Cognitive Systems, Predictive Processing, and the Self

Robert D. Rupert
U. of Colorado, Boulder
March 1, 2021

I. 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 arguments developed herein 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, those challenges give us reason to favor CPC, partly by increasing the chances that our ultimate theory of cognition will appeal to a heterogenous collection of mechanisms and processes.

[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, which comprises lots of components, stitched together, and has a reasonable claim to *being* the subject or self.

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 issues within the debate about extended cognition and extended mind. Section III presents CPC, a particular view of the human cognitive system – of what makes it a single integrated system – and identifies the deliverances of CPC with the cognitive self. Section IV develops the line of reasoning briefly described in the opening paragraph, 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.

II. Motivation for CPC: Extended Mind? Extended Cognition?

Why should one take interest in a theory of cognitive systems and their boundaries? One path runs via an 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

[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 related to the self," to accommodate the possibility of eliminativism about the self.

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, rather than on the status of individual states or processes taken in isolation. I emphasize extended *cognition* rather than extended mind, because, in philosophy, discussions of the mind tend to draw 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, the sort of perspectives that, in other domains, have been overturned or radically revised by careful scientific investigation. Thus, I approach questions about the mind as questions in the philosophy of science, philosophy cognitive science, in particular.

How, then, should we proceed, if we approach the boundaries of cognition as a topic in philosophy of science? Throughout the history of cognitive science, it has been assumed that two contrasting kinds of causes 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. The proponent of HEC should thus want to identify a natural kind or

[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.

property *(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 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 consistent and theoretically central distinction between two kinds of causes, one of which covers 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" versus "contributing from outside of that system" appears central. It runs through all manner of 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, 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 causes that contribute to the production of intelligent behavior, principled in the sense that it rests on successful practice across the cognitive science, this would seem to be our best bet – to distinguish two kinds of causes as, on the one hand, those that contribute from within the relatively persisting, relatively integrated system and, on the other hand, those that that contribute from beyond the boundary of that system.

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

[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 kind in question is a generic kind.
[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).

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.

III. The CPC

Cognition is a scientific kind, hypothesized by the relevant sciences to explain (what we take to be) a particular domain of phenomena, those concerning, in the first instance, various forms of intelligent human behavior: conversation in real time, patterns of similarity in the treatment of objects (what we might think of as 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, which I am calling 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, and I proposed 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 in the cognitive system and the ones 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.

[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 stable 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. Because CPC is location-neutral, though, if CPC does what it's meant to, anything that satisfies CPC is a cognitive system, which allows at least the possibility of extended cognitive systems in addition to an organism-bound cognitive system.

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 that do not.[8] Discard the rest.

7. The integrated cognitive system comprises all and only those mechanisms 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

[8] It's possible that multiple significant gaps appear on this list – at Step 6 as well as at Step 4 – which might seem to muddy the waters hopelessly. 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 the interesting version of HEC to project or extend causes of the "narrowest band" – causes from the most tightly integrated cognitive system – into the world beyond the organism, 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, possibilities for the identification of different kinds of causes 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 at least partly beyond the boundary of the organism are of the same kind of cause as what have traditionally been viewed as 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.

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

[9] Note, too, that integration is likely an irreducible property; in which case the measure of integration spelled out in the text should be thought of as highly diagnostic without providing a set of necessary and sufficient conditions. For typical subjects, with a significant amount of worldly experience, CPC accurately delineates the cognitive system and captures what it is about an integrated cognitive system that allows it to play its distinctive role, as the producer of flexible, adaptive behavior.

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 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, one 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 Steps 5 and 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 learning, 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 humans only 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 sort of process, subnetworks with overlapping members wrest control from each other via competitive processing. When two functional subnetworks have overlapping members, only a small amount of differential stimulus can shift the agent's activity from the performance of one task to the performance of another. 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 emphasizes what seems likely to be a central trait of such a system – that any given mechanism is capable of cooperating with various other subsets of mechanisms to complete a variety of tasks.

