalism, the representations in question are *anything but* action-neutral. Remember that considered in the context of FEP, the process of minimizing the prediction error is merely a way of achieving a pragmatic goal of keeping an organism

[3] That is unless, of course, one is committed to full-blown antirepresentationalism.

within conditions that help maintain it in a far-from-thermodynamic-equilibrium state. This is directly achieved through action, construed in PP as minimizing the prediction error by engaging reflex arcs to quash proprioceptive prediction error. And perception (perceptual inference) is there to provide guidance for action; estimating the causes of the sensory signal functions to enable adaptive engagement with environment. In other words, on the PP view of things, building a structural representation of environment is not an end in itself but a tool of self-maintenance (Williams, 2017). This is in line with those approaches in the literature that try to recast representationalism so that it becomes not an alternative but an ally to 4E approaches (Bickhard, 1999; Rosenberg & Anderson, 2004).

Third, the content of representations postulated by PP is organism-relative and shaped by the organism's embodiment. To see this, PP once again must be considered within the proper context of FEP. Given that perception is ultimately a tool for self-maintentance, the content of the internal models is naturally expected to be strategically selective (Burr & Jones, 2016; Clark, 2013, 2015; Williams, 2017). What is 'reconstruced' in internal models of prediction-error-minimizing-agents are those aspects of the environment which constitute the organism's Umwelt, i.e. the ones which the organism depends on in its practical engagements with the environment. Furthermore, given that one situation can be associated with different suprisals for different types of organisms (what has large surpirisal for a human phenotype may not be surprising for a cod phenotype), it is natural to hypothesize that the content of those models will differ from species to species (Williams, 2017). Also, the organism's body plays a non-trivial role in constraining the contents of generative models. To learn the causal structure of its surroundings, the prediction-error-minimizing-agent needs to intervene on the environment, where those interventions serve as 'experiments' that enable the system to disambiguate between alternative hypotheses. The body plays a crucial role here, as it serves as a reliable, readily-available 'laboratory' (Burr & Jones, 2016). The sort of statistical patterns most readily accessible and learnt are those that the depend on the bodily interactions with environment.

Fourth, consider the question of the vehicles of representations in PP. Here, of particular interest is how PP deals with the idea of detached

representations, that is representations used for off-line cognition instead of for perception or action control. On PP view of things, imagery, counterfactual reasoning, action planning or dreaming could be understood in terms of generative models run is simulation mode – in a way that is fully or partially freed from the "sensory enslavement" of direct interaction with the environment (see e.g. Hobson & Friston, 2012; Pezzulo, 2017; Seth, 2014). Simulations of this sort could generate a cascade of top-down sensory signals, activating levels relatively low within the hierarchy. This way, generative models could run simulations that span multiple levels of the processing hierarchy and bring about patterns of neural activity that resemble to those that accompany perception and action. If this is so, then representational vehicles underlying off-line cognition will not comprise of amodal, body-neutral neural code, but will rather involve neural machinery primarily involved in modality-specific (this includes interoception, see Seth, 2013) on-line cognition. This again connects nicely with what some proponents of the embodied approach have argued for (Barsalou, 1999; Goldman, 2012).

2. The commitment to inference

The second conservative commitment of PP relates to the notion of inference. The motivation for it stems from the idea of the external world as a sort of 'black box' for the skull-bound brain (Clark, 2013; Hohwy, 2013). On this story, to do its job as a controller of action, the brain needs to generate movements that accord with the layout of the organism's immediate surroundings. A real-life snake and a snake-looking cucumber mandate different reaction on part of the agent. However, all that the brain has direct access to are the effects that the external things impinge on the sensory apparatus of the organism. The input is ambiguous, as sensory states are underdetermined by the world: in many realistic circumstances, the sensory effects of a snake and a cucumber may be quite similar. Hence, the task of perception is to recover the *most likely* external causes of the sensory signal – out of a range of some alternatives – so that adaptive action can be initiated. This 'recovery' is construed in terms of an inference under uncertainty. The brain abductively 'infers' environmental causes of the sensory input, that is, it comes up with hypotheses that best explain (given

a larger model of the environment) the sensory patterns by citing their worldly causes. This, of course, places PP within a longer history of thinking about perception in terms of an abductive inference (Gregory, 1980; Helmholtz, 1860/1962).

