Review of Philosophy and Psychology (2022) 13:947–972
https://doi.org/10.1007/s13164-021-00577-y

Cognitive Systems, Predictive Processing, and the Self

Robert D. Rupert [1]

Accepted: 21 July 2021/Published online: 3 August 2021

Abstract

This essay presents the conditional probability of co-contribution account of the individuation of cognitive systems (CPC) and argues that CPC provides an attractive basis for a theory of the cognitive self. The argument proceeds in a largely indirect way, by emphasizing empirical challenges faced by an approach that relies entirely on predictive processing (PP) mechanisms to ground a theory of the cognitive self. Given the challenges faced by PP-based approaches, we should prefer a theory of the cognitive self of the sort CPC offers, one that accommodates variety in the kinds of mechanism that, when integrated, constitute a cognitive system (and thus the cognitive self), to a theory according to which the cognitive self is composed of essentially one kind of thing, for instance, prediction-error minimization mechanisms. The final section focuses on one of the central functions of the cognitive self: to engage in conscious reasoning. It is argued that the phenomenon of conscious, deliberate reasoning poses an apparently insoluble problem for a PP-based view, one that seems to rest on a structural limitation of predictive-processing models. In a nutshell, conscious reasoning is a single-stream phenomenon, but, in order for PP to apply, two streams of activity must be involved, a prediction stream and an input stream. Thus, with regard to the question of the nature of the self, PP-based views must yield to an alternative approach, regardless of whether proponents of the predictive processing, as a comprehensive theory of cognition, can handle the various empirical challenges canvassed in the preceding section.

Keywords Cognitive systems · Predictive processing · Self · Extended cognition · Conscious thought · Extended mind

✉ Robert D. Rupert
robert.rupert@colorado.edu

[1] Department of Philosophy and Institute of Cognitive Science, University of Colorado, Boulder, CO, USA

1 Introduction and Overview

This essay plumps for a theory of the cognitive self. The theory rests on what I have elsewhere (Rupert 2019) called the 'conditional probability of co-contribution' (or CPC) account of the cognitive system (Rupert 2009, 2010). In what follows, I argue largely in an indirect way for this view of the cognitive self, by emphasizing empirical challenges faced by a competing approach, one that relies entirely on predictive-processing models and mechanisms to ground a theory of the cognitive self. In broad strokes, the argument runs as follows: given our current epistemic position, we should prefer a theory of the cognitive self of the sort CPC provides, one that accommodates variety in the kinds of mechanism that, when integrated, constitute a cognitive system,[1] to a theory according to which the cognitive self is composed of essentially one kind of thing, for instance, prediction-error minimization (PEM) mechanisms. Although it is apt to take a predictive processing based account of the self as foil, central threads of the argument generalize. To the extent that homogeneous views of cognition – according to which intelligent behavior is produced by a single kind of process or mechanism – face significant empirical challenges, we have reason to favor CPC, for it allows heterogeneity in mechanism and process.

The remainder of the paper consists of four sections. The first of these, Section II, motivates the discussion of cognitive systems, largely by situating the issue within the debate about extended cognition and extended mind. Section III presents CPC, a particular view of the human cognitive system and identifies the deliverances of CPC with the cognitive self. Section IV develops the line of reasoning briefly described above. As a comprehensive theory of cognition, the predictive processing based approach (PP, hereafter)[2] faces significant empirical challenges. Given the extent of these challenges, we would do better to embrace a more pluralistic and less empirically committed account of the nature of the cognitive self. The CPC offers such an account, delineating the cognitive self via a measure of integration that appeals to the degree of clustering of the mechanisms that contribute to the production of intelligent behavior, without any commitment to the nature of those mechanisms (PEM or otherwise). In the final section, it is argued that one of the core functions of the cognitive self – to engage in deliberate, conscious reasoning – poses an apparently insoluble problem for PP, one that seems to rest on a deep structural limitation of PP, as a mechanistic story of human cognition. Thus, with regard to the question of the nature of the self,[3] PP must yield to an alternative approach, regardless of whether PP can handle the empirical challenges canvassed in Section IV.

[1] Not a cognitive *sub*system, mind you. Questions about cognitive subsystems, such as a face recognition system or speech-parsing system, are a different matter, which I do not address here. (See Rupert 2019 for further discussion.) Rather, I'm interested in *the* cognitive system, writ large.

[2] Given limitations of space, I do not provide an introduction to PP or the core computational process associated with it, PEM. See Wiese and Metzinger (2017) for the relevant background.

[3] Or, what might be more carefully described as "processes normally associated with the self," to accommodate the possibility of eliminativism about the self.

2 Motivation for CPC

Why should one take interest in a theory of cognitive systems and their boundaries? One path runs via evaluation of the extended mind hypothesis, the claim that a significant proportion of human mental states are realized by, implemented in, or take the form of physical states appearing at least partly beyond the boundary of the human organism.[4] In Rupert (2004), I argued that resolution of the debate rests on an account of cognitive systems – that the only (or at least the most promising) way to resolve the debate about the extended mind is to focus, in the first instance, on the boundaries of the human cognitive system as a whole, rather than on the status of individual states or processes. I emphasize extended *cognition* rather than extended mind because, in philosophy, discussions of the mind tend to draw heavily on commonsense intuitions or on everyday ways of thinking and talking about mental states. Such discussion of the mind is, by my lights, too wedded to pretheoretic, folk perspectives, of the sort that scientific progress has tended to overturn or radically revise in other domains. Thus, I approach philosophical questions about the mind as questions in philosophy of cognitive science.

How, then, should we proceed? Throughout the history of cognitive science, it has been assumed that two contrasting kinds of cause contribute to the production of intelligent behavior, for example, the current state of the chessboard, on the one hand, and a computation-governed search for the next move to make, on the other. Moreover, it has been assumed that causes of the latter kind appear only inside the organism, while causes of the former kind appear both inside and outside the organism. Having set the stage in this relatively neutral way, the Hypothesis of Extended Cognition (HEC) can be stated thusly: many external contributors to the production of intelligent behavior are, surprisingly, of the same scientific kind (call it 'cognitive', if you like, but the label is not essential) as causes of the kind whose instances, in humans, were previously thought to be found only internally. Thus, the proponent of HEC should want to identify a scientific kind *(a)* of central importance to cognitive science and *(b)* that is shared by the paradigmatic internal states of interest, on the one hand, and the bio-external states and processes focused on by proponents of HEC, on the other.[5]

Proponents and skeptics alike should thus set out in search of a theoretically central distinction between two kinds of cause, one of which includes the paradigmatic states historically thought to be of the kind that occurs only within the organism. Among the possibilities, the distinction between "contributing from within the relatively integrated, relatively persisting system" and "contributing from outside of that system" stands out. It runs through all forms of successful cognitive modeling – from computationalist to connectionist to dynamicist to subsumption-based views to straightforwardly biological views (cf. Ross and Ladyman 2010). (This distinction is closely related to Margaret Wilson's distinction between facultative and obligate systems [2002, 630], roughly,

[4] Early philosophical statements of the extended view can be found in Clark and Chalmers 1998, Hurley 1998, Rowlands 1999, and Wilson 1994; the current century has seen an explosion of work on the topic – see Clark 2008, for a representative and influential example.

[5] This does not require that the such inner and outer states share a fine-grained causal profile; see Rupert (2004, section VII) for discussion of the possibility that the shared natural (or scientific) kind in question is a generic kind.

systems formed on the fly and those that are relatively persisting.) It is also reflected in computationalists' emphasis on cognitive architecture (Pylyshyn 1984).[6] If we are after a principled way to distinguish between two kinds of cause that contribute to the production of intelligent behavior, this would seem to be our best bet – to distinguish between causes that contribute from within the relatively persisting, relatively integrated system and those that contribute from beyond the boundary of that system. Thus, the identification of the boundary of the cognitive system becomes the central question in our attempt to evaluate HEC; if, for instance, the boundary of the cognitive system (typically) does not extend beyond the boundary of the organism, then HEC would appear to be false, because in that case no (or very few) causes of the cognitive sort – i.e., those that are part of the cognitive system – would fall beyond the boundary of the organism.

Various bits of the preceding reasoning might be, and have been, challenged; and much more can be said in their defense (for a recent treatment of the issues, see Rupert 2019). For present purposes, I leave the matter here, but I hope to have piqued the reader's interest in a theory of cognitive systems, by connecting it to one of the most important recent debates in philosophy of cognitive science.

3 The CPC

Cognition is a scientific kind, hypothesized by the relevant sciences to explain (what we take to be) a particular domain of phenomena, including such forms of human behavior as conversation in real time, patterns of similarity in the treatment of objects (e.g., reidentification and categorization), the production of works of art, the formulation and testing of scientific theories, the playing of chess, performance on reading comprehension exams, and so on. Whether these phenomena are ultimately of a piece – whether, in every instance, the explanation of the phenomenon in question appeals substantively to a kind of state or process appealed to in the explanation of all of the others – remains to be seen. This is the nature of scientific enquiry. But, our working hypothesis is that all of these phenomena deserve to be grouped together because they (or at least the lion's share of them) result distinctively from the activity of an integrated system of a certain sort, a cognitive system.

