Synthese
https://doi.org/10.1007/s11229-018-1768-x

CrossMark

Predictive coding and thought

Daniel Williams[1,2]

Received: 5 October 2017 / Accepted: 15 March 2018

Abstract Predictive processing has recently been advanced as a global cognitive architecture for the brain. I argue that its commitments concerning the nature and format of cognitive representation are inadequate to account for two basic characteristics of conceptual thought: first, its generality—the fact that we can think and flexibly reason about phenomena at any level of spatial and temporal scale and abstraction; second, its rich compositionality—the specific way in which concepts productively combine to yield our thoughts. I consider two strategies for avoiding these objections and I argue that both confront formidable challenges.

Keywords Predictive processing · Predictive coding · Free-energy principle · Probabilistic graphical model · Hierarchical probabilistic generative model · Concepts · Thought · Higher cognition · Bayesian brain

> In philosophy, ever since Plato, the mainstream opinion has been that the mind is the organ of thought… Psychologists, for the last hundred years or so, have mostly viewed that sort of 'intellectualism' as an embarrassing remnant of the Enlightenment (Fodor 2011)

✉ Daniel Williams
dw473@cam.ac.uk

1 Faculty of Philosophy, Trinity Hall, University of Cambridge, Wychfield Site, Storeys Way, Cambridge CB3 0DZ, UK

2 Cognition and Philosophy Laboratory, Monash University, Clayton, Australia

Published online: 24 March 2018

1 Introduction

An increasingly influential thesis in cognitive science is that the brain is fundamentally a *probabilistic prediction machine*, continually striving to minimize the mismatch between self-generated predictions of its sensory inputs and the sensory inputs themselves (Clark 2016; Friston 2010; Friston et al. 2017a; Hohwy 2013; Seth 2015). When repeated up the hierarchical structure of the neocortex, this process of *prediction error minimization* is alleged to implement approximate Bayesian inference and generate the full spectrum of psychological phenomena that make up the mind. Drawing on work from statistical physics, theoretical biology, machine learning, and cognitive and computational neuroscience, this extremely ambitious emerging framework—variously titled "predictive processing," "predictive coding," and the "free-energy principle" in the broader literature (I will use "predictive processing"—see Sect. 2 below)—is currently creating both intense excitement and scepticism in the mind sciences, alongside a growing literature that explores its implications for foundational debates in the philosophy of mind and science (Hohwy 2013; Clark 2016 for overviews).

As even its most enthusiastic proponents acknowledge, one of the most important challenges for predictive processing is whether the mechanisms it posits can be extended to capture and explain *thought* (Clark 2016, p. 299; Hohwy 2013, p. 3; see also Roskies and Wood 2017).[1] In addition to perceiving and acting upon the world, we also *reason, deliberate, plan,* and *reflect*. These cognitive processes implicate concepts that can be flexibly recombined in productive ways and applied across an indefinite number of domains and tasks. If predictive processing cannot capture such manifest psychological capacities, it fails as a "grand unified theory of the brain" (Huang 2008)—as a "*complete framework*… for explaining perception, *cognition*, and action in terms of fundamental theoretical principles and neurocognitive architectures" (Seth 2015, p. 1; my emphasis).

This paper has two aims. First, I argue that extant views about the nature and format of cognitive representation within predictive processing are unable to capture two basic characteristics of conceptual thought: first, its *generality*—the fact that we can think and reason about phenomena at *any* level of spatial and temporal scale and abstraction; second, its *rich compositionality*—the specific way in which concepts productively combine to yield our thoughts.[2] Second, I identify and evaluate two strategies for overcoming this problem. The first dissociates the ambitious prediction error minimization framework from the problematic claims about cognitive representation I identify, and abandons the latter. The second attempts to explain away these alleged characteristics of thought as idiosyncrasies of natural language. I argue that both possibilities should be explored, but that they confront formidable challenges: in the first case, it strips predictive processing of substantive empirical content and thus much of its explanatory ambition; in the second case, it is not clear how the suggestion would help even if it were true, and there are good reasons to suppose it is false.

[1] I use "thought" as an intuitive catch-all term for a suite of "higher" (i.e. non-perceptual) cognitive capacities: reasoning, planning, deliberating, and so on (see Sects. 3 and 4 below).

[2] The current paper thus complements a growing body of literature that criticises the explanatory ambition of predictive processing as a grand unified theory of the brain (Colombo and Wright 2016; Klein 2016).

I structure the paper as follows.

In Sect. 2, I identify the central commitments and core theoretical structure of predictive processing, and distinguish these commitments from more schematic (and thus less constraining) frameworks in cognitive science. In Sect. 3, I outline what I call the "standard strategy" among advocates of predictive processing for extending the mechanisms it posits to the domain of "higher" cognition. In Sect. 4, I identify two related reasons why the standard strategy fails. I then conclude in Sect. 5 by outlining and evaluating the two most plausible strategies one might pursue in response to these objections.