The idea that PP is in fact committed to inferentialism about perception faces two sorts of criticism. On the one hand, it may be argued that the view presented above gets the 'epistemic' situation of the brain completely wrong. Perception is not underdetermined by sensory stimulation because all the required information is already present in the physical energies affecting the sensorium; and/or because the brain is, in virtue of its wiring, attuned to statistical patterns in the environment to the degree where no disambiguating inference is needed (Anderson, 2017; Orlandi, 2016, 2017). There is no motivation for postulating inference in the first place. I will not address this sort of criticism here, as it seems to be properly aimed not at the *conservative* reading of PP, but the whole PP framework itself. It arguably makes obsolete the very postulate that the brain implements a nesting, hierarchical model engaged in generating top-down sensory prediction. On the other hand, it may be argued that the notion of inference at play in PP it is either trivially liberal or misconstrues what the framework actually postulates (Bruineberg, Kiverstein & Rietveld, 2016). That is, the 'inferences' involved are not genuine inferences or the inferential approach is not justified by what the PP says about the machinery underlying perception, action, and cognition. To address this sort of criticism, I want to first elucidate what 'inference' as postulated in PP amounts to, and then proceed to show that it is neither excessively liberal nor does it get PP wrong (for a similar, in-depth defence of the inferential nature of PP and related computational models of perception, see Kiefer, 2017).

There are three crucial ingredients that make PP genuinely inferential. First, note that the commitment to inference is strongly tied to the commitment to representation. Given that inference constitutively involves transitions between *contentful* states, the former commitment presupposes the latter. In fact, it seems that to treat inference as postulated in PP literally, we should go with the stronger, realist brand of representationalism. Assuming strong representationalism, there *are*

transitions between genuinely contentful states in PP, as the internal hierarchical generative model is changing to keep track of the environment at different time-scales. This amounts to updating an action-guiding, detachable, error-detection-affording structural representation of the environment. Two general transitions involved are (1) revising the current estimate to match the current sensory input; (2) learning through perception, that is, revising the overall structure of the model ('priors') so that the prediction error is better minimized over longer periods. The model goes from one representational state to another by revising, adding, or dropping current hypotheses and long-standing beliefs.

Second, these representational transitions are approximately Bayesian without explicitly representing the Bayes rule. It is reasonable to hypothesize that a system that minimizes prediction error is a system that performs approximate Bayesian inference by maximizing the posterior probability of its model of the environment (see Hohwy, in print; Hohwy, Roepstorff & Friston, 2008; Kiefer & Hohwy, 2017). This means that a system updating its generative model to minimize prediction error is a system that updates its internal estimates of the environment in a way that conforms with Bayes rule. As such, given that Bayesian inference embodies a rational rule for revising one's beliefs or subjective probabilities, perception (and action, see Hohwy, in print) on PP view turns out to conform to a normative principle. Its rationality stems from the fact that Bayesian inference is truth-preserving (for a more detailed discussion, see Kiefer, 2017). And truth-preservation is another constitutive feature of inference.

Third, it seems that a kind of autonomy is implied in truly inferential processes. Suppose that there is succession of events A, B and C and that each of those events produces, in turn, an internal representational state A', B' and C' in some cognitive agent. Suppose that the move from A' to B' to C' conforms to some truth-preserving rule like modus tollens. Because of how the transition between the representational states is completely determined by external events, it does not seem to count as inference. Inference is constitutively an act, a part of agent's cognitive *activity*. Importantly, representational transitions involved in PP meet this criterion of inference. The way that perceptual hypotheses and priors are updated is not a matter of passively registering external states. Rather, it is co-shaped by the

internal states and dynamics of the prediction-error-minimizng system. The perceptual inference and perceptual learning are not completely determined by the driving, sensory signal, but actively shaped and constrained by the system's prior 'knowledge'. So, inference properly counts here as an active, not just reactive process.