Can anything more precise be said about the integrated nature of the system, anything that sheds light on its role as a cognitive system, as a system that flexibly produces a wide range of forms of intelligent behavior? In previous publications (Rupert 2009, 2010, 2011, 2013), I develop the idea that a cognitive system consists of a collection of mechanisms that co-contribute in overlapping subsets to the production of a wide range of forms of intelligent behavior, as well as a mathematical measure meant to cash out the requirement "in overlapping subsets," thereby accounting for integration. The measure is location-neutral; it distinguishes between two kinds of causal contributor to the production of intelligent behavior – the contributors appearing

[6] Compare a point made, in a different context, by Gabriel Segal: "Whole subjects plus embedding environments do not make up integrated, computational systems . . . the whole subject is the largest acceptable candidate for the supervenience base because it is the largest integrated system available" (1991, 492).

in the cognitive system and those that do not, without any regard to whether the cognitive system straddles the boundary of the organism.[7]

Here, then, is CPC, now refined so as to clarify its structure and commitments. Bear in mind that, although the description to follow has a procedural flavor – as if it were a recipe for carrying out a construction – it is meant to reveal something significant about the property of cognitive integration itself. It is not a construction that any human is likely to (or is likely ever to have the resources to) carry out:

1. Take an organism at a given time, and form every non-singleton subset of the mechanisms that have distinctively causally contributed to the production of any form of intelligent behavior exhibited by that organism.

2. For each such subset, relative to each form of intelligent behavior, there is, for each of its proper subsets, a probability of its being a causal contributor to the production of that form of behavior conditional on every member of the complement of that set's contributing causally.

3. Rank order all such conditional probabilities.

4. Take the natural cut-off between the higher probabilities and lower ones. If something's being an integrated system is a natural kind (that is, a scientific property or kind), and the current proposal is on the right track, we should expect such a statistically significant gap to appear. Discard all entries below that gap.

5. For each mechanism appearing on the list of sets with higher conditional probabilities (that is, the sets above the gap referred to at Step 4), simply count the number of times that mechanism appears. Then, rank order individual mechanisms accordingly (that is, according to their number of appearances above the gap on the list produced by Step 4.).

6. A statistically significant gap separates those mechanisms that appear higher on this second list from those appearing lower on the list.[8] Discard the latter mechanisms.

[7] Independent arguments – for instance, from the existing successes of organism-oriented cognitive science – can be given for the claim that there is at least one cognitive system (not subsystem) appearing entirely within the boundary of the organism (Rupert 2009, 2010). Thus, if CPC is on the right track, we should expect the body-bound cognitive system – that is, whatever relatively persisting, relatively integrated cognitive architecture plays a role in successful cognitive modeling and appears inside the boundary of the organism – to satisfy the conditions laid down by CPC (and if it doesn't, that would seem to count against CPC). Because CPC itself is location-neutral, though, anything that satisfies CPC is certified as a cognitive system, and thus CPC itself does not beg the question against HEC; it allows that additional, possibly extended cognitive systems could be discovered, a discovery that might undermine appeals to simplicity and conservatism that are meant to show that, once we've established the presence of one cognitive system, within the organism, we should limit commitment to just that one.

[8] It's possible that multiple significant gaps appear – at Step 6 as well as at Step 4 – which muddies the waters. I contend that, if multiple significant gaps appear, then the collection of causes of intelligent behavior divides into three or more types, rather than two. But, I take HEC to concern causes of the type determined by the "narrowest band" – causes from the most tightly integrated cognitive system – which for the purposes of the debate at hand, yields a set of determinate questions to pursue. Weaker versions of HEC might refer to causes of intermediate status – those not below the lowest significant gap on Step 6's list but not above the highest gap. And, given that Step 6 builds on Step 4, possible kinds of cause multiply. Nevertheless, the radical cachet of HEC, which has attracted so much attention, derives from its strong version, according to which many states and processes beyond the boundary of the organism are of the same kind as those traditionally taken to be ur-cognitive states (which I expect would be among those that fall above the highest gap in Step 6's list, itself presupposing the importance of the highest gap on Step 4's list). Thanks to Luke Roelofs for pressing me to address these issues.

7. The integrated cognitive system comprises all and only those mechanisms left standing, that is, all and only those appearing above the gap on the second list.

Presented in this formal way, CPC's implications may remain obscure.[9] Consider an example, then. The typical subject is quite good at avoiding obstacles as she moves about, and if orthodox computational theories of vision are on the right track, a visual edge-detection mechanism has almost certainly causally contributed to such behavior. A mechanism that computes distance from retinal disparity will likewise have contributed to obstacle avoidance in the typical subject, as will have a mechanism that calculates shape from detected shading (Marr 1982). With regard to the avoidance of obstacles, many further mechanisms have contributed, for instance, various motor control mechanisms. To keep matters relatively simple, let us add only one such motor-control mechanism to the mixture of mechanisms under consideration. The resulting set of four mechanisms allows the possibility of six two-membered sets, four three-membered sets, and one four-membered set. For each two-membered set, two conditional probabilities are relevant: the first-mechanism's contributing conditional on the second's, and vice versa; this yields a total of twelve entries on the rank-ordered list constructed at CPC's Step 3. For each of the four three-membered sets, there are six relevant conditional probabilities: each single mechanism's contributing conditional on the other two's, and each combination of two's contributing conditional on the third's; this yields a total of twenty-four additional entries on the rank-ordered list constructed at CPC's Step 3. For the four-membered set, there are fourteen relevant conditional probabilities (which thus represent fourteen further entries to the rank-ordered list in question). For any one of the four, we must include the probability of its contributing conditional on the contribution of the remaining three, and vice versa, which yields eight entries. The remaining proper subsets of the four-membered set are pairs, as are the complements in all such cases. For any such pair, and there is a conditional probability of its contributing given that its complement pair is contributing. That yields six entries, which together with the eight from our lopsided divisions of the four-membered set, equals a total of fourteen entries contributed by the four-membered set. Relative to only this one kind of behavior and only these four elements, we already have fifty entries on the rank-ordered list associated with CPC's Step 3. Now go through this procedure – in principle! – for every grouping of all causally contributing mechanisms relative to each form of intelligent behavior that has been exhibited by the subject in question.

With regard to the example at hand, each of the four mechanisms will presumably appear in many subsets with high conditional probabilities (in the sense that the probability of a proper subset of a set's contributing will be high given that the complement of the set is contributing). This is a function of the mechanisms and the form of behavior chosen. For instance, one might reasonably think that the probability of the edge-detection mechanism's contributing given that the shape-from-shading mechanism is contributing is close to one; it would seem that every time the shape-from-shading mechanism contributes to the avoidance of obstacles, the edge-detection mechanism also contributes, at least for the typical subject, partly

[9] Integration may well be an irreducible property, in which case the measure of integration spelled out in the text should be thought of only as highly diagnostic, not as providing a set of necessary and sufficient conditions: for the typical subject, with a significant amount of worldly experience, CPC accurately delineates the cognitive system as it plays its distinctive role – as the producer of flexible, adaptive behavior. Alternatively, one might view integration as reducible, take CPC to reduce it successfully, and hold that the probabilities playing a role in CPC are propensities (rather than historically determined frequencies).

because, as we might say informally, they are both fundamental mechanisms of visual processing. Similarly for P(edge detection|shape-from-shading & distance from retinal disparity) and for P(distance from retinal disparity & edge detection|shape-from-shading). Notice, however, that sets including only the three visual mechanisms may well deliver higher conditional probabilities than sets that mix the motor-control mechanism with the visual mechanisms, particularly where the motor-control mechanism is being conditioned upon. It seems highly probable that if the visual mechanisms are guiding obstacle avoidance, then the motor-control mechanism is. But, perhaps the motor control mechanism also contributes to obstacle avoidance in cases in which, for example, the subject successfully navigates a familiar room in the dark, from memory, with little visual guidance. Thus, P(shape-from-shading|motor control) may be significantly lower than the conditional probabilities just considered. This will likely not be the case when the motor-control mechanism is being conditioned upon alongside a visual mechanism. For example, P(shape-from-shading|motor control & edge detection) is not likely to be any lower than conditional probabilities involving only our three visual mechanisms; for, if the motor-control mechanism in question is contributing along with the edge detection mechanism to obstacle avoidance, then we're almost certainly talking about visually guided obstacle avoidance, in which case shape-from-shading is almost certain to be contributing as well. As a result, consideration of our four mechanisms in connection with obstacle avoidance would presumably yield many subsets with high conditional probabilities (those that appear above the cut-off point at CPC's Step 4), even if the motor-control mechanism shows up in fewer than do the other three.

If cognition must occur within the boundaries of the cognitive system, as delineated by CPC, it would seem that for most individual human subjects at most times, cognitive processing occurs within the boundaries of the subject's body; for, generally speaking, the preceding characterization of the cognitive system cuts against the inclusion of special-purpose tools and one-offs, which tends to be the status of causal contributors beyond the boundary of the body. (A special purpose tool will likely appear in many sets with high conditional probabilities relative to a single form of intelligent behavior, but will not appear in such sets relative to other forms of intelligent behavior, putting that special-purpose mechanism at a significant disadvantage at Step 5, and thus 6, relative to mechanisms that contribute to a variety of forms of intelligent behavior.) The location of individual human cognition is largely an empirical matter, though. The systems-based proposal CPC leaves open the possibility that a tool – perhaps an iPhone (Chalmers 2008) – that consistently contributes to the production of a variety of forms of intelligent behavior across a variety of contexts, alongside a shifting set of co-collaborators that themselves have similar standing, is part of a human's cognitive system.

But why think CPC is correct? Flexible and adaptive behavior – that is, intelligent behavior – is the heart of cognition. This includes flexibility in knowledge acquisition, in the acquisition of concepts and skills, in problem-solving, and in the deployment of a variety of resources in the pursuit of and revision of goals in an oft-changing environment, among much else. It is this flexibility – and the accompanying high-degrees of social coordination and environmental modification as means to achieve a wide variety of goals – that attracts attention to certain forms human behavior and performance, and motivates the development of a distinctive science (cognitive science) to study them, in contrast to tropes and other stereotyped forms of behavior. It is the lack of such flexibility that drives

continuing complaints about extant forms of artificial intelligence. "It's not intelligence at all," one is tempted to say about such systems, "It wouldn't have any idea what to do if an unexpected situation were to arise! It does only that one thing!" – whether that one thing is playing chess, answering quiz-show questions, or controlling an automobile.