A further advantage of CPC, of special importance in the current context, is its neutrality with respect to the nature of 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 what the core feature of cognitive systems is, regardless of a given model's theoretical orientation. This high level of abstraction invites pluralism about the set of the possible mechanisms that compose any given cognitive system. Perhaps a single cognitive system contains some mechanisms best understood as connectionist. Perhaps others in the same system are computational, in a traditional sense, engaging in logic-based inference. Perhaps some in the same system are best understood using the tools of PP. CPC leaves the matter open, partly for the epistemic reasons to be emphasized below, that is, because we are not currently in a position to know exactly what kinds of mechanisms and processes produce intelligent behavior in humans, and we have reason to doubt that the set of mechanisms is homogeneous.

The CPC offers a theory of the integrated cognitive system, and, I maintain, thereby offers an account of the cognitive self. The cognitive system is the cognitive self. One might reasonably resist, however, holding that the self is only a proper part of the cognitive system. It might be only the part that constructs narratives (Schechtman 2007, 2011) and maybe 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 part of the cognitive system that manages the deployment of information (Metzinger 2009). I am not optimistic about such attempts to carve out a distinct portion of the cognitive system, treating only those delineated 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), working in tandem with various other parts of the cognitive system, and 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 phenomenon of interest, some parts of the cognitive system will play the role of especially active contributors.

IV. 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?[10]

My goal in this section is to foment doubt, in particular, 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 mechanisms. To the extent that PP has difficulty handling

[10] 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 weakness of an appeal to FEP's unifying power, in the context of theories of cognition, see Klein (2018) and Sims (2017). 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.

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.

I begin by noting some concerns already in the literature. 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, commits only to 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 (2018) 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 concerns, 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.

In what follows I'll add to the list of concerns canvassed above, 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. I do not aim to prove that comprehensive, or maximal, PP is false or hopeless as a research program, but rather 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:

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 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 impinges on the cognitive system; its features are detected or represented; and various of them become associated with each other.

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 functioning (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).

This approach may well pan out, but one should want some reassurance that the suggested accommodations are not *ad hoc*, that 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. The greater the extent to which one includes non-PP mechanisms in one's conception of PP-processing, 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-processing is the "canonical computation" (Walsh et al. 2020, 257) in human cognitive processing. In contrast, CPC has no problem taking 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.

2. *Feature representations fed forward: The case of the inveterately bad guesser*

On PP, only prediction-error signals are fed forward. In contrast, on standard non-PP views of the role of sensory input, one's beliefs are informed by the content of perceptual representations generated by stepwise extraction of increasingly abstract features or structured pieces of information from the sensory signal, as it moves inward from the periphery. Vision science has produced a wealth of evidence in support of this standard, 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 active 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 a correction of a bad guess; but that's largely incidental to the nature of the process, if 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 that deliver perceptual information in a standard feedforward fashion.

As noted, Orlandi and Lee (2019) press Clark on the importance of forward-flowing feature-related information in this kind of case. 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 (that is, which gist to arrive at, in Clark's terms). Here one might 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 does not seem to suffice to create determinate enough expectations to initiate a thoroughly PP-process.

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 can 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 fed-forward representation of that feature 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. *The neural evidence*

There is a genuine question of neural evidence in support of PEM. A recent extensive survey (Walsh et al. 2020) emphasizes cases in which support for PEM has been found, but the authors

make 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.

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. If the information arrives in the form of representations, it would appear quite plausible that such information is then integrated into the model by the operation that updates the *content* of the model in light of the *content* of the representation.[11] Standard methods of modeling such content-sensitive processes involve the transformation of content-laden units by computational operations, nothing specifically to do with PP-based mechanisms, particularly when this happens very quickly rather than by habituation – for instance when extensive model changed is caused immediately, by, for example, linguistic input of the sort, "Sorry, mate, you've got that all wrong; the world is *this* way instead."

Similarly, where there is neural evidence of the transmission of error-signals, error neurons sometimes seem to be saying only "Wrong! Try harder!" (Walsh et al. 2020, 254). In which case, the changing of the model now becomes entirely a process of "reflective cogitation," a process by which the subject must reason her way through the prediction to try to trouble-shoot her model (or set of standing beliefs). Reasoning about the strengths and weakness of a model

[11] 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.

based on one's having been told that one is wrong does not seem to be an especially PP-based process.