Taken together, this amounts to a view of inference as an act of representational change that (approximately) conforms to a truth-preserving rule. *This*, and nothing more, is the sense in which conservative PP is committed to inference. Notably, there may be other considerations in favor of the claim that literal inference is involved in PP. For example, Kiefer (2017) argues that – in line with some influential treatments of inference in philosophical literature – representational transitions in PP (and related frameworks) are such that they increase the overall coherence of representations involved, that is, their consistency and the number of inferential connections between them. Another point might be that because the generative model reduces the prediction error relative to the sensory signal (as caused by the external world), the representational change can be also seen as maximizing the 'empirical adequacy' of the model. Nonetheless, it must be conceded that the sort of inferential processes postulated in PP also *lack* some of the features that characterize many paradigmatic instances of inference. In particular, they are not consciously accessible or goal-directed in the sense of being driven by personal-level intentions. But it is doubtful whether any of those features is *necessary* for a cognitive process to count as inference (see Kiefer, 2017).

As mentioned, the idea that full-blown inference is involved in PP can raise some skepticism. One reason for this stems from a close inspection of the way that the notion of inference is employed in the literature on FEP. As some authors point out (Bruineberg, Kiverstein & Rietveld, 2016), 'inference' as used in the work of Karl Friston (e.g. 2013) boils down to a dynamic coupling between the organism and its environment in which the mutual information between the internal (organismal) and external ('hidden') states is maximized. Because, almost by definition, every organism falls under FEP (to live is to actively avoid surprising and seek unsurprising states), every organism can count 'inferring' the states of the environment in this sense. Furthermore, this notion applies to non-living

coupled systems, for example, to a system composed of two coupled pendulum clocks (Bruineberg, Kiverstein & Rietveld, 2016) There clearly is something misleading about treating bacteria or synchronized clocks as engaged in *literal* inference. This very minimal, relaxed usage of the notion diverges from a more cognitivist sense that most associate with inferentialist view of perception. However, as mentioned at the outset of this paper, we need to be careful to distinguish between PP and FEP. This raises the possibility that the notion of inference at play in PP is different than the one sometimes used in discussions of FEP. And it seems that this is exactly the case. 'Inference' at use in PP is significantly stronger: it entails far more than the coupling of two dynamic systems. It involves an endogenously controlled transition between genuinely representational states that approximately conforms to a truth-preserving rule. Hence, the concerns about trivialization of the notion of inference which can be reasonably raised in the context of FEP do not apply to PP.

Another way to challenge the inferential reading of PP is by trying to show that the processes the framework postulates have features that prevent them from counting as truly inferential. In particular, some authors (Bruineberg, Kiverstein & Rietveld, 2016) point to the fact that traditional inferential theories of perception rely on an analogy between perception and scientific hypothesis testing. But this analogy collapses once we consider PP in the context of FEP. When properly construed, the job of the perceptual system is not to generate representations that 'objectively' capture the environment. Perception is a fundamentally biased sort of hypothesis-testing enterprise:

> If my brain really is a scientist, then it is heavily invested in ensuring the truth of a particular theory, which is the theory that "I am alive". This is a fundamental prior belief that drives all action; namely, I exist and I will gather all the evidence at hand to prove it. It will only make predictions whose confirmation is in line with this hypothesis. It does not give competing hypotheses a fair chance and is extremely biased in the way it interprets the data. It decides on the outcome of an experiment beforehand (my staying alive) and manipulates the experiment until the desired result is reached. If my brain is a scientist, it is a crooked and fraudulent scientist (…) (Bruineberg, Kiverstein & Rietveld, 2016, pp. 14-15).