2 Predictive processing as a cognitive architecture

The term "predictive processing" is used in many ways across the cognitive sciences. As I will use the term, it refers to an extremely ambitious theory that combines two ideas: first, a vision of the brain as an "organ for prediction error minimization" (Hohwy 2014, p. 259)[3]; second, a specific account of the representations and algorithms—a "cognitive architecture" (Thagard 2011)—implicated in this process of prediction error minimization (see Clark 2016; Hohwy 2013; Seth 2015).

The first of these ideas relates predictive processing to the "free-energy principle," according to which prediction error minimization is a special case of a more fundamental imperative in biological agents to self-organize (Friston 2010). The free-energy principle is of immense scientific and philosophical interest in its own right, and is supposed to provide the rationale for thinking that all neural activity is orchestrated around a single, overarching function (Hohwy 2015). Nevertheless, I will have little to say about it in what follows. My aim is to evaluate predictive processing understood as a theory in cognitive neuroscience concerning the information-processing strategies exploited by the brain. This theory should be able to stand on its own two feet (although see Sect. 5).

The second idea connects predictive processing to "predictive coding," probably the most influential name for the framework under discussion in the broader scientific literature (Bogacz 2017; Rao and Ballard 1998). Strictly speaking, however, predictive coding is an encoding strategy whereby only the unpredicted elements of a signal are fed forward for further stages of information processing (Clark 2013, p. 182). This idea plays a pivotal role within predictive processing, but it should not be confused for predictive processing itself, which situates this strategy within a much broader theoretical context (Clark 2016, p. 25).

2.1 Bayesian inference and predictive coding

To introduce predictive processing, I will first present the necessary background on Bayesian inference and predictive coding in this sub-section, and then outline more technical details on the nature of probabilistic graphical models more generally in the next (Sect. 2.2). Both sub-sections—and especially the latter—are more technical

[3] Note that it is in fact a vision of the brain as an organ for long-term, average prediction error minimization (see Hohwy 2015). For convenience I drop this qualifier in the text.

than is customary in introductions of predictive processing in the philosophical literature. Nevertheless, as will be clear, this level of detail is crucial for evaluating its ability to explain thought.

First, then, there has been something of a "Bayesian revolution" in cognitive science in recent years—an "explosion in research applying Bayesian models to cognitive phenomena" (Chater et al. 2010, p. 811). This revolution has been motivated both by a growing consensus that one of the fundamental problems that cognitive systems solve is inference and decision-making under *uncertainty*, alongside an appreciation of how Bayesian statistics and decision theory can be used to model the solutions to such problems in mathematically precise and empirically illuminating ways (Chater et al. 2010; Griffiths et al. 2010; Tenenbaum et al. 2011).

In the case of perception, for example, the brain must exploit the statistical patterns of activity at the organism's sensory transducers to recover the structure of the distal environment. This evidence, however, is notoriously "sparse, noisy, and ambiguous" (Tenenbaum et al. 2011, 1279). Specifically, it dramatically underdetermines the complex structured environmental causes that generate it and is transmitted via input channels that are vulnerable to corruption by random errors. Mainstream perceptual psychology understands this problem as an inference problem: the brain must infer the causes of its sensory inputs from the sensory inputs themselves (Helmholtz 1867). Because this is a paradigmatic instance of inference under uncertainty, an influential contemporary tradition models this inferential process as *statistical inference* in accordance with Bayes' theorem (Rescorla 2013).

Bayes' theorem is a derivation of probability theory, and states the following:

$$(Bayes' Theorem)\ \ \mathrm{P}(H|E) = \mathrm{P}(E|H)\,\mathrm{P}(H)\,/\mathrm{P}(E)\,.$$

Bayesian perceptual psychology models the perceptual system as a mechanism that infers the causes of its sensory inputs in accordance with this theorem (Geisler and Kersten 2002). In a typical model, the perceptual system assigns probabilities both to hypotheses estimating possible environmental states P(H) (its *priors*) and to different patterns of sensory evidence given such hypotheses (the *likelihood* P(E/H)). When confronted with new evidence (i.e. activity at the organism's sensory transducers), it then updates these priors to posteriors P(H/E) in conformity with Bayes' theorem.

Bayesian perceptual psychology is mathematically precise and has produced empirically impressive models of a range of perceptual phenomena (Rescorla 2013). More generally, its emphasis on *prior expectations* and optimal statistical inference provides attractive explanations of a range of general features of perception: how brains overcome the noise and ambiguity in their sensory evidence, how they effectively integrate evidence from across the sensory modalities, how they generate perceptual *constancies*, and why they systematically produce specific perceptual *illusions* (cf. Chater et al. 2010; Geisler and Kersten 2002; Kersten and Yuille 2003; Rescorla 2013). Further, although perception has plausibly been the domain where Bayesian modelling has gained the most widespread acceptance in cognitive science (see Chater et al. 2010), the frameworks of Bayesian statistics and decision theory have been applied to a much broader range of phenomena, reflecting the fact that most if not all cognitive tasks can fruitfully be understood in terms of abductive inference and decision-making

under uncertainty: language, memory, sensorimotor processing, judgements of causal structure and strength, and much more (cf. Chater et al. 2010; Tenenbaum et al. 2011).