CPC is grounded in the idea that flexibility is achieved in human cognition by the presence of many units, circuits, and mechanisms poised to work together in various combinations. There's plentiful evidence that this sort of thing happens in the human brain (Anderson 2010, 2014; Cole et al. 2013; Botvinick and Cohen 2014). On some accounts of this kind of process, subnetworks with overlapping members wrest control from each other via competitive processing. In such cases, a shift in task doesn't require an entirely new network to take control from a previously dominant one; more subtle shifts in the co-activation of elements, some of which are already active, can more smoothly effect such a transition. The systems-based view CPC thus emphasizes what seems likely to be a central trait of a system capable of flexible, adaptive cognition – that any given mechanism is capable, via direct interaction or by co-contribution, of collaborating with various other mechanisms to drive various kinds of task performance.[10]

Of special importance in the current context is CPC's neutrality with respect to the nature of the mechanisms in question and the processes in which they participate. CPC is meant to capture what it is for a cognitive system to be integrated, or to provide at least an illuminating diagnostic measure of this core feature of cognitive systems. This high level of abstraction invites pluralism about the set of the possible mechanisms that compose any given cognitive system. Perhaps a cognitive system contains some mechanisms the contribution of which to task performance is best understood as connectionist. Perhaps others in the same system engage in logic-based inference. The contribution of still others might be best understood using the tools of PP. CPC leaves the matter open. Given our current epistemic situation, this seems desirable.

The CPC offers a theory of the integrated cognitive system, and, I maintain, thereby offers an account of the cognitive self. One might reasonably resist, however, holding that the self is only a proper part of the cognitive system – the part that constructs narratives (Schechtman 2007, 2011), and perhaps only to the extent that such narratives serve a certain purpose, for example, to smooth over social interactions (Dennett 1991). Or, perhaps the self is only the proper part of the cognitive system that manages the deployment of information (Metzinger 2009). I am not optimistic about such attempts to carve out distinct portions of the cognitive system, treating only those parts as *the* self. Many distinct self-models appear in the cognitive system, and they play a variety of computational roles (Flanagan 1994, Velleman 2005, de Vignemont 2018), working in tandem with various other parts of the cognitive system; thus, it seems quixotic to attempt to draw a principled boundary between the "self-y" part of the cognitive system and the rest of it. The behavior associated with the self is so varied and produced by a such wide variety of mechanisms, in overlapping subsets, that we do best to treat the entire cognitive system as the self, while recognizing that, with regard to a given

[10] Though speculative in some ways, this perspective helps to make sense of a large body of experimental evidence: of neural reuse (Anderson 2014); of competition-based task-switching (Yeung et al. 2006, Teskey and Masson 2017); of correlations between different degrees of network integration (across networks with overlapping nodes) and task performance (Shine et al. 2016); and of the extent to which task performance changes by degree as a result of changing degrees of priming, inhibition, or integration. The fleshing out of this line of argument must be left for another occasion.

phenomenon of interest, some parts of the cognitive system might be especially active contributors. Narrower conceptions of the self seem to ignore the breadth of the everyday conception of the self as the agent responsible for action. On the everyday conception, the self performs all manner of action: engaging in conversation, raising children, designing an experimental device, pitching a no-hitter in baseball, getting a university degree, cooking a meal, writing an academic paper, and so on. Upon successful completion of any one of these things, one might reasonably say "I did it!" But, the suite of cognitive capacities and mechanisms that contribute to this wide range of behavior seems to have no principled boundary short of the entire cognitive system; so far as I can tell, the entire collection of integrated cognitive mechanisms is at work across the range of cases – mechanisms involved in reasoning, creativity, the managing of social interactions, as well as the continuous use of various memory systems and language skills, not to mention motor ability and perceptual acuity. Only when one limits discussion to a specific sort of question – e.g., "What kind of person are you?" – is one tempted to think there's a distinct proper part of the cognitive system that counts as the self, or so I maintain.[11]

4 Unsolved Problems, Empirical Challenges

The epistemic remarks in support of CPC naturally raise questions about the empirical plausibility of PP, as a comprehensive theory of the mechanisms of human cognition. After all, CPC's neutrality doesn't count for much if PP has proven itself to be, or is plausibly on track to be, the correct view of human cognition. How promising is PP, then, as a comprehensive theory of cognition?[12]

[11] As one referee points out, these observations fall well short of a conclusive argument against all competing views of the self. I hope, however, they suffice to motivate the comparison to PP-based views. A careful comparison to a full range of other views of the cognitive self remains a project for another day.

[12] When objections are raised to PP's role as a comprehensive theory of cognition, advocates for PP often appeal the Free Energy Principle (FEP), as the ultimate, unifying ground, claiming that it is fundamental principle of all life, and perhaps much beyond. Examination of FEP would expand the scope of this essay unmanageably, and so I limit myself to a more traditional and circumscribed discussion of mechanisms responsible for intelligent behavior in humans. For some discussion of the strengths and weaknesses of an appeal to FEP's unifying power, in the context of theories of cognition, see Klein (2018) and Sims (2017). Of particular concern is that FEP is incredibly flexible and, at least as currently applied, would seem unfalsifiable; more to the point, it seems not to have generated specific process models that are homogeneously PP and can be used to account for the kinds of behavioral data (simultaneously constrained by neural data) that stand at the center of the cognitive-scientific enterprise. With reference to Sims's taxonomy, then, I take Maximal Predictive Processing (ibid., 10), rather than Free Energy Principle (ibid., 13), as CPC's foil.

In this context, it's worth considering the possibility that, as one referee's comments suggest, PP is not meant to provide an account of the boundaries of the cognitive system or the nature of the cognitive self. Instead of pitting CPC against Maximal PP, one might make a natural marriage of them, with CPC identifying the boundary of the cognitive system and PP identifying what goes on inside such a cognitive system. That possibility is certainly on the table, but at least in so far as PP is meant to ground process models – that is, models of the computational mechanisms that produce intelligent human behavior – the gist of this paper's main argument applies. It might turn out that CPC, once applied to the human case, delivers an integrated collection of mechanisms each of which is, in its role as part of the cognitive system, a PEM mechanism. But, there's some reason to think it will not turn out that way; for example, the mechanisms involved in reading comprehension might be accounted for, in some very abstract sense, by FEP models, but they may not have the character of PEM processes; one of CPC's primary virtues, relative to the current context, is that it can accommodate such an eventuality, irrespective of optimism on the part of the proponent of Maximal PP.

My goal in this section is not to banish PP from cognitive science, but rather to increase the plausibility of a pluralist hypothesis: that the production of intelligent behavior is best explained by the activity of a system constituted by a variety of kinds of mechanisms. To the extent that PP has difficulty handling the phenomena to be canvassed below, the pluralist hypothesis increases in plausibility; and to the extent that the pluralist hypothesis increases in plausibility, so does CPC, for the latter allows (though it does not mandate) a situation in which a variety of kinds of mechanism "band together" to form an integrated cognitive system. In this context, it's worth laying out some of the potential challenges to PP – even without making an effort at a complete review and evaluation of the literature on PP – because plausibility comes in degrees. If this were a binary matter, I might simply make the (relatively) uncontroversial observation that Maximal PP has not been confirmed empirically and leave matters there. But, that kind of argument gives us little sense of the scope, nature, or extent of the challenges in question, which are not binary matters fully captured by the observation that PP is, or is not, empirically confirmed.

I begin by noting some extant objection. Roskies and Wood (2017) worry about the ability of PP to account for "decision making, volitional action and long-term planning" (852), as well as creativity, free will, and "how we establish goals and life-plans" (856). They also express the concern that, if one distinguishes passive prediction (closely related to model-free or association-based learning, more on which below) from active prediction, the evidence seems to support significant presence of the former, to the detriment of claims to PP's comprehensiveness (855); for PP, in its concrete form, comprises only active prediction. Rescorla (2017) and Orlandi and Lee (2019) express the concern that some proponents of PP – Clark especially – fail to distinguish clearly enough between support for Bayesian views (of perception, for example) and particular implementations that deploy PEM; much of the evidence supporting the former, does not automatically support the latter. Orlandi and Lee also worry that, even some legitimate ways of representing PEM do not require that the PP architecture feed forward error signals only, but rather allow the feeding forward of visual features: "In particular, in a novel environment, at least initially, the visual system's priors will be neutral between many possibilities, and the bottom-up signal will do most of the work" (Orlandi and Lee 2019, 210), which runs contrary to what is supposed to be one of PP's most striking and novel theoretical claims. Williams (2020) challenges PP models to account for compositionality and the generality and productivity of thought. Klein (2018) worries particularly about the role of desires. And, Sims (2017) expresses concern about the ability of PP to account for playful and exploratory behavior. Clark (2019, 2020) responds to some of these objections, but his responses go only so far, painting the picture of a research program that rests on significant successes but also a large stack of promissory notes, particularly regarding so-called higher cognition. Note, too, that Clark and colleagues (Walsh et al. 2020) have recently made an extensive survey of cases in which support for PEM has been found, while at the same time making no bones about the upshot of their survey: that neural evidence in support of PEM

is decidedly mixed, with many experiments seeming to provide evidence contrary to Maximal PP.[13]

In what follows, I expand on this list of concerns, in some cases elaborating on existing ones and in other cases developing new entries or significantly different versions of concerns perhaps only gestured at thus far in the critical literature. It is not my purpose to review the state of the evidence for and against maximal PP; neither is it my goal to establish that Maximal PP is false or hopeless as a research program. Rather, I aim only to emphasize PP's empirical riskiness and thereby to increase the plausibility of the pluralistic hypothesis about cognitive mechanisms and, correspondingly, the plausibility of CPC:

4.1 Model-Free and Associative Learning

Consider the prevalence of model-free learning and the closely related phenomena of associative learning, Hebbian learning, and perceptual learning (Roskies and Wood 2017, 855; Sims 2017, 9). Associative mechanisms appear to play a robust role in human cognition but do not fall under PP's ambit. The essential PP scheme – the one that Maximal PP takes to exhaust the inventory of human cognitive mechanisms – presupposes a generative model that issues in predictions, which can lead to prediction errors and subsequent revision of the model. In cases of model-free and other forms of associative learning, generative models seem to play no role. Stimuli impinge on the cognitive system, their features are detected or represented, and various of them become associated with each other as a result of their co-occurrence.