One might feel tempted to use these considerations as an argument against the involvement of inference in PP. But this criticism would beg the question. Of course, according to PP, the perceptual system is *not* interested in truth for the sake of it. As mentioned before, it is selective in the way it recapitulates the structure of the environment. It is natural to expect that it changes its representational states in a way that is systematically biased toward the overarching aim of keeping the organism in unsurprising states, which sometimes means sacrificing truth or accuracy. Furthermore, it has been forcefully argued on PP view of things, action initiation is based on systematically *misrepresentational* precision estimations (Wiese, 2016). Yet, it is far from clear why the fact that the way the perceptual system works diverges from idealized norms of scientific rationality could prevent the system in question from counting as *inferential*. Because of social factors and cognitive biases, the way *scientists* update their hypotheses in light of evidence sometimes (perhaps often) deviates from idealized norms of scientific rationality. This hardly makes the updating process non-inferential. To generalize, crooked inference is inference nonetheless. And as I take it, conservative rendering of PP (charitably interpreted) is *only* committed to the idea of perception as inference, *not* to an importantly different and stronger claim that perceptual inference functions to uncover truth for the sake of it.

3. The commitment to internalism

The last commitment often associated with conservative construal of PP is to an internalist view of the mind. Here, 'internalism' means a claim that, contrary to extended and (strong incarnations of) embodied views, the constitutive basis for cognition does not go beyond the boundary of the central nervous system. This 'neurocentric' or 'seclusionist' reading of PP is defended by appealing to the notion of a Markov blanket (Hohwy, 2016, 2017, in print). The concept comes from causal network models and refers to nodes of the network such that, given some node X, the state of X is statistically fixed (can be fully predicted) by the states of those nodes. The Markov blanket of X will thus include its neighboring nodes: its 'parents' (proximal nodes that activate X), its 'children' (proximal nodes activated by X) and the parents of its children (Friston, 2013). Now, the point is that

internal sensory and 'active' (motor) states constitute a Markov blanket for a prediction-error-minimizing agent. Less technically, to fully predict how agent's internal states will evolve in time, all that is required is knowledge about its internal dynamics and what happens at the sensorimotor Markov blanket. Assuming that on the PP view of things cognition is prediction-error-minimization, the generative-model-based machinery involved in minimizing the error is situated within the Markov blanket thus construed. This way, the brain and spinal cord emerge as the sole seat of mindedness. Relatedly, this also opens up the possibility of skepticism, whereby an agent can enjoy a rich cognitive life even if it is being fed its sensory states not by the external world (nor does it output its active states to actual body), but rather by a misleading demon.

As noted by the opponents of the conservative reading of PP, this way of defending internalism in PP turns out problematic (Clark, 2016a, 2017; Fabry, 2017). One particularly forceful criticism points out that the concept of a Markov blanket is a technical notion that can be applied to any dynamical system to demarcate it from its environment (Clark, 2017). There will be Markov-blanketed systems *within* the prediction-error-minimizing agent, from single neurons to particular levels within the hierarchical generative model implemented in the brain. In addition, Clark argues that nothing prevents us from postulating Markov-blanketed systems that encompass the (embodied) brain *and* parts of the external, technological environment. That is, a system that comprises the biological agent equipped with technological extensions or interfaces could count as prediction-error-minimizing agent enclosed within a Markov blanket. In fact, Clark (2017) opts for a view that the boundaries of minds change 'metamorphically' through life as technological extensions are added and subtracted.