Despite these attractions, Bayesian cognitive science confronts at least three big challenges. First, it is implausible that the brain explicitly follows Bayes' theorem. Exact Bayesian inference can be extremely slow and sometimes computationally intractable (Penny 2012a, b; Tenenbaum et al. 2011). Much work in statistics and machine learning is thus devoted to developing algorithms for *approximate* Bayesian inference (Penny 2012a, b, p. 2). Second, it needs an account of where the relevant priors and likelihoods come from. Without independent constraints on their content, there is a significant risk of post hoc model-fitting (i.e. *any* cognitive phenomenon can be modelled by identifying the right priors and likelihoods "just-so") (cf. Bowers and Davis 2012; Jones and Love 2011). Finally, once we have a story of the aetiology of priors and the algorithms responsible for approximate Bayesian inference, how do such processes get implemented in the brain's neural networks (Chater et al. 2010)?

A useful although slightly misleading[4] pathway to understanding predictive processing is as an answer to these questions. Very roughly, its answer is: by installing and updating a hierarchical probabilistic generative model through predictive coding and precision-weighted long-term prediction error minimization.

To understand how this works, one must grasp two basic ideas: first, how Bayesian inference can be formalized in terms of precision-weighted prediction error minimization; second, how the brain (and especially the neocortex) can do precision-weighted prediction error minimization and thereby implement approximate Bayesian inference through hierarchical predictive coding.

First, then, recall that Bayesian inference is a matter of identifying the hypothesis P(H/E) that best predicts the evidence P(E/H) weighted by its prior probability P(H). If we assume the hypothesis space and evidence are Gaussian (and can thus by described by their sufficient statistics), one can calculate this by comparing the mean value *m* of the prior distribution with the mean value *e* of the evidence to compute a *prediction error*—namely, the distance between these two values (Hohwy 2017; Seth 2015).[5] In this way the prediction error varies inversely with the likelihood of the hypothesis: the greater the hypothesis' likelihood, the less prediction error it generates. Bayes' theorem then answers the question: how much should the prior be updated in light of the prediction error it generates?

To calculate this requires some means of weighting the relative reliability or *uncertainty* of the two sources of information. With Gaussian distributions, this can be computed from their relative *precisions* (the inverse of the *variance* of the two density functions), the ratio of which determines the *learning rate* in Bayesian inference: the more precise the priors are relative to the evidence, the less the agent *learns* about the world—the less the prediction error influences the posterior—and vice versa (Hohwy 2017). Thus as one learns more about a domain, one's priors will become increasingly *precise* (less uncertain), and prediction errors will be weighted less in driving infer-

[4] Misleading because it wrongfully suggests Bayesian inference is the *function* of prediction error minimization (see Williams 2017).

[5] I use "probability distribution" to include density functions (i.e. continuous distributions).

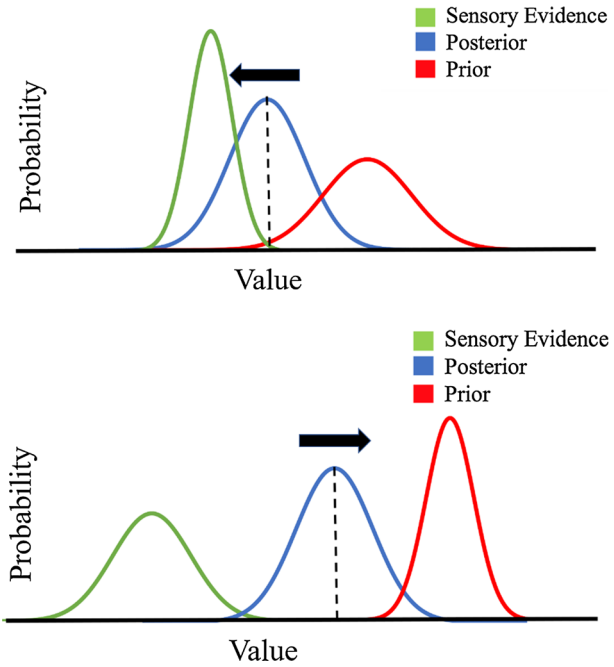

Fig. 1 Posterior beliefs are a function not just of the means but also the precisions of the sensory evidence and prior distribution. On the top graph, the comparatively greater precision of the sensory evidence has a proportionately greater influence on the posterior than the prior. On the bottom, this is reversed

ence—just as Bayes' theorem dictates. For a simple visual illustration of these ideas, see Fig. 1.