Proponents of PP have, in response, argued that associative mechanisms can be incorporated into a PP-based account. One approach sets model-free learning within a grander scheme of trade-offs between model-based prediction and model-free learning (Clark 2016, 253–255). On this view, the system governing the trading-off itself selects the learning mechanisms it does – model-based or model-free – as a way of minimizing prediction error. Another PP-based approach treats what might appear to be model-free learning as, in fact, the activity of a generative model: a higher-level generative model itself represents that an association exists between, for instance, bell-ringing and the interoceptive stimulation that corresponds to salivation (to invoke a classic case), which is then physiologically instituted by active inference (Pezzulo et al. 2015).

These approaches may well pan out, but one should want some reassurance that the suggested accommodations are not ad hoc, that, for example, one salivates because one expects to salivate when one sees food, rather because of a historically built-up association between the sight of food and gustatory stimulation. Location of neural mechanisms that play the hypothesized roles (in control, for instance) would take us some distance in this direction, though that strategy is still a work in progress.

Keep in mind, too, the risk to Maximal PP of welcoming any significant amount of genuinely model-free learning into the fold, as Clark seems to, even when it is managed by a control structure the goal of which is PEM; for, the greater the extent to which one

[13] Note the distinction between the sorts of cases in question – in which imaging data is collected while subjects perform experimental tasks, which is then tested against predictions made by concrete PP models – and the sort of "proof of concept" discussions of PP, FEP, and neural processing that predominate in (Parr et al. 2019) and (Da Costa et al. 2020), where the authors are especially concerned to show that, for example, certain forms of message passing are consistent with some of what's known about neural processing.

includes non-PP mechanisms in one's conception of PP, the less plausible it becomes to claim that neural processing "performs homogeneous computations at all hierarchical levels" (Pezzulo et al. 2015, 32) or that PP is the "canonical computation" (Walsh et al. 2020, 257) in human cognitive processing.[14] In contrast, CPC easily takes on board a variety of potential components of the cognitive system, regardless of how they fit into a PP scheme and regardless of whether the cognitive system is, in fact, a thoroughly PP-system or merely one that can be seen as such when one focuses on certain aspects of its functioning.

4.2 Feature Representations Fed Forward and the Case of the Inveterately Bad Guesser

According to PP, only prediction-error signals are fed forward. In contrast, on a common non-PP view of perceptual processing, perceptual representations are generated by stepwise extraction of increasingly abstract features from the sensory signal, as it moves inward from the periphery. Vision science has produced a wealth of evidence in support of this feature-detection-based approach (Palmer 1999). How convincing is PP's alternative? Can feature-detection be cast aside so easily?

To be sure, it is sometimes claimed that, on the PEM scheme, error signals carry specific pieces of information by dint of the physiologically determined functional role of the specific neural structures or assemblies signaling errors: "Importantly, prediction errors are not regarded as general surprise or arousal signals but rather, the source, connectivity, and stimulus preferences of an error unit imbue its output with specific information about the nature of the mismatch between predicted and actual input" (Walsh et al. 2020, 243). This sounds suspiciously like feed-forward feature detection, though; as talk of representation is typically understood in cognitive neuroscience, a unit's "stimulus preference" just amounts to what the unit represents.

Imagine that a system proceeds through its life making incredibly noncommittal predictions or predictions that are very unlikely to be correct. As it encounters the environment, prediction signals with substantive content revise, or otherwise significantly parameterize, the relevant generative model. While this can all be fit into a PP framework, the PP aspect of the process is not terribly enlightening; instead it sounds as if what is at work is a standard flow of feature-representations being fed forward, with one exception: that the process is typically preceded by a bad (or at least unhelpfully indeterminate) guess. On that view, one can think of the incoming information as being

[14] Note, too, that in order to evaluate a claim to canonical status or the universality of a cognitive mechanism, one may have to attend to differences between explanatory levels (even when one is operating wholly at the subpersonal level, multiple explanatory levels within the subpersonal are relevant). It may be that a human-style use of modus tollens can be implemented by a connectionist network, but the question arises nonetheless whether such structures as activation-passing from one unit to another and summation of input to a unit explain human behavioral results or whether, instead, the performing of modus tollens explains the behavioral results. So, the question is not only whether a certain kind of computation is always present neurally across different problem-solving contexts; that might be so, even if the explanatorily relevant computational strategies or processes differ radically across cases. Thus, a substantive claim to universality, heterogeneity, or canonical status must go a step farther and show that the kind of computing relevant to the explanation of all manner of human performance – in reading comprehension tasks, mathematical problem-solving, the planning of vacations, the making of decisions about one's child's education, the writing of academic papers, and so on – manifests a single kind.

a correction of a bad guess; but that's largely incidental to the nature of the process. Rather, that process is driven by stimulation at the periphery – almost none of which is dampened by a prediction signal, given the badness of the guess – and operations on that stimulation, which deliver perceptual information in a standard feedforward fashion.

As noted above, Orlandi and Lee (2019) press Clark on this kind of point. In response, Clark claims that a robust pattern of model-based prediction and error-detection can be initiated by only the most vague of feature representations: "First, very general, extremely rapidly processed (low spatial frequency) features of the sensory input enable an initial guess at the rough gist of the scene — is it a natural scene, a face, animals, an industrial landscape?" (2019, 290–291). On Clark's view, this "initial guess" initiates a standard PEM process that satisfactorily accounts for the perceptual phenomena to follow. This response does not obviously put the concern to rest. The initial signals fed forward may not be semantically transparent – they may not correspond to features that we have natural-language designations for – but they do not seem to be mere error signals either, signals that simply say "the prediction was wrong, by this much" (along some unspecified dimension). Moreover, it constitutes a risky empirical bet on Clark's part that such signals suffice to get a more thoroughly PEM process off the ground, without those signals being processed in a stepwise, feed-forward fashion that extracts enough of the right kind of information to determine which high-level model to activate beyond some insignificant baseline (which gist to arrive at, in Clark's terms). Here one might also wonder about cases in which baseline is set well enough – I'm in the library in an English-speaking country – but the range of possibilities remains enormous, so large that any particular possible sensory experience is incredibly low. I can read any book I pick up, at whim, presumably by detecting the features of the text on the page in a feedforward manner. A vague signal indicating that what I see is English text (the combinatorial possibilities of which are enormous) does not seem to suffice to create determinate enough expectations to initiate a thoroughly PP-process, in contrast to one that is driven almost entirely by feature-detection. (And note that many of the sampling techniques developed by computer scientists to circumvent combinatorial explosion are not PP in character.)

4.3 Associative Learning and Features Fed Forward

To connect points 1 and 2, consider again the efforts of Pezzulo et al. (2015) to bring model-free learning under the umbrella of PP. On this view (ibid., 20–22), we characterize a canonical form of model-free learning – classical conditioning – as the activation of amodal representations in prefrontal cortex. These "central representations" (ibid., 20) guide the exploitation of relations between, say, sensory stimulation and interoceptive states.

How do central representations become active, though? In cases in which one is habituated to an unlikely event – perhaps bells being rung has a very low prior probability in one's environment – a representation of that feature, fed forward, activates the amodal representation of a bell ringing, in essence activating a mini-model of the relation between that stimulus and the responses associated with it. It appears that what needs to be fed forward is something like "*this* is what happened, instead of what you predicted," where '*this*' represents a feature detected – being the

ringing of a bell – in which case PP-based theorists seems to need to appeal to a story about feed-forward, feature-detecting processes that is not PP-based in any direct sense.

4.4 Pseudo-Conditionalization

In a similar vein, consider the problem of updating a model conditionally, for the purpose of hypothetical reasoning (so-called pseudo-conditionalization – Staffel 2019). A subject can take on a commitment – an outright belief – that she doesn't actually hold and reason from that supposed belief. The subject tentatively updates her commitments – including various conditional commitments, such as the acceptance of certain likelihoods – conditional on the supposed belief, which she has accepted only for the sake of argument or for hypothetical reasoning; moreover, and this is a point of special importance in much of what follows, the process seems particularly sensitive to the *content* and *logical structure* of the belief (or belief-like state) being supposed.

4.5 Model-Updating Processes

One should want to know about the process by which a model is updated in response to the information provided by prediction error. Of course, there's an extensive literature laying out how, at least in principle, such PP-based revision might work. But, it's a fair distance from there to successful models of human performance. Consider, for instance, cases in which updating is extensive and fast – for instance when extensive model change is caused immediately by linguistic input from a teacher – rather than being the effect of habituation. Individual subject's models are often radically revised by one-shot learning sensitive to the content and logical structure of what is said,[15] whereas typical algorithms sensitive to error signals account much more naturally for habituation and learning from large sets of data. (Compare the operation of backpropagation, a widely known error-responsive connectionist learning algorithm. It assigns more responsibility for error to connections that contributed more to the production of an erroneous output. But it's not obvious how this jibes with the way in which, for example, a geocentric thinker can, almost instantly, revise her model of the universe when she is informed of the heliocentric theory and the evidence for it.)