Assuming there is a nesting hierarchy of Markov-blanketed systems that go both within and outside the brain, natural questions arise. Which Markov blanket is the privileged one when it comes to delineating the mind? And why think that the boundary coincides with the blanket that secludes the central nervous system? In fact, these considerations leave us with three options regarding the idea of a Markov blanket as cognition- or mind-delineating boundary: (1) there is one, stable, unique blanket that delineates cognition and it is the blanket that surrounds the central nervous system;

(2) the boundaries of a cognitive system are enclosed by a Markov blanket that metamorphically changes to include factors that go beyond the central nervous system alone; (3) no Markov blanket serves as a unique, cognition-demarcating one. Only option (1) counts as genuinely conservative. However, the most important lesson is that the technical notion of the Markov blanket *as such* is not enough to decide between these three options (Clark, 2017). This means that the justification for internalist reading of PP, if it is to be found at all, presumably will not come from the conceptual resources of the framework itself.

Internalism turns out to constitute a soft underbelly of conservatism about PP, the one commitment that seems the least justified in light of the framework (for other arguments against the internalist reading of PP, see Clark, 2016a, 2017; Fabry, 2017). However, two things need to be pointed out before the conservatist admits defeat on this front. First, the internalist commitment is logically independent from the other two. Most importantly, neither representationalism nor inferentialism about PP presuppose the truth of internalism. There is nothing contradictory about the idea of a system that trades in representations and engages in inferences but whose boundaries do not coincide with the boundaries of the central nervous system. So even if we do drop the internalist commitment, the other two can remain intact, leaving us with what is still a recognizably (albeit weakly) conservative outlook on the nature of cognition. Second, even if internalism cannot be defended by pointing to the notion of a Markov blanket alone, there may be other, independent considerations in favor of internalism. In particular, it might be interesting to see how PP meshes with other, independent theoretical proposals that support delineating cognition in internalist, skull-bound way. A full, in-depth discussion of this subject is beyond the scope of this paper. However, let me briefly sketch out the connections beyond PP and some of the well-known, pro-internalist conceptions of where cognition ends.

i. PP and non-derived content

On one view, what distinguishes cognition from non-cognition is the fact that only the former involves processes that make use of *non-derived intentional content* (Adams & Aizawa, 2001, 2010). This is the content that

is intrinsic to the content-bearing state rather than derived from conventions or interpretative/explanatory practices. Note that when applied to PP, this approach would connect internalist commitment to the representational one. Because on the weak, pragmatist/instrumentalist reading of representationalism in PP, content is clearly *derived* (it depends in its existence on the explanatory practices of scientists), the connection would have to be with the *strong* branch of representationalism. The internal, resemblance-based, action-driving model that the strong reading of representations in PP appeals to seems like a good seat for non-derived content. The content of this model is based on the structural resemblance between the representational vehicle and some (represented) part of the environment, such that the degree to which the resemblance holds is causally relevant for the success of model-guided actions (Gładziejewski, 2016; Kiefer & Hohwy, 2017; Williams, 2017). Neither the structure of the vehicle, the structure of the represented state of affairs nor the resemblance relation itself are observer-dependent; this view of content is realist through and through (see also Gładziejewski & Miłkowski, 2017). Hence, it is reasonable to assume that content here is not of derived nature. Assuming further that the generative model that serves as representation turns out properly situated within the confines of the skull, we end up with an internalist view. The weakness of this proposal lies in the non-derived-content-based strategy of delineating cognition itself. By definitionally linking cognition with representational content, this criterion is hardly ecumenical towards 4E approaches. More importantly, it seems to deflate or trivialize representationalism by *a priori* precluding the truth of anti-representationalism about cognition (Ramsey, 2015).