As stated, however, this process is far too simple to work in the case of real-world perception. Specifically, the sensory input received by the brain is not a linear function of environmental causes. If it *were*, recursively employing this process of Bayesian inference would result in arbitrarily precise priors and the optimal minimization of prediction error in the manner just described (Kiefer and Hohwy 2017). The real world, however, is richly structured, dynamic, and volatile, with causes interacting at multiple levels of spatial and temporal scale (Hohwy 2017). Under such conditions, an inferential system must be able to flexibly adjust the *learning rate* as conditions change. For example, it is no good holding steadfast to one's prior expectations about the location of an object if it is moved. This requires *hierarchical Bayesian inference*: if the evidence is a function of a complex hierarchically structured dynamic environment, the perceptual system must effectively *invert* this function, deconvolving the structural elements of the environment and forming expectations at multiple temporal scales in a manner that regulates the learning rate in context-sensitive ways. Such a system induces a *hierarchical Gaussian filter*, extracting regularities at different time scales from the non-linear time series of sensory input (Mathys et al. 2014; Hohwy 2017).

In formal terms, this means building a hierarchy of hypothesis spaces capturing regularities in the environment at increasingly greater spatial and temporal scale and abstraction (see Sects. 3 and 4 for more on this idea of a hierarchy). The hypotheses at

any level L + 1 thereby function as the *priors* for the level below L and the "sensory evidence" for the level above L + 2. The basic process of *prediction error minimization* is thus iterated up this hierarchy, ensuring that hypotheses higher up can guide inferences lower down, which in turn influence such higher-level expectations (Hohwy 2013, p. 31). In addition, such a system can build expectations for the *precisions* of sensory evidence at multiple levels, such that longer-term expectations can be exploited to adjust the learning rate in context-sensitive ways (Hohwy 2017). For example, such a system might learn that visual evidence is differentially reliable in various lighting conditions, and modulate the relative precision of visual evidence accordingly. A hierarchical Bayesian system of this kind is thus responsive not just to the causal structure of the environment but to the state-dependent level of noise and uncertainty as this structure is conveyed through sensory evidence (Hohwy 2013, p. 66).

So far I have described *exact* Bayesian inference, albeit of a specific, hierarchical form: such a system explicitly follows Bayesian inference in learning and updating its priors, and thereby keeps prediction error at an optimal minimum. We saw above that exact Bayesian inference is implausible both computationally and biologically, however.

This is where *predictive processing* comes in. Instead of claiming that brains follow exact Bayesian inference and thereby minimize prediction error, it states just the opposite—that brains are configured to minimize long-run, average prediction error and thereby come to *approximate* the results of exact Bayesian inference (Kiefer and Hohwy 2017, p. 18). The significance of this idea is this: whilst it is implausible to suppose that the brain explicitly follows Bayes' rule, it is not obviously implausible to suppose it might be structured in such a way as to minimize long-run prediction error. Minimally, what this requires is some way of matching predictions of incoming sensory input against the sensory input itself in a way that is responsive to the hierarchical structure of its environmental causes, and that factors in the uncertainty of these evidential sources.

This is where *predictive coding* comes in. As noted above, predictive coding is an encoding strategy whereby only the unpredicted elements of a signal are fed forward for further stages of information processing. In the case of the deeply hierarchical neocortex (Rolls 2016), predictive processing contends that "backwards" or "top-down" connections in the brain match predictions from "higher" cortical areas (e.g. frontal or temporal) against activity at "lower" cortical areas (functionally closer to the proximal sensory input), and that—as per predictive coding—only the deviations from such predictions (i.e. the prediction errors) are then fed forward (Kiefer and Hohwy in press).

Computing prediction errors through predictive coding in this way provides the brain with an internally accessible quantity to minimize. Predictive processing's proposal is that the brain's ceaseless efforts to minimize this quantity installs a rich, structured body of information about the environmental causes of sensory input—a *hierarchical generative model* (see below Sect. 2.2)—which can be exploited in guiding approximate Bayesian inference. The features of exact Bayesian inference required to minimize prediction error are thus thought to emerge naturally and in an unsupervised manner from the single imperative to minimize long-run prediction error (Kiefer and Hohwy 2017).

What is the evidence for this ambitious view of cortical information processing? Its motivation comes mostly from simulations of artificial neural networks either explicitly following this approach (e.g. Rao and Ballard 1998) or closely related approaches in artificial intelligence involving generative models and prediction-based learning (Dayan et al. 1995; Clark 2016, Ch. 1). At a more general level, although predictive processing is advanced approximately at Marr's (1982) "algorithmic" level of description (Clark 2016, p. 319), the mechanisms it requires—prolific feedback connections, hierarchical representation, functionally distinct units for predictions and prediction errors, and gating mechanisms for adjusting the influence of prediction errors as a function of their precision—are thought by at least some to map on convincingly to brain structures (Friston 2002, 2003, 2005; Gordon et al. 2017; Weilnhammer et al. 2017). Further, its commitment to a single principle of cortical functioning exploits evidence for the surprising uniformity and radical plasticity of the neocortex, alongside a widely held conviction that what functionally differentiates cortical areas is their input–output profile, not their computational architecture (George and Hawkins 2009; Hawkins and Blakeslee 2005; Mountcastle 1978).[6] When these facts are combined with both the increasing quantity of applications of the theory to different psychological and neurocognitive phenomena (see Clark 2016) and the immense popularity of Bayesian modelling in cognitive science (Chater et al. 2010), it is not surprising that predictive processing is currently enjoying the level of interest and excitement it does.