[15] This kind of processing is prima facie at odds with the standard brute-physical descriptions given in the PP literature of the contribution of error signals: "On most accounts of predictive processing...prediction errors drive representations in higher levels of the cortical hierarchy to provide better predictions – and thereby suppress prediction error signals in lower levels...prediction error signals also drive associative plasticity to update the generative model" (Barron, Auksztulewicz, and Friston 2020, 2). Driving associative plasticity isn't necessarily at odds with logical inference and content-based model-revision, but much work must be done to square the former and the latter in a way that preserves PP's explanatory primacy, in contrast to the role of logical inference or content-based construction of narratives. And, note that, throughout this discussion, talk of logical inference and content-sensitive processing is not tied specifically to conscious thought or personal-level processing. Plenty of cognitive processing is sensitive to logical structure yet is opaque to conscious reflection and fails various other tests meant to place a process at the personal level. In other words, the concerns expressed in the main text do not presuppose that Maximal PP is meant to be a theory of so-called personal-level cognition.

4.6 Context, Discourse Models, and Reading Comprehension

Proponents of PP frequently invoke the effect of context, where context is set by higher levels in the hierarchy of generative models. Consider one particular dimension to context-setting, the role of linguistic input and, perhaps more importantly, inferences from linguistic input (for, in many cases, linguistic input does not itself provide the contextual information, but only provides information from which contextualizing information is inferred). The point here is partly to suggest that PP-based approaches themselves rely, for their plausibility, on language-based context-setting; but my point is also to indicate the great extent to which language-processing requires the subject's maintenance of a model of the relevant discourse or text being read (Graesser et al. 2003, Rayner and Reichle 2010). Here we find a cluster of challenges concerning the maintenance of content-based coherence in the model, among, for instance, sets of propositions, and also concerning a variety of inferential processes, some required for the maintenance of coherence and some for the deployment of such models for further tasks (including context-setting).[16]

It's not as if no PP-based work has been done on the topic of linguistic communication (see, for example, Friston and Firth 2015), but PP-based work so far seems to capture only rudimentary aspects of linguistic exchange, not the cognitively demanding phenomena of content-based discourse-modeling or the drawing of inferences from such models, so as to set context or solve further problems, such as answering questions about a past conversation or, in the case of text consumption, answering questions on a reading-comprehension exam.

4.7 Cognitive Negotiation and Confirmational Holism

Models alone – in the sense in which one thinks of models in philosophy of science – do not make predictions, not even in the form of probability densities. When we take into account the need for auxiliary hypotheses and ancillary assumptions, in order to generate predictions, the classic Duhem-Quine concerns arise. And when error results, it can be unclear which aspect of this complex prediction-generating apparatus is to blame. I would emphasize that the decision what to change in these cases is *itself* a cognitive process. The data must be processed and reasoned about. A representation of what picture the data seem to present must be constructed (whether personally or subpersonally), and then the subject must decide, upon some mulling over (or cognitive negotiation) whether the data really does conflict with the model, or whether something else went wrong. That process, however, requires a feed-forward representation of the features of the data, a construction of a representation of the data and what they seem to entail. Prediction-error signals alone do not seem up to this task. What is needed is a period during which the data is represented neither as prediction-error nor as simply being "what was predicted," but rather "how things seem to be." During this period, the subject asks, "Does the way things present themselves as being actually conflict with the model in question or is the apparent mismatch the result of some other glitch?"

[16] The problem of one-shot word learning (Bloom 2000) presents a challenge as well, which might overlap with this one, to the extent that it involves thematic and content-based interpretation of the intention of other speakers in an environment.

Perhaps this process can be managed entirely by adjusting precision, but it would appear to require something cognitively richer. As part of this process, further bits of information are called up holistically (information from any cognitive quarter might be relevant), and such information is put to use to try to decide whether the new data can be made consistent with the model or whether it does, in fact, count against the model; and so on. This raises two concerns: first, one wonders whether PP has a way to handle the holistic nature of model revision and the interaction *among* generative models that such confirmational holism entails, and, second, one wonders whether this process can be modeled without the assumption of feedforward feature-analysis in the construction of how things seem or of what one is entertaining as a possibility based on the data, which must then be analyzed so as to make a judicious decision concerning model revision.[17]

4.8 Appearance of the Raw Materials and Relations between them

Mysterious, too, is the generation of the elements presupposed by PEM. How do subjects acquire new models? How do new ideas – flashes of insight in art, architecture, and science – come about? How are likelihoods determined, and how do they get revised, a particularly mysterious matter when those revisions cannot be anticipated (Paul 2014, Rupert 2016)? (Can a PP-based approach model Paul's transformation on the road to Damascus?) Many of these phenomena are difficult to account for on anyone's theory, but that fact does little to increase the plausibility of a purely PP-based account of them.[18]

Consider, too, the possibility that some of these processes, as well as other forms of reasoning, are driven by relations between models themselves applicable to different domains. Analogical reasoning, for instance, might play a strong role in the shaping of a model, by introducing into one model relations that are patterned after relations in a model of a different domain (Gentner 2003). Other forms of generalization and abstraction seem to play a role in learning and reasoning as well, in ways that might lead to the formation of or revision of a given model and that do not obviously yield to a PP-based treatment. Such

[17] Part of what emerges from the list of challenges raised in the text is a concern about a qualitative gap, between the widely reported concrete modeling successes of PEM – concerning, for instance, bistable perception and self-tickling – and phenomena that seem to involve deductive inference and content-based interpretation and reasoning. The latter sorts of case dogged connectionism's claims to a comprehensive and universal revolution in cognitive science (Marcus 2001) and continue to play a role in debates about machine learning and artificial intelligence (Marcus 2018). It is partly for this reason that it has been worth running through some of the potentially problematic cases, to show that Maximal PP's problem in modeling human data may well not be merely a matter of degree or just a matter of continuing to make incremental progress. Maximal PP seems to face something more akin to a problem of a limited domain of application. That large swaths of phenomena connected to semantic content and so-called higher cognition don't seem to fall within the scope of PEM models – as models of actual human performance – suggests a qualitative problem, and thereby reinforces a skepticism that increases CPC's standing appreciably.

[18] In the present dialectical context, it is largely beside the point that alternative views do not handle a phenomenon of interest, that the phenomenon is, so to speak, "everybody's problem." (Compare, for example, Clark's remarks about the fact that neither proponents of PP nor their detractors have a convincing story about the "origins of idiosyncratic desires" – Clark 2020, ms p. 8.) If a problem is everybody's problem, then it presents an empirical challenge to PP and thereby bolsters, to at least some degree, the prospects of the pluralist thesis, by increasing the likelihood that, when someone finds a solution, it will not harness exclusively PP-based mechanisms.

"model-melding" might just as well be a useful by-product of interference (because two different models share neural circuits that correspond to the computing of structurally similar operations in the two models) as it is the result of PEM.

4.9 Optimality Assumptions

Humans deviate from optimal performance in all manner of ways. The mechanisms responsible for cognition-related performance are messy and exhibit a variety of forms of limitation – from the ignoring of base-rates to various forms of interference in memory to failure at the Wason card task. Investigating such shortcomings may lead us to understand better the mechanisms responsible for producing human performance. It's possible that such investigation will reveal only glitchy implementations of PEM mechanisms, but it seems at least as likely that we'll discover otherwise: that the kludge-y interaction of a variety of not-entirely-PEM mechanisms produces the many and varied patterns of deviation from optimality as well as the positive extent to which human cognition approximates (e.g., Bayesian) optimality.

Compare the view of Griffiths et al. 2010., who emphasize Bayesian idealization and who hope ultimately to contribute to cognitive science's search for models of human performance: "Although cognitive modeling and machine learning are two different enterprises, a basic challenge for both is to match human-level performance in domains such as language, vision, and reasoning" (ibid., 363). And, Bayesian cognitive modelers match human performance by guiding the search for mechanisms that perform Bayesian inference "in a variety of implicit and approximate ways" (ibid., 362). More generally speaking, Griffiths et al. hope for a "synthesis with more bottom-up, mechanistically constrained approaches to modeling the mind" (ibid., 362). On this view, "Probabilistic models are a tool for exploring different sets of assumptions about representations and inductive biases, making it possible for data to lead us to an account of human cognition" (ibid., 363).

One must thus be circumspect when faced with fruitful, research-guiding idealizations. Even when we focus on PP's many extant successes, we should not neglect the sort of question that the comments from Griffiths et al. raise, which is whether an idealizing assumption plays only the role of an instrument facilitating the efficient search of hypothesis space – perhaps leading us to a mess of mechanisms that do not, in fact, fit very well the specifications of the idealizing assumption – or whether the success to which an idealization gives rise should be seen as evidence that the cognitive system is deeply of the sort that the idealization specifies.

Let me be clear about the dialectic. I do not claim to have proven that PP cannot handle the phenomena canvassed above, and I have not presented a thorough review of the literature in connection with the points raised. As many readers will be well aware, an enormous amount of impressive work has been done in the now established PP tradition. Nevertheless, such reviews and synthetic pieces as exist (Costa et al. 2020) leave many questions unanswered. They provide some grounds for optimism, based on in-principle results, but with regard to concrete PP-based models of human performance of the sorts flagged above, the reader is struck by the preliminary nature of the PP-based work that's reported. Thus, the challenges in question are genuine and, in light of them, optimism about Maximal PP requires no small leap of faith. For this reason, CPC's account of cognitive systems – and, thus, CPC's account of the self – carries with it the advantages of pluralism about cognitive mechanisms and processes.