ii. PP and cognitive systems

Another internalist way of demarcating cognition appeals to the notion of a cognitive system (Rupert, 2009). Roughly, 'cognitive systems' are physical systems that causally underlie collections of cognitive capacities and skills. These systems are integrated and persisting, and the collections of cognitive capacities and skills they give rise to are stable across different contexts. Because of their persisting and stable nature, it is cognitive systems that enable successful psychological or cognitive-scientific explanation by

making possible reliable generalizations about cognition. They give rise to stable patterns of cognitive behavior that can be studied under a wide range of independent experimental paradigms. The proposal is that only brains (or central nervous systems) count as 'cognitive systems' in this sense. For example, it is argued that 'extended' systems which comprise the (embodied) brain and parts of the environment are too ephemeral to afford successful, generalizable scientific inquiry (Rupert, 2009). Now, it might be hypothesized that the central nervous system *qua* prediction-error-minimizing mechanism counts as a cognitive system in this sense. It persists across different contexts and gives rise to cognitive phenomena. Furthermore, it might be argued that although there are extended prediction-error-minimizing systems enclosed by technology-based Markov blankets, these are not cognitive, as they are not stable enough to underlie successful scientific generalization. If this is true, it could rule out Clark's metamorphically extended predictive minds. The obvious problem, however, is that there are Markov blanketed, prediction-error-minimizing mechanisms *within* the central nervous systems. These may be even more stable and persisting error minimizing mechanisms within the agent. So, there remains something arbitrary about treating the peripheries of the central nervous system as the peripheries of cognition.

iii. PP and pseudo-closed-loop control

Grush (2003) defends internalist or 'Cartesian' demarcation of cognition by employing notions from control theory. To put Grush's sophisticated account in a nutshell, the idea is that brains count as sole seats of cognition because they are systems for which 'the world is not enough'. Due to the temporal delay that separates the sending of a motor command to the body and the sensory feedback resulting from the performed action, the brain is unable to perform motor control based on the feedback alone. Rather, it uses a pseudo-closed-loop architecture, where an efference copy of motor command is sent to an emulator, an internal structure that mimics the dynamics of the environment and the muscle-skeletal system. The sensory predictions endogenously derived from the emulator are essential, on Grush's account, for planning and fine-tuning ongoing movement. Furthermore, the emulator can be employed for purely off-line purposes,

like imagery. The upshot is that because of its reliance on an internal emulator, the brain emerges as 'potentially self-contained' – a system firmly distinct from the external environment and (given some other assumptions, see Grush 2003) a unique seat of cognition. Now, there is a recognizable kinship between emulation theory (and other efference-copy-based approaches to motor control) and PP (Dolega, 2017). Most importantly, note how perception and action are crucially guided in PP by endogenously generated, top-down sensory predictions. Obviously, because of the crucial role that the sensory input and error signals have in shaping the internal processing, the prediction-error-minizing system is far from being closed-off from the environment. This does not, however, diverge from Grush's original emulation framework, as, on his view, the sensory feedback constantly corrects the emulation-based predictions. Notice also how, in PP, when the precision of the sensory signal is predicted to be low (and so the sensory input's influence on hypothesis-revision is also low), or when the generative model is used purely off-line, the brain will appear as largely causally decoupled from the external environment. Because of those considerations, there is potential in PP to construe the brain (or the central nervous system) in Grushian way, as a largely self-contained seat of cognitive phenomena.

Conclusions

When seen within the proper context of the Free Energy Principle, minimizing the prediction error with the use of hierarchically structured generative models turns out to serve as a tool for self-organization. This strong pragmatic and organism-oriented spin on PP naturally invites an interpretation of the framework that is much closer to 4E approaches than to more orthodox, internalistic and intellectualist approaches in cognitive science. However, in the present paper I attempted to elucidate what 'conservative' reading of PP amounts to, hoping to show that this way of understanding PP is not ungrounded or completely alien to the spirit of the 4E approaches. I showed how PP is representational, both in a weak (pragmatic) and strong (realist) sense. Even on the strong reading, the representations postulated in PP are not just passive mirrors of nature, but action-guiding map-like structural representations that largely use