Much more could be said about the theory than the foregoing overview provides—for example, about how precision-weighting is supposed to align with the functional role of *attention* (Feldman and Friston 2010), and how *action* is supposed to arise through an inversion of the core process of prediction error minimization (cf. Clark 2016, Ch. 4). Nevertheless, this overview of its commitments will suffice for this paper. Next I outline the specific assumptions about cognitive representation that fall out of such commitments.

2.2 Hierarchical probabilistic generative models

So far, I've presented the broad architecture of predictive processing's view about cortical information processing: as brains minimize precision-weighted prediction error through hierarchical predictive coding, they acquire a structured body of knowledge about the ambient environment and exploit this knowledge to guide approximate Bayesian inference. In this sub-section I get clearer about how the nature of representation is understood within this framework—both because it further illuminates how this story is supposed to work, but also because it will be crucial for evaluating its capacity to explain conceptual thought.

First, then, the core representational structure in the predictive mind is the *hierarchical probabilistic generative model* (henceforth HPGM) (Clark 2016; Kiefer and Hohwy 2017). Conceptually, a generative model is a compact structure capable of generating a range of phenomena in a way that models (when accurate) the actual process by which the relevant phenomena are generated. In most engineering and theoretical

[6] See Marcus et al. (2014) for forceful dissent on this point.

applications of generative models, the process is *causal* (see below), and the phenomena are *data*—for example, the activation patterns provided as input to an artificial neural network (Kiefer and Hohwy 2017, in press, pp. 3–4). In predictive processing, the brain's generative model is encoded in top-down and lateral[7] synaptic connections in the neocortex. This model both *generates* predictions carried via "top-down" synaptic connections in a manner that represents or *simulates* the process by which the brain's sensory input is actually generated by the environment (Gładziejewski 2015; Williams and Colling 2017).

As noted above, this process of inversion is *hierarchical*. A generative model is *hierarchical* if it contains multiple levels of "latent variables" (variables whose values are not provided directly by the data). In effect, this amounts to a hierarchy *of* generative models, with each level's data represented as dependent on the latent variables at the level above (Kiefer and Hohwy 2017). Finally, a generative model is *probabilistic* if it represents not just the dependencies among the values of random variables but the dependencies among the probabilities of such values. For example, in a classification task in which the challenge is to assign classes C to examples E, a generative model determines the probability of examples given classes P(E/C), as well as a prior distribution P(C) over classes (Lake et al. 2016).

In many presentations of predictive processing, this overview of HPGMs would be sufficient. For our purposes, however, it is not. Specifically, the mere claim that the brain commands a HPGM is consistent with a variety of ways of *expressing* the information encoded in that generative model: as a simple neural network, a more structured graph, or a functional program, for example (Goodman et al. 2015; Tenenbaum et al. 2011). Such alternative representational formats have important consequences for both the utility and computational power of a generative model, as well as a theory's consilience with contemporary views about the structure and dynamics of biological neural networks (Tenenbaum et al. 2011). This point will be important below (Sect. 4.2). To flag the issue in advance, however: this paper does not address whether generative models can capture the basic characteristics of thought, but whether predictive processing with its specific commitments concerning the structure and format of the brain's generative models can.

Although this point is not often made clear in the literature, the structure of HPGMs in predictive processing can be understood in terms of *probabilistic graphical models* (Bastos et al. 2012; Denève and Jardri 2016; Friston et al. 2017b; Kiefer and Hohwy 2017; Lee and Mumford 2003; Penny 2012a, b). Graphical models were first introduced into cognitive science and artificial intelligence largely by Judea Pearl's seminal work (Pearl 1988). These data structures are *graphical* because they express the set of dependencies between the elements of a domain with a *graph*, a mathematical structure of *nodes* and *edges* that can be directed (A → B), undirected (A–B), or bidirected (A ← →).

Such graphs effectively comprise compact representations of dependence or *relevance* relationships (Danks 2014, p. 39). The graph itself captures the *qualitative* structure of a domain: namely, the pattern of dependence relationships between its

[7] I ignore the role of lateral connections in this overview (see Friston 2005) for space constraints and because it is not directly relevant to my argument.

elements. In most cases, this qualitative structure is complemented with quantitative parameters that represent the strength of the dependencies between these nodes. For example, in the graphical models that will concern us, the nodes represent variables and the edges represent probabilistic dependencies between them. As such, the qualitative information encoded in the graph must be complemented with a set of probability distributions that quantify how the probabilities assigned to some variables systematically depend on the values of others.