Consider now an objection, that I have been examining matters at too fine a grain. Andy Clark writes:

> ...not every proper part of an integrated free-energy minimizing system (e.g. a cognitive agent) that does implement such an online prediction error minimizing process need itself be directly involved in that process...By the same token, a system that minimizes free energy using online prediction error minimizing techniques (a 'PEM system') could be part of a larger free energy minimizing whole that includes multiple sub-systems that do not work that way...The moral is that not every part of the full cognitive economy needs itself to display the full PEM profile. (Clark 2017, pp. 8-9).

Clark's point is *not* my pluralist point. Rather, his point would seem to be that, even if not every part of a system is best understood as using the tools of PP, it might still be best understood as a PP system; in Clark's terms, PEM might serve as "the all-purpose adhesive" (2016, 262) binding together various resources that contribute to problem-solving.[19] The idea seems to be that at the correct level of abstraction – when one tries to grasp what a system is really up to, what its ultimate purpose, goal, or operating principle is – the human cognitive system is best understood as a PP system, even if some parts of it are not directly involved in the minimization of prediction error.

I resist this idea, partly because it waters down Maximal PP to a great extent; it takes a claim about actual cognitive processes and mechanisms and recasts it as something more nebulous and interest-relative. In what follows, I articulate further concerns about this kind of response.

First, we should be wary of teleological thinking in cognitive science. The fundamental goal of cognitive science is to formulate and test models that account for human performance[20]; claims about the ultimate purpose of a given hybrid architecture or pluralist collection of mechanisms seem tangential to cognitive science's goal. Second, cognitive science is ultimately in search of mechanisms, broadly understood. This is the coin of the realm, because the discovery of mechanisms, implemented in biological wetware (or extended resources), integrates cognition and mind into the scientific world view; this is what qualifies cognitive science as a contributor to science's overall project of understanding the natural world. So, we should be suspicious of theoretical claims that separate themselves too much from the nature of the mechanisms in play,

[19] Importantly, for Clark, the pluralistic collection of problem-solving resources bound together by an overarching drive toward PEM can extend beyond the boundaries of the organism; if the logic of the inclusion of external resources is itself PP-based – that is, if it results from the drive to minimize prediction error – then the entire system should be seen as an extended cognitive system, one that is, by its nature, a PP system.

Clark claims that adopting the framework of PP alters little the contours of the debate about extended cognition (2016, 260). I tend to agree. In the PP-based context, the question naturally arises whether recruitment of external resources in order to help to minimize prediction error renders those external resources genuinely cognitive. I would argue that recruitment can do so only by creating a new cognitive system (or expanding the current one). And, if CPC, or something like it, represents our best account of the cognitive system, then much of what gets recruited by the cognitive system does not thereby become part of a cognitive system, regardless of whether the rationale for that recruitment is PP-based.

[20] What questions should a theory of problem solving answer?" Newell, Shaw, and Simon asked at the dawn of cognitive science; and responding to their own question, "First, it should predict the performance of a problem solver handling specified tasks" (Newell, Shaw, and Simon 1958, 151).

without some explicit justification (such justification as does exist in the case of CPC, that CPC tracks a distinction of causal-explanatory importance in all successful styles of cognitive-scientific modeling). Third, to the extent that it is helpful to think in terms of function or optimal performance, often it is because doing so guides the discovery of mechanisms or the best models of performance (see 4.9 above). In such cases, thinking in terms of function, optimality, or ultimate purpose plays a merely epistemic role in cognitive science (and in science more generally); such thinking provides a guide to the discovery of the actual processes that produce the phenomena in question, rather than a picture of the reality so discovered. Fourth, we should beware of claims to comprehensive-ness that sound bold, but are ultimately too weak to be of much interest. Any physical system or process *can* be modeled as a dynamical system (Wheeler 2005), and a vast array of physical system can be seen as computational systems (Putnam 1988). The mere fact that a system can be understood, at a certain level of abstraction, as a Phi system does not entail that the ultimate truth about the system is that it's a Phi system. So, even if there's a way to see all of cognition as PP-oriented, this does not automatically yield a substantive version of Maximal PP. Fifth, if we take questions about ultimate purpose to bear on questions of comprehensiveness, why not ask what prediction-error minimization is in the service of? Why not think, for instance, that the unifying account of cognition is that cognition builds models, which is not inherently a PP account of cognition, even if accurate model-building in humans is largely (or even entirely) served by PEM? At an even more abstract level, perhaps the point of cognition is to have an accurate account of the world (that will be good for all sorts of purposes, many of them unanticipated), encoded in generative models or otherwise. It seems arbitrary to claim that the single, ultimate purpose of cognition is to construct something with regard to which one can minimize prediction error, given that these other goals or interests make just as much (or more) sense.

Finally, we might note the complex relations between what are sometimes thought of as levels in the philosophy of cognitive science. In a recent paper, Clark (2020) repeatedly emphasizes the distinction between the personal level and the level at which the PP-story holds, which he characterizes as the subpersonal level. I myself have argued against the importance of a distinction between the personal and subpersonal levels (Rupert 2015, 2018); I propose that philosophers of cognitive science leave the distinction behind. But, most philosophers of mind and cognitive science embrace it. Taking the distinction on board, then, we might argue in the style of Fodor and Pylyshyn's critique of connectionism (Fodor and Pylyshyn 1988). Imagine that traditional, cognitivist models of such phenomena as reading comprehension and holistic evaluation of evidence emerge as the clear winners at the personal level. If PP-based mechanisms merely realize or implement such cognitivist models, it would be misleading to describe cognition as ultimately or fundamentally PP-based. This point is reinforced if, at the implementational level – that is, the subpersonal level – the PP-based account itself draws on mechanisms such as model-free learning that are not particularly PP-like in their operation.[21]

[21] Note well, however, that questions about the correct level of explanation – and the correct form of explanation at that level (cognitivist, connectionist, PP, or whatever) – arise even if we limit ourselves to a discussion of the subpersonal. See note 15.

In summary, many central aspects of cognition have yet to yield to PP, and at present we do not have compelling reason to believe they will. Instead, it seems eminently reasonable to think that the cognitive system is a mixture of kinds of cognitive mechanisms, and thus we should want a theory of cognitive systems – and of the cognitive self – such as CPC.

5 A Puzzle about Predictive Processing and Conscious Reasoning

Is there a fallback position for the proponent of a PP-based account of the self, perhaps a positive PP-based story that can be dissociated from Maximal PP? The proponent of a PP-based view of the self might, on this approach, concede CPC as a theory of the cognitive system, but might argue, nevertheless, that the *self* is a proper part of the cognitive system as a whole, a part that is distinctively PP in nature.

In this section, I argue on relatively general grounds that the strategy leads to a dead end. A fundamental structural commitment of PEM seems to stand in PP's way, preventing PP from accounting for a phenomenon central to our conception of the self – deliberate, conscious reasoning. Thus, a PP-based theory of the self, whether as part of Maximal PP or as a more narrowly focused PP theory of the self, does not appear to be in the cards.

Consider a recent PP-based account of the self, put forward by Jacob Hohwy and John Michael (2017) (H&M, hereafter). PP-based accounts appeal to the idea of a generative model, which plays the fundamental role of passing its predictions to a lower-level in PP's inferential hierarchy. Such models identify latent causes, for example, causes of sensory or interoceptive signals. The self might be the thing that a high-level generative model represents as the latent cause of a certain inputs or relations between inputs: "Our proposal is to conceive of this internal model of endogenous causes as a representation of *the self*. The suggestion then, is that agents model the self as a hierarchy of hidden, endogenous causes, and further, that the self is identical to these causes" (ibid., 369).[22]

Moreover, there's a reflexive aspect to H&M's proposal. According to H&M, part of the overarching generative model (or family of models) is, itself, the self. This part of the model represents various things that are responsible for certain patterns in the input. One of those things is a part of the model itself, the part that, via active inference, drives the relevant patterns in the input, by driving action that leads to the sensory and interoceptive stimulation that needs to be accounted for. In H&M's words, "...the part of the model that is involved in active inference is the self: this part of the model (the active states and their more deeply hidden causes) are the very endogenous causes that can be inferred in perceptual inference, which therefore become part of the self-model that in turn, in a dynamic downstream manner, shape active inference" (H&M, 375).[23]

[22] The body is also represented by (part of) a generative model, but it is not represented as identical to the self: "What we label the 'self' is constituted by more deeply hidden causes than what is represented in the body-model specifically" (H&M, 371).

[23] A central aspect of H&M's contribution, though tangential to present purposes, focuses on the social aspect of this process, the way in which inference in social contexts – regarding other minds as latent causes, as well as what one infers from what one is told about minds in general – can influence one's model of oneself, via social feedback loops.

Difficulties arise when we ask about misrepresentations of oneself, which H&M want their approach to allow for:

> Finally, the account is better positioned than narrative accounts to explain how we can be wrong in our self-representations. The self is not merely the fictitious subject of a narrative. Instead, it is the set of endogenous causes being referred to by self-models, which are constrained by their embeddedness in a positive feedback loop constituted by worldly causes, bodily states, sensory states, internal states at various levels of causal depth, and active states (H&M, 385).