6. *Pseudo-conditionalization*

In a similar vein, consider the problem of updating a model conditionally, for the purpose of hypothetical reasoning, which is sometimes referred to as 'pseudo-conditionalization' (Staffel 2019). For example, I can ask myself how I would update my commitments – including various conditional commitments, such as the acceptance of certain likelihoods – were I were to believe a given proposition fully (even though, as a matter of fact, I do not believe it fully). How does this process proceed on a PP-based view? Computer scientists have, over the decades, devised various sampling methods and techniques for simplifying search that might be relevant, but it's not clear whether these processes will fit into a PP framework; some of the proposed methods – for instance, the use of particle filters (Griffiths, Vul, and Sanborn 2012, 266) – might require the brain to keep track of series of features in the previous input, which doesn't naturally fit into the PP framework. They involve something more like the dynamic binding of variables and storage of patterns of such bindings.

7. *Context, discourse models, and reading comprehension*

Proponents of PP frequently invoke the effects 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 information from which the contextualizing information must be inferred). The point here is

partly to suggest that the PP-based approaches themselves rely, for their plausibility, on language-based context-setting processes, but my point is also to indicate the great extent to which language-processing requires the subject's maintenance of a model of the discourse the subject is engaged in or of the 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).[12]

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, questions on a reading-comprehension exam.

8. *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. Taking into account the need to add auxiliary hypotheses and ancillary assumptions, in order to generate predictions, classic Duhem-Quine concerns arise. When error results, it can be unclear which aspect of the prediction-generating apparatus to blame. What I especially want to emphasize is that the decision what to change in these cases is *itself* a cognitive process. The data must be processed

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

and reasoned about. A representation of what picture the data seem to present must be constructed, 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 I predicted," but rather "how things seem to be." During this period, the subject asks, "Does the way things present themselves as being actually jibe with my model of the world?" 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, contradict what the model would predict; 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 or, more generally, feedforward hypothesis construction – the construction of how things seem or 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.

9. *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)? (What happened to Paul on the road to Damascus? Can a PP-based approach model it?) 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.[13]

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.

10. *The non-optimal and the role of 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 limitations – 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

[13] In this 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 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 PP-based 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 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. Rather, my point is that the challenges in question are genuine and

that, in light of them, optimism about PP's unmitigated and universal success requires no small leap of faith. For this reason, CPC's account of cognitive systems – and, thus, CPC's account of the self – has certain advantages. It is to its distinct credit that it is consistent with 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.[14] 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

[14] 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.

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 such a great extent; it takes a claim about actual cognitive processes and mechanisms and recasts it as something much more nebulous and interest-relative. In what follows, I articulate some further concerns about this kind of response.

First, we should beware of teleological thinking in cognitive science. The fundamental goal of cognitive science is to formulate and test models that account for human performance;[15] 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, without some explicit justification (such justification as does exist in the case of CPC, that it 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, it is because doing so guides the discovery of mechanisms or the best models of performance (see point 10 above). 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

[15] 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).

the phenomena in question, rather than the reality so discovered. Fourth, we should be wary 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), just as any physical system can be seen as a computational system (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 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 does seem a bit arbitrary to think that the purpose of cognition is 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 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.

In summary, many central aspects of cognition have yet to yield fully to PP, and at present we do not have compelling reason to believe that 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.

V. 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 narrowly focused 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 range of inputs, portions of 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).[16]

Moreover, there's a reflexive aspect of H&M's proposal. According to H&M, part of the overarching generative model (or family of models) is, itself, the self. The self-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).[17]

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

---

[16] 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).

[17] 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 inferences from one is told about minds in general – can influence one's model of oneself, via social feedback loops.

> 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 more 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). 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.

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 becomes someone that misrepresents 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 resulting 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. More needs to be said, however. Additionally, we would like to be able to accommodate cases in

which a subject right now represent herself as, for instance, generous, even though her generative model is not disposed to cause generous behavior. And, it remains unclear how H&M's 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 active inference.

This suffices, I think, to put on display the oddness of a PP-based view as it relates to 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 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). But, 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?

How might PP accommodate focused, conscious thought? 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.