What makes graphical models so powerful is their graphical component. Imagine, for example, that you want to model some domain that contains 100 variables—a meagre size when contrasted with the complexity of the world we must navigate. A joint distribution over such a set of variables will thus be 100-*dimensional*, which is extremely difficult both to learn and manipulate (Jacobs and Kruschke 2010, p. 10). What a graph enables you to do is capture the probabilistic dependencies between these variables in terms of *direct relevance* (i.e. dependence) (Pearl 1988). Specifically, graphical models capitalize on the fact that whilst everything is in some sense relevant to everything else, not everything is *directly* relevant to everything else (Danks 2014). For example, which country someone lives in is relevant to how wealthy they are. But it is not directly relevant: once one factors in other information—their assets, the amount of money in their bank account, and so on—the knowledge of where they live becomes uninformative. In the vocabulary of probability theory, such variables are *conditionally independent* given observation of these other variables.

In a graph, this notion of direct relevance is captured in its edges. Roughly, if nodes are adjacent in the graph—that is, they are connected to one another by some edge between them—they are directly relevant to one another; if not, they are conditionally independent. As such, one can effectively factorize a high-dimensional joint probability distribution into a set of much lower-dimensional distributions in which the probability assignment of each variable depends only on the values of adjacent nodes (Jacobs and Kruschke 2010). When the parameters of a model accurately reflect the network of conditional independence relationships implied by its graphical structure in this way, the model satisfies the *Markov assumption*, which states that "if two nodes are not adjacent in the graph, then they are independent in the quantitative component [i.e. model parameters]" (Danks 2014, p. 43). This factorization comes with considerable representational and computational benefits (Fig. 2).

In predictive processing, the HPGMs are best understood as neurally instantiated hierarchical probabilistic graphical models of this kind (Bastos et al. 2012). Specifically, the variables at each level of the generative model (on out to the observable variables provided directly by the sensory input) are represented as directly dependent on the values of variables at the level above (i.e. their *parents*), such that each level effectively determines a conditional probability distribution on the level below. In this way, non-adjacent levels of the model are conditionally independent of one another in a manner that satisfies the Markov assumption. Predictive coding and prediction error minimization are then supposed to offer a computationally tractable account of how this hierarchical generative model gets both installed and updated in real time (cf. Bastos et al. 2012; Kiefer and Hohwy 2017).

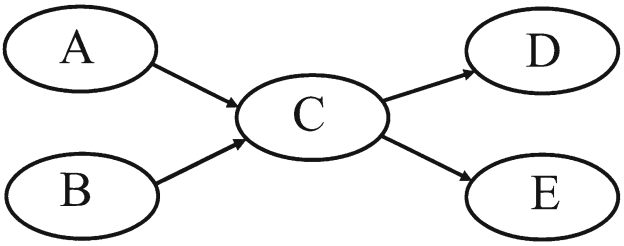

Fig. 2 A maximally simple directed graphical model capturing the conditional dependence structure between five random variables. In this case, even though the values of D and E are dependent on the values of A and B, learning the values of A and B tells one nothing about the values of either D or E *once one knows the value of C*. D and E are thus conditionally independent of A and B given C

Predictive processing is by no means alone in its commitment to the graphical structure of cognitive representations (Danks 2014; Gopnik and Wellman 2012; Griffiths et al. 2010; Hinton 2011; Sloman 2005). Nevertheless, it is not noted often enough that it *is* committed to this structure.

3 The standard strategy

In this section I set out the *prima facie* obstacles in the way of extending the cognitive architecture just outlined to capture and explain conceptual thought. I first set out the challenge (Sect. 3.1) and then outline the standard strategy in the literature for addressing this challenge (Sect. 3.2), before turning in Sect. 4 to explain why it is inadequate.

3.1 The challenge

First, then, I noted in Sect. 1 that predictive processing is advanced as a global theory of brain function. The aspect of this extreme explanatory ambition that I want to focus on here is its ability to explain distinctively conceptual *thought* or what is sometimes called "*higher cognition*." *That* it can explain such psychological phenomena is often asserted in the literature (my emphases):

> … the first truly unifying account of perception, *cognition*, and action (Clark 2016, p. 2).

> This mechanism is meant to explain perception and action *and everything mental in between* (Hohwy 2013, p. 1).

> Arguably, PP provides the most complete framework to date for explaining perception, *cognition*, and action in terms of fundamental theoretical principles and neurocognitive architectures (Seth 2015, p. 1).

However, despite the frequency with which such claims are made in the literature, there have been few systematic attempts to show how predictive processing might be

extended to capture and explain conceptual thought.[8] Further, it is not at all clear how any such extension might work. As presented above, the framework seems first and foremost to provide an account of unsupervised learning and perception (and action (Clark 2016, Ch. 4)). Clearly, however, there is much more to our mental lives than such perceptual responsiveness to the structure of the distal environment. We don't just perceive (and act). We reason, deliberate, plan, and reflect, and such psychological processes are often strongly decoupled from immediate perception.