I find this remark puzzling. Missing from H&M's account is a detailed story of the kind of misrepresentation in play, what, in particular, the *relata* are in the (mis)representing relation. H&M's account sometimes sounds as if it's an endorsement of straightforward self-reference: part of the self-model just is the self, so the self-model refers to the self-model. And, straightforward self-reference – "I am me" – does not seem to make room for the two distinct relata necessary for misrepresentation (necessary so that one of the relata can incorrectly describe or depict the other – here setting aside complications to do with the essential indexical, which are not relevant to my concern). Thus, H&M seem to owe us a more nuanced account of how the thing that causes action in active inference can come apart from one's model of the cause of the portion of one's sensory and interoceptive input that is identified as the self.[24]

Perhaps H&M mean to distinguish between the fully fleshed-out model of the self and the part of the model that contributes to active inference. Perhaps the model's general long-term expectations are sometimes adjusted based on a mistaken application of a learning algorithm, so that the subject comes to misrepresent itself post hoc. For example, the self might develop a narrative that rests on an erroneous parsing of past causes of its behavior. The parsing can be erroneous because there is a fact of the matter concerning which part of the model did, at a past time, contribute to the action that created input that then needed to be modeled; and perhaps, in response to the input, the model was adjusted in ways that mistakenly describe the portion of the past model that caused the action that led to the input in question. Even if this can be made to work on H&M's behalf, we should also want to accommodate cases in which a subject right now represents herself as, for instance, generous, even though her current generative model is not disposed to cause generous behavior. And, it remains unclear how H&M can allow such misrepresentation, given that, on their view, a substructure in the current model is identical to the very thing with the current dispositions to produce behavior via

[24] Wiese (2017) has argued that, in active inference, the execution of action relies on misrepresentation (of, for example, where one's hand is). These are misrepresentations of one's body, however, not misrepresentations of one's self, for H&M (see note 22). In other words, Wiese shows how one's self can engage in misrepresentation (of one's bodily activity or location), but without showing that the self misrepresents the self (or the self-model). Perhaps, though, H&M might hold that the self-model represents itself as *accurately representing the body*. This would be a misrepresentation of the self by saying, in effect, "I myself am the cause of such-and-such signals, and those signals are brought about by a process in the self that accurately represents the activity or location of the body." That, however, introduces only a kind of relational misrepresentation of the self, that is, by misrepresenting how the self is related to the things it represents beyond the self, e.g., one's hands.

active inference. If that part of the model represents itself as generous, then the subject will, after all, perform generous acts, because that part of the model is what produces action, and it does so in virtue of how it represents itself. So, H&M's model seems to preclude one's misrepresenting one's current traits.

This suffices, I think, to introduce a certain oddness about PP-based views of self-oriented cognition, including self-reflection. That oddness, as I see it, results partly from a structural problem. It would seem that PP lacks the right form to account for a central aspect of self-related cognitive processing: conscious, deliberate reasoning, the kind of thing one does when one sits down to think through a decision carefully, but which also occurs throughout the day – while driving, showering, riding the bus, and so on. In a nutshell, the problem is this: reverie is single-stream, but PP processing is inherently dual-stream, involving both prediction and input. Conscious, deliberate thought is simply a flow of logically connected ideas (or merely associated ones, or some combination of the two). It's decidedly not the generation of a prediction, even in a loose sense, to be measured against and corrected by an incoming stream; there's nothing – no second-stream – for it to be a prediction about, mismatch with which might generate an error signal. An essential aspect of PP models is the way a generative model is improved by the effects of prediction error, ultimately generated by a mismatch between predictions – passed through the hierarchy to the sensory level – and sensory input. But how can the structure of the model apply when there is no sensory (including interoceptive) input that might play such a corrective role?

Perhaps, in the spirit of the model of associative learning proposed by Pezzulo et al. (2015), the proponent of PP might pursue the following approach. Imagine one has a generative model that predicts that one will have certain internal auditory stimulation (in the "voice in one's head," which is often active during conscious reasoning) and predicts that such stimulation will be followed by certain actions. Here I have in mind a high-level model that is part of the subject's cognitive resources and that predicts the relation between the experienced voice in the subject's reverie at a given time and actions that will occur at later times. This creates the possibility of prediction error, because the actions in question might not be performed, and in this manner, the model that produces the complex expectation "internal speech + proprioceptive feedback" might be altered.

Although there's conceptual room for such an approach, it seems problematic for at least two reasons. First, it assigns a dubious role to the internal auditory signal. That signal is itself a product of active inference, and thus the auditory stream doesn't seem central to an account of deliberation itself, but rather of something downstream from it. A model might make a prediction of the sort "I will say P in my head," but in order to make that prediction it has to have already done the reasoning that "P" is meant to constitute or be the conclusion of. So, the thinking itself doesn't have anything obviously to do with PEM; rather, it's the shuffling around of structures in one's model of the world, for instance, in one's set of beliefs – and that would seem to be an entirely orthodox, non-PP account of deliberative thought, proceeding by, for instance,

deductive inference and then, after the fact, giving rise to a prediction of what the voice in one's head will say.[25]

Second, note that the results of an internal thought process have an infinite number of possible consequences, depending on what else happens – what one thinks later, what new evidence comes into one's possession, how other people behave, and so on. The computational demands involved in the present PP-based proposal – which would seem to require representing all of these contingencies – are overwhelming, though. It is also worth remarking that not all conscious thought is associated with an internal sensory stream, which seems to undermine the PP-based proposal in question.[26]

In the essay, I hope to have sown seeds of doubt about Maximal PP and thereby to have indirectly put the merits of CPC on display – in particular, its easy accommodation of pluralism about cognitive mechanisms and processes. Bear in mind, too, the potential of the present strategy to generalize. Wherever one finds support for hybrid cognitive models (Jilk et al. 2008) or reason to doubt one cognitive-scientific research program's claim to comprehensiveness, one also finds an argument in support of CPC

[25] Much of what is said in this section explicitly addresses questions about conscious reasoning. That is, of course, no surprise given the theme of this special issue; moreover, this framing matches well enough the way in which Hohwy and Michael present their PP-based account of the self. At the same time, the focus on conscious reasoning might seem unfair in the context of a comparison to CPC, the development of which is largely motivated by a focus on what are typically thought of as subpersonal states and processes. Such concerns about fairness are easily turned. Let's take CPC and PP both to be pitched at the subpersonal level (which is how many proponents of PP would characterize their work). We should still ask about the subpersonal account of personal-level phenomena, such as conscious, deliberative thought, for the latter are grounded in the former. CPC can treat the stream of conscious experience as the reflected image of a stream of subpersonal states, with the conscious nature of the former states explained – to the extent that subpersonal processes ever explain so-called personal-level phenomena – by the relation of their subpersonal counterparts to other subpersonal states, where that relation could, in principle, take any of a wide variety of possible forms. Compare: The global workspace is a subpersonal functional module – at least that seems to be the appropriate placement of it if one is playing along with the idea that the personal-subpersonal distinction serves a purpose in philosophy of cognitive science. It's possible, then, that a state is conscious if and only if it is the personal-level analogue to a state in the global workspace, and such states at the personal-level are conscious precisely because of the relation between their corresponding subpersonal states and other subpersonal states, such as motor commands or states in early visual processing (relations the holding of which is what a state's being in the global workspace amounts to). In contrast, PP-based explanation places severe constraints on the allowable subpersonal explanatory resources. If one's explanation of the personal-level conscious experience (or phenomenon) of deliberative reasoning is to be a distinctively PP-based explanation, there must be something distinctively PP about the process or structure – at the subpersonal level – that accounts for (or makes intelligible, or explains) the personal-level experience (or phenomenon) of conscious reasoning (because there's nothing distinctively PP about the personal-level phenomenon). I contend that the arguments pursued in the main text, appropriately modified to focus explicitly on the subpersonal context, cast doubt on such an approach.

[26] Another possible approach would exploit the oft-made (though perhaps dubious) distinction between System 1 and System 2 (Evans and Frankish 2009), arguing that processing in the style of System 1 rests on a model that can be tuned by processes occurring in System 2. On this view, both System 1 and System 2 produce streams of internal speech, one of which predicts the other. But, which one is prediction and which determines the error signal? Is one stream meant to represent the "real" self"? And, doesn't this approach require that both streams reason independently about the problem under consideration, prior to each stream's production of a signal suitable for comparison? Perhaps what's needed here is something like a balancing act between the two sources, determining on any given occasion which of the two sources acts as input and which as prediction. But one should ask whether the resulting picture doesn't violate the spirit of PP-based theorizing. This is a picture of two models, fine tuning each other, rather than a model being brought into alignment with some further reality meant to be the target of that model.

as a theory of the cognitive system, thereby also bolstering CPC's claim to account for the self.

Finally, does pluralism extend to CPC itself? Isn't CPC empirically risky? Yes, but we know that the assumption of some sort of relatively integrated, relatively persisting system has been a source of success across a wide range of research programs in cognitive science. This recommends an inference to the best explanation: that this assumption appears in virtually all successful cognitive-scientific models is best explained by its tracking something genuine, an important target kind *integrated cognitive system*. CPC represents a reasonable attempt to characterize this kind in a rigorous way and in a way that accommodates the plausibility of pluralism about mechanisms and processes. But, just as we should continue to explore the possibilities presented by PP, we should also explore other ways to characterize integration.

References

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

Anderson, Michael L. 2014. *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.

Barron, Helen C., Ryszard Auksztulewicz, and Karl Friston. 2020. Prediction and memory: A predictive coding account. *Progress in Neurobiology* 192: 1–13.

Bloom, Paul. 2000. *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Botvinick, Matthew M., and Jonathan D. Cohen. 2014. The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science* 38: 1249–1285.

Chalmers, David. 2008. Foreword to Andy Clark's *supersizing the mind* (see Clark [2008]).

Clark, Andy. 2008. *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford: Oxford University Press.