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 prediction of the internal sensory pattern. It is one thing to make the conditional prediction "if someone thinks "P," they're likely to do x, y, or z," but it's another to predict the conjunction, "I will think P and I will do x, y, and z at certain specified times." The first sort of process seems not implausible; it's the sort of conditional often associated with a theory of mind. But, what seems to be required in the present context is a structure of the second sort, in keeping with which "P" is itself a product of active inference. That doesn't seem to be an account of conscious thought, though, 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. So, the thinking doesn't really have anything obviously to do with PEM; rather, it's the shuffling around of structures in one's model of the world, that is, in one's set of beliefs – which is an entirely orthodox, non-PP account of thought. Perhaps the production of "P" – silently voiced in one's head – by active inference can be given some sort of PP-related role; but the thinking that is an expression of the self is the activity of the model, activity that can just as well not give rise to "P" in internal voice[18] and is not a matter of predicting what I myself am going to think; I've already thought it by generating the prediction of "P"!

[18] Bear in mind, too, that deliberate, conscious thought stands as a paradigmatic example of content-sensitive and structure-sensitive logical reasoning, which harks back to concerns detailed in section IV.

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 of the current proposal – which would seem to requires representing all of these contingencies – are overwhelming and the storage and memory mechanisms are underexplained. 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.[19]

In the end, a host of problems and promises remain. This essay aims not at the conclusive. I do hope, however, 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. Bear in mind, too, the potential of the present reasoning to generalize. Wherever one finds support for hybrid systems (Jilk et al. 2008) or reason to doubt one cognitive-scientific research program's claim to comprehensiveness (Fodor 1975), one also finds an argument in support of CPC as a theory of the cognitive system and perhaps also as an account of 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

[19] Another possible path 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? How does such structured processing take place in both streams? Is one stream meant to represent the "real" self"? It's one thing to say that the cognitive system switches back and forth between the use of System 1 or System 2 (as is hypothesized by proponents of PP in the case of model-free versus model-based processing). What's needed here is something more 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. One should also ask whether the picture this approach delivers 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 a model.

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 modeling endeavors is best explained by its tracking something genuine, an important target kind *integrated cognitive system*. CPC is a reasonable attempt to track 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.

Works Cited

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: 10.15502/9783958573031 .

Clark, Andy. 2019. Replies to critics: In search of the embodied, extended, enactive, predictive (eee-p) mind. In M. Colombo, E. Irvine, and M. Stapleton (Eds.), *Andy Clark and his critics*. Oxford: Oxford University Press
DOI: 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, DOI: 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.

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

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

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

Fodor, Jerry. 1975. *The language of thought.* Cambridge, MA: Harvard University 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 D. Gentner and S. Goldin-Meadow (Eds.), *Language in Mind: Advances in the Study of Language and Thought* (Cambridge, MA: MIT Press), pp. 195–235.

Goodman, Noah. D., Joshua B. Tenenbaum, and Tobias Gerstenberg. 2015. Concepts in a probabilistic language of thought. In E. Margolis and S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (Cambridge, MA: MIT Press), pp. 623–654.

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.

Griffiths, Thomas L., Edward Vul, and Adam N. Sanborn. 2012. Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science* 21, 4: 263–268.

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

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.

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 M. Colombo, E. Irvine, and M. Stapleton (Eds.), *Andy Clark and his critics* (Oxford: Oxford University Press), pp. 206–221.

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

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

Pezzulo, Giovanni, Francesco Rigoli, & 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 doi:10.1093/analys/anx083 .

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

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: 343–56.

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 (2013): 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.

Schechtman, Marya. 2011. The narrative self. In S. Gallagher (Ed.), *The Oxford Handbook of the Self* (Oxford: Oxford University Press), pp. 394–416.

Segal, Gabriel. 1991. Defence of a reasonable individualism. *Mind* 100, 4: 485–494.

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: 10.15502/9783958573246 .

Staffel, Julia. 2019. How do beliefs simplify reasoning? *Noûs* 53, 4: 937–962.

Velleman, David. 2005. The self as narrator. In J. Christman and J. Anderson (Eds.), *Autonomy and the challenges to liberalism: New essays* (Cambridge: Cambridge University Press), pp. 56–76.

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, 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: 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.