In fact, there are really two problems here. The first is how *any* theory of the mechanisms implicated in perception could simply be extended to capture the mechanisms implicated in thought as well. If perception and cognition are characterized by different information-processing characteristics and capacities, as is commonly believed in cognitive psychology (Firestone and Scholl 2015, p. 1), it is not clear how a single theory could subsume both domains of psychological phenomena. Second, it is not clear how the *specific* mechanisms posited by predictive processing could capture the kinds of capacities exhibited in thought. What could precision-weighted hierarchical prediction error minimization have to do with planning and writing an academic paper, worrying about what your date thinks about your offer to pay the bill, figuring out how best to arrange a surprise birthday for a friend, and so on?

Of course, so far this at best amounts to a challenge—a vague expression of incredulity—and not a substantive objection. Before I turn in Sect. 4 to substantiate this incredulity, I first outline what I think is plausibly the standard strategy for answering this worry among advocates of predictive processing.

3.2 The standard strategy

The "standard strategy" can be extracted from the work of Clark (2013, 2016), Hohwy (2013) and Fletcher and Frith (2008), among others. It has three components, which I outline in turn.

First, it begins by either denying or deflating any *distinction* between perception and thought. Clark (2013, p. 190), for example, contends that predictive processing dissolves "at the level of the implementing neural machinery… the superficially clean distinction between perception and knowledge/belief," and Hohwy (2013, p. 72) notes that "it is *normal* for us to distinguish between on the one hand perception and experience, and on the other hand belief and thought… *Yet the perceptual hierarchy does not suggest that there is such a difference*" (my emphasis).

This challenge comes from at least two sources. The first concerns representation: as outlined in Sect. 2, there is a sense in which all information processing within predictive brains implicates *beliefs*—specifically, "Bayesian beliefs" encoding probability distributions over possible states of the world (Seth and Friston 2016). As such, *the same fundamental kind of representation* underlies representation in both sensory cortices and cortical regions responsible for intuitively "higher-level" cognition. The second concerns inference: according to predictive processing, *all* representations are updated through (approximate) *Bayesian inference.* Given that Bayesian inference

[8] Clark (2016) is an exception to this (see Sect. 5).

is a paradigmatically cognitive process, this encourages the idea that what we ordinarily distinguish as perception and thought are in fact subsumed under a broader information-processing regime.

If this is right, what accounts for the strong intuition of a difference between perception and thought? The second component of the standard strategy *explains away* this intuition as a response to different levels of the brain's hierarchical generative model. Specifically, what we ordinarily characterise as perception tracks representations in the generative model computationally closer to the sensory periphery estimating environmental causes at fine-grained spatiotemporal scale, whereas more intuitively *cognitive* activities correspond to representations of worldly phenomena implicated in larger-scale, longer-term regularities (Clark 2013; Hohwy 2013). Hohwy (2013, p. 72), for example, suggests that "perceptions are… basically shorter-term expectations and concepts longer-term expectations," and Clark (2013, p. 190) asserts that "in place of any *real* distinction between perception and belief we get variable differences in the mixture of top-down and bottom-up influence and differences of temporal and spatial scale in the internal models that are making the predictions" (my emphasis). In addition, Hohwy (2013, pp. 72–73) suggests that more intuitively conceptual representations comprise discrete distributions over categorical entities, whereas continuous densities capture more intuitively perceptual experience (see also Friston et al. 2017b).[9]

Importantly, these components are not restricted to philosophical analyses of the implications of predictive processing. In a highly influential article on the emergence of hallucinations and delusions in schizophrenia, for example, Fletcher and Frith (2008) draw on predictive processing to offer a unified account of *both* phenomena in terms of a common underlying computational architecture. Exemplifying the first component of the standard strategy, they deny that "perception and belief formation… [comprise] distinct processes," citing "recent advances in computational neuroscience" (i.e. predictive processing) to argue that "the unusual perceptual experiences of patients and their sometimes bizarre beliefs… [are] part of the same core abnormality…" (Fletcher and Frith 2008, p. 48). Expressing the second component, they then claim that there is "a hierarchy of… inferencing devices in the brain where lower levels of the hierarchy are more relevant to perception and upper levels and more relevant to beliefs" (Fletcher and Frith 2008, p. 56).

As stated, however, these first two aspects of the standard strategy do not address a major component of the worry raised above—namely, that much of our psychological lives is decoupled from direct perceptual engagement with our environments.