modality-specific vehicles and whose content is constrained by the way the organism is embodied and embedded in its environmental niche. Furthermore, I argued that the notion of inference that (the conservative rendering of) PP trades in is non-trivial, yet liberal. The inferential nature of perception amounts to the fact that the way the perceptual representations are (actively, not just reactively) updated conforms to a truth-preserving rule. There is no commitment here to an overly intellectualist claim that prediction-error-minimizing agents cognize in accordance to some inflated principle of rationality. Lastly, I attempted to show that whatever grounds there might be for treating PP in internalist terms, they are probably not to be found in the conceptual resources of the framework itself. However, I sketched out how PP might fit with some other, independent ways of delineating the mind in a skull-bound way. The resulting view is that PP is representational and inferential in what might be the most 4E-friendly way possible, and it does not have to be considered internalist (at least not on its own terms). Taken together, these considerations show that conservative reading of PP is well-grounded and not *that* conservative after all.

REFERENCES

Adams, F., Aizawa, K. (2001). The bounds of cognition. *Philosophical Psychology*, 14, 43–64.

Adams, F., Aizawa, K. (2010). *The Bounds of Cognition*. Oxford: Wiley-Blackwell.

Allen, M., Friston, K. J. (2016). From cognitivism to autopoiesis: towards a computational framework for the embodied mind. *Synthese*, doi: 10.1007/s11229-016-1288-5.

Anderson, M. L. (2017). Of Bayes and bullets. In T. Metzinger, W. Wiese. *Philosophy and Predictive Processing*. MIND Group, ISBN: 9783958573055.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 577-609.

Bickhard, M. H. (1999). Interaction and Representation. *Theory & Psychology*, 9, 435–458.

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

Burr, C., Jones, M. (2016). The body as laboratory: Prediction-error minimization, embodiment, and representation. *Philosophical Psychology*, 29, 586–600.

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

Clark, A. (2015). Radical predictive processing. *The Southern Journal of Philosophy*, 53, 3–27.

Clark, A. (2016a). Busting out: predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Nous*, doi: 10.1111/nous.12140.

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

Clark, A. (2017). How to knit your own Markov blanket : Resisting the second law with metamorphic minds. In: T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*, MIND Group, ISBN: 9783958573031.

Dolega, K. (2017). Moderate predictive processing. In T. Metzinger, W. Wiese (eds), *Philosophy and Predictive Processing*. MIND Group, ISBN: 9783958573116.

Downey, A. (2017). Predictive processing and the representation wars: a victory for the eliminativist (via fictionalism). *Synthese*, doi: 10.1007/s11229-017-1442-8.

Egan, F. (2010). Computational models : a modest role for content. *Studies in History and Philosophy of Science*, 41, 253–259.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, 170, 115–135.

Fabry, R. E. (2017). Transcending the evidentiary boundary: Prediction error minimization, embodied interaction, and explanatory pluralism. *Philosophical Psychology*, 30, 391–410.

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

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

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

Gładziejewski, P. (2015). Explaining cognitive phenomena with internal representations: A mechanistic perspective. *Studies in Logic, Grammar and Rhetoric*, 40, 63–90.

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

Gładziejewski, P. (2017). The evidence of the senses: A Predictive Processing-based take on the Sellarsian dilemma. In T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*, ISBN: 9783958573161.

Gładziejewski, P., Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors. *Biology and Philosophy*, 32, 337–355.

Goldman, A. I. (2012). A moderate approach to embodied cognitive science. *Review of Philosophy and Psychology*, 3, 71–88.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 290, 181–97.

Grush, R. (2003). In defense of some 'Cartesian' assumptions concerning the brain and its operation. *Biology and Philosophy*, 18, 53–93.

Helmholtz, H. (1860/1962). *Handbuch der Physiologischen Optik*. J P. C. Southall (ed), Vol. 3. New York: Dover.

Hobson, J. A., Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, 98, 82–98.

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

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

Hohwy, J. (2017). How to entrain your evil demon. In T. Metzinger, W. Wiese (eds). *Philosophy and Predictive Processing*, MIND Group, ISBN: 9783958573048.