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

Clark, Andy. 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*: 3. Frankfurt am Main: MIND group. Doi: https://doi.org/10.15502/9783958573031 .

Clark, Andy. 2019. Replies to critics: In search of the embodied, extended, enactive, predictive (eee-p) mind. In *Andy Clark and his critics*, ed. M. Colombo, E. Irvine, and M. Stapleton. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190662813.003.0020.

Clark, Andy. 2020. Beyond desire? Agency, choice, and the predictive mind. *Australasian Journal of Philosophy* 98 (1): 1–15. https://doi.org/10.1080/00048402.2019.1602661.

Clark, Andy, and David Chalmers. 1998. The extended mind. *Analysis* 58: 7–19.

Cole, Michael W., Jeremy R. Reynolds, Jonathan D. Power, Greg Repovs, Alan Anticevic, and Todd S. Braver. 2013. Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience* 16 (9): 1348–1355.

Costa, Da, Thomas Parr Lancelot, Noor Sajid, Sebastijan Veselic, Victorita Neacsu, and Karl Friston. 2020. Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology* 99 (102447): 1–24.

de Vignemont, Frédérique. 2018. *Mind the body: An exploration of bodily self-awareness*. Oxford: Oxford University Press.

Dennett, Daniel C. 1991. *Consciousness explained*. Boston, MA: Little, Brown and Company.

Evans, Jonathan St, B. T, and Keith Frankish. 2009. *In two minds: Dual processes and beyond*. Oxford: Oxford University Press.

Flanagan, Owen. 1994. Multiple identity, character transformation, and self-reclamation. In *Philosophical psychology*, ed. G. Graham and G.L. Stephens, 135–162. Cambridge, MA: MIT Press.

Fodor, Jerry, and Zenon Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28: 3–71.

Friston, Karl J., and Christopher D. Firth. 2015. Active inference, communication and hermeneutics. *Cortex* 68: 129–143.

Gentner, Dierdre. 2003. Why we're so smart. In *Language in mind: Advances in the study of language and thought*, ed. D. Gentner and S. Goldin-Meadow, 195–235. Cambridge, MA: MIT Press.

Graesser, Arthur C., Morton Ann Gernsbacher, and Susan R. Goldman. 2003. *Handbook of discourse processes*. Mahwah, NJ: Lawrence Erlbaum Associates.

Griffiths, Thomas L., Nick Chater, Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. 2010. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences* 14: 357–364.

Hohwy, Jacob, and John Michael. 2017. Why should any body have a self? In *The subject's matter: Self-consciousness and the body*, ed. F. de Vignemont and A.J.T. Alsmith, 363–391. Cambridge, MA: MIT Press.

Hurley, Susan L. 1998. Vehicles, contents, conceptual structure, and externalism. *Analysis* 58 (1): 1–6.

Jilk, David J., Christian Lebiere, Randall C. O'Reilly, and John R. Anderson. 2008. SAL: An explicitly pluralistic cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence* 20 (3): 197–218.

Klein, Colin. 2018. What do predictive coders want? *Synthese* 195: 2541–2557.

Marcus, Gary. 2001. *The algebraic mind*. Cambridge, MA: MIT Press.

Marcus, Gary. 2018. Deep learning: A critical appraisal. *arXiv:1801.00631v1*.

Marr, David. 1982. *Vision*. New York: W. H. Freeman and Company.

Metzinger, Thomas. 2009. *The ego tunnel: The science of the mind and the myth of the self*. New York, NY: Basic Books.

Newell, Allen, J.C. Shaw, and Herbert A. Simon. 1958. Elements of a theory of human problem solving. *Psychological Review* 65 (3): 151–166.

Orlandi, Nico, and Geoff Lee. 2019. How radical is predictive processing? In *Andy Clark and his critics*, ed. M. Colombo, E. Irvine, and M. Stapleton, 206–221. Oxford: Oxford University Press.

Palmer, Stephen E. 1999. *Vision science: Photons to phenomenology*. Cambridge, MA: MIT Press.

Parr, Thomas, Dimitrije Markovic, Stefan J. Kiebel, and Karl J. Friston. 2019. Neuronal message passing using mean-field, Bethe, and marginal approximations. *Scientific Reports* 9 (1889): 1–18.

Paul, L.A. 2014. *Transformative experience*. Oxford: Oxford University Press.

Pezzulo, Giovanni, Francesco Rigoli, and Karl Friston. 2015. Active inference, homeostatic regulation, and adaptive behavioural control. *Progress in Neurobiology* 134: 17–35.

Putnam, Hilary. 1988. *Representation and reality*. Cambridge, MA: MIT Press.

Pylyshyn, Zenon. 1984. *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: MIT Press.

Rayner, Keith, and Erik D. Reichle. 2010. Models of the reading process. *WIREs Cognitive Science* 1: 787–799.

Rescorla, Michael. 2017. Review of Andy Clark, *surfing uncertainty: Prediction, action, and the embodied mind*. *Notre Dame Philosophical Reviews* 2017 (01): 15 https://ndpr.nd.edu/reviews/surfing-uncertainty-prediction-action-and-the-embodied-mind/.

Roskies, A.L., and C.C. Wood. 2017. Catching the prediction wave in brain science. *Analysis Reviews* 77 (4): 848–857. https://doi.org/10.1093/analys/anx083.

Ross, Don, and James Ladyman. 2010. The alleged coupling-constitution fallacy and the mature sciences. In *The extended mind*, ed. R. Menary, 155–166. Cambridge, MA: MIT Press.

Rowlands, Mark. 1999. *The body in mind: Understanding cognitive processes*. Cambridge: Cambridge University Press.

Rupert, Robert D. 2004. Challenges to the hypothesis of extended cognition. *Journal of Philosophy* 101: 389–428.

Rupert, Robert D. 2009. *Cognitive systems and the extended mind*. Oxford: Oxford University Press.

Rupert, Robert D. 2010. Extended cognition and the priority of cognitive systems. *Cognitive Systems Research* 11: 343–356.

Rupert, Robert D. 2011. Cognitive systems and the supersized mind. *Philosophical Studies* 152: 427–436.

Rupert, Robert D. 2013. Memory, natural kinds, and cognitive extension; or, Martians don't remember, and cognitive science is not about cognition. *Review of Philosophy and Psychology* 4 (1): 25–47.

Rupert, Robert D. 2015. Embodiment, consciousness, and neurophenomenology: Embodied cognitive science puts the (first) person in its place. *Journal of Consciousness Studies* 22: 148–180.

Rupert, Robert D. 2016. Embodied concepts, conceptual change, and *a priori* knowledge; or, justification and the ways life can go. *American Philosophical Quarterly* 53 (2): 169–192.
Rupert, Robert D. 2018. The self in the age of cognitive science: Decoupling the self from the personal level. *Philosophic Exchange* 47: 1–36.
Rupert, Robert D. 2019. What is a cognitive system? In defense of the conditional probability of co-contribution account. *Cognitive Semantics* 5: 175–200.
Schechtman, Marya. 2007. Stories, lives, and basic survival: A refinement and defense of the narrative view." Royal Institute of Philosophy Supplement 60: 155–178, Stories, Lives, and Basic Survival: A Refinement and Defense of the Narrative View.
Schechtman, Marya. 2011. The narrative self. In *The Oxford handbook of the self*, ed. S. Gallagher, 394–416. Oxford: Oxford University Press.
Segal, Gabriel. 1991. Defence of a reasonable individualism. *Mind* 100 (4): 485–494.
Shine James, M., Patrick G. Bissett, Peter T. Bell, Oluwasanmi Koyejo, Joshua H. Balsters, Krzysztof J. Gorgolewski, Craig A. Moodie, and Russell A. Poldrack. 2016. The dynamics of functional brain networks: Integrated network states during cognitive task performance. *Neuron* 92: 544–554.
Sims, Andrew. 2017. The problems with prediction: The dark room problem and the scope dispute. In T. Metzinger & W. Wiese (Eds.) *philosophy and predictive processing*: 23. Frankfurt am Main: MIND group. Doi: https://doi.org/10.15502/9783958573246 .
Staffel, Julia. 2019. How do beliefs simplify reasoning? *Noûs* 53 (4): 937–962.
Teskey, Morgan L., and Michael E. J. Masson. 2017. Components of competitor priming in task switching. Memory & Cognition 45: 1384–1397.
Velleman, David. 2005. The self as narrator. In *Autonomy and the challenges to liberalism: New essays*, ed. J. Christman and J. Anderson, 56–76. Cambridge: Cambridge University Press.
Walsh, Kevin S., David P. McGovern, Andy Clark, and Redmond G. O'Connell. 2020. Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences* 1464: 242–268.
Wheeler, Michael. 2005. *Reconstructing the cognitive world: The next step*. Cambridge, MA: MIT Press.
Wiese, Wanja. 2017. Action is enabled by systematic misrepresentations. *Erkenntnis* 82: 1233–1252.
Wiese, Wanja, and Thomas Metzinger. 2017. Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger & W. Wiese (Eds.) *philosophy and predictive processing*: 23. Frankfurt am Main: MIND group. Doi: https://doi.org/10.15502/9783958573246 .
Williams, Daniel. 2020. Predictive coding and thought. *Synthese* 197: 1749–1775.
Wilson, Margaret. 2002. Six views of embodied cognition. *Psychonomic Bulletin and Review* 9: 625–636.
Wilson, Robert A. 1994. Wide computationalism. *Mind* 103 (411): 351–372.
Yeung, Nick, Leigh E. Nystrom, Jessica A. Aronson, and Jonathan D. Cohen. 2006. Between-task competition and cognitive control in task switching. *The Journal of Neuroscience* 26 (5): 1429–1438.