Hohwy, J. (in print). The predictive processing hypothesis and 4e cognition. In A. Newen, L. Bruin, S. Gallagher (eds), *The Oxford Handbook of Cognition: Embodied, Embedded, Enactive and Extended.*

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

Hutto, D. D. (2017). Getting into predictive processing's great guessing game: Bootstrap heaven or hell? *Synthese*, doi: 10.1007/s11229-017-1385-0.

Kiefer, A. (2017). Literal perceptual inference. In T. Metzinger, W. Wiese (eds) *Philosophy and Predictive Processing*, ISBN: 9783958573185.

Kiefer, A., Hohwy, J. (2017). Content and misrepresentation in hierarchical generative models. *Synthese*, doi: 10.1007/s11229-017-1435-7.

Klein, C. (2016). What do predictive coders want? *Synthese*, doi: 10.1007/s11229-016-1250-6.

Orlandi, N. (2016). Bayesian perception is ecological perception. *Philosophical Topics*, 44, 255–278.

Orlandi, N. (2017). Predictive perceptual systems. *Synthese*, doi: 10.1007/s11229-017-1373-4.

Ramsey, W. (2015). Must cognition be representational? *Synthese*, doi: 10.1007/s11229-014-0644-6.

Pezzulo, G. (2017). Tracing the roots of cognition in predictive processing. In T. Metzinger, W. Wiese (eds), *Philosophy and Predictive*

Processing, MIND Group, ISBN: 9783958573215.

Rescorla, M. (2013). Bayesian perceptual psychology. In M. Matthen (ed.), *The Oxford Handbook of Philosophy of Perception* (694–716). Oxford University Press.

Rosenberg, D. G., Anderson, M. L. (2004). Content and action: The guidance theory of representation. *The Journal of Mind and Behaviour*, 29, 55–86.

Rupert, R. (2009). *Cognitive Systems and the Extended Mind*. Oxford: Oxford University Press.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17, 565–573.

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

Sprevak, M. (2013). Fictionalism about neural representations. *The Monist*, 96, 539–560.

Wiese, W. (2016). Action is enabled by systematic misrepresentations. *Erkenntnis*, doi: 10.1007/s10670-016-9867-x.

Wiese, W. (2017). What are the contents of representations in predictive processing? *Phenomenology and the Cognitive Sciences*, 16, 715–736.

Wiese, W., Metzinger, T. (2017). Vanilla predictive processing for philosophers: A primer on predictive processing. In T. Metzinger, W. Wiese (eds), *Philosophy and Predictive Processing*. MIND Group, ISBN: 9783958573024.

Williams, D. (2017). Predictive processing and the representation wars. *Minds and Machines*, doi: 10.1007/s11023-017-9441-6.

ABSTRACT

JUST HOW CONSERVATIVE IS CONSERVATIVE PREDICTIVE PROCESSING?

Predictive Processing (PP) framework construes perception and action (and perhaps other cognitive phenomena) as a matter of minimizing prediction error, i.e. the mismatch between the sensory input and sensory predictions generated by a hierarchically organized statistical model. There is a question of how PP fits into the debate between traditional, neurocentric and representation-heavy approaches in cognitive science and those approaches that see cognition as embodied, environmentally embedded, extended and (largely) representation-free. In the present paper, I aim to investigate and clarify the cognitivist or 'conservative' reading of PP. I argue that the conservative commitments of PP can be divided into three distinct categories: (1) representationalism, (2) inferentialism, and (3) internalism. I show how these commitments and their relations should be understood and argue for an interpretation of each that is both non-trivial and largely ecumenical towards the 4E literature. Conservative PP is as progressive as conservatism gets.

KEYWORDS: embodied cognition; enactivism; Free Energy Principle; inference; internalism; Predictive Processing; mental representation