The third part of the standard strategy is to try and capture such decoupled psychological phenomena in terms of "offline simulation" (Clark 2016; Pezzulo 2017). As Clark (2016) argues, once a predictive brain acquires a rich, structured generative model with which it can perceive the world, a capacity to *imagine* that world—to engage in a "deliberate imaginative exploration of our own mental space (273)—is attractively co-emergent (8). Within the context of predictive processing, such purely simulative capacities are made possible by removing the constraints of incoming sen-

[9] Strictly, this would require a slightly different architecture to standard predictive processing (see Sect. 2, and Friston et al. 2017b), although discrete distributions are plausibly *easier* to compute with than density functions.

sory evidence on cortical processing, such that top-down predictions are not made to minimize the prediction errors they generate. Taking the model "offline" thus enables the system to simulate hypothetical causal processes and engage in counterfactual reasoning, thereby facilitating cognitive capacities directed at temporally distant or counterfactual states of affairs (Clark 2016, Ch. 3).

Unfortunately, concrete testable models within this ballpark are for the most part yet to be developed. Nevertheless, there is a wide range of speculative accounts concerning how this potential aspect of the predictive processing architecture might be able to illuminate an array of psychological phenomena—phenomena such as dreaming (Hobson and Friston 2012), social cognition and coordination (Kilner et al. 2007), memory (Henson and Gagnepain 2010), planning (Clark 2016, Ch. 3), and more (cf. Clark 2016). Whatever the details of such accounts, predictive processing clearly must exploit this potential of its cognitive architecture if it is to have any hope of accounting for the range of cognitive capacities we exhibit that are decoupled from online perceptual activity or action.

I don't pretend that the foregoing three components of the standard strategy capture *all* work within predictive processing when it comes to accounting for higher cognition—indeed, I return in Sect. 5 to work that points to the transformative impact of public symbol systems and structured social practices on the predictive mind (cf. Clark 2016)—but I do think these three components are representative of the basic strategy by which advocates of predictive processing have sought to address the challenge with which I began this section. I next turn to render this challenge precise and argue that the standard strategy does not successfully answer it.

4 Two challenges to predictive processing

Despite the attractions of the standard strategy, it doesn't work. Specifically, predictive processing is unable to capture two basic and related characteristics of conceptual thought: first, its *generality*—the fact that we can think and reason about phenomena at any level of spatiotemporal scale and abstraction; and second, its *rich compositionality*—the specific way in which concepts are productively combined to yield our thoughts. These characteristics undermine predictive processing's commitment to hierarchical representation and probabilistic graphical models as the whole truth about cognitive representation in the brain. Before I advance the arguments in defence of these claims, I first briefly note two qualifications.

First, and most obviously, I make no claim to originality in identifying these cognitive capacities. In fact, they plausibly stand at the centre of most prominent historical debates about cognitive architecture—about the prospects of behaviourism, connectionism, dynamicism, and embodied cognition, for example. Specifically, they are those capacities most amenable to the flexible combinatorial symbol systems of "classical" cognitive science (Newell and Simon 1976). What is novel here is the analysis of why predictive processing cannot accommodate them. This is not a simple rehashing of classical critiques of connectionism (Fodor and Pylyshyn 1988). Predictive processing represents a genuine improvement over such previous models, and Clark (2015, 2016) explicitly claims that hierarchical generative models within predictive

processing can answer the worries about structure and representational flexibility that motivated such critiques. I think he is wrong (see below), but it must be shown why.

Second, and hopefully equally obviously, I do not claim that these are the *only* characteristics of cognition or human psychology more generally that pose a challenge for predictive processing.[10] Instead, the claim is more modest: that the characteristics I identify are easily grasped, are central to conceptual thought, and transparently highlight deep structural problems with the cognitive architecture it posits.

4.1 The generality of thought

The first challenge is straightforward: we can think and reason about phenomena at *any* level of spatial and temporal scale or abstraction in a way that flexibly combines representations across such levels; extant views about cognitive representation within predictive processing are unable to accommodate this fact; therefore, such views are mistaken.

To see the problem here, begin with Vance's (2015) observation that predictive processing seems to be committed to two ideas about the representational hierarchy: first, that it tracks computational distance from sensory surfaces; second, that it tracks representations of phenomena at increasingly larger spatiotemporal scales.[11] As Vance (2015) notes, these two ideas seem to be straightforwardly in tension with one another: "person-level beliefs can be computationally far from the sensory surfaces and can also represent things at small spatiotemporal scales"—a fact he illustrates by reference to our ability to think about sub-atomic particles moving close to the speed of light. Specifically, if beliefs are supposed to exist higher up the hierarchy (see Sect. 3.2) and moving higher up the hierarchy is supposed to result in representations of phenomena at larger spatiotemporal scales, it should be impossible to have beliefs about extremely small phenomena implicated in fast-moving regularities. Given that this is evidently not impossible, the standard story about the representational hierarchy must be mistaken.

The problem, however, is much deeper than this, and it is worth unpacking it in some detail.

First, we can think and flexibly reason about phenomena at *any* spatiotemporal scale. The simplest way to see this is to note that we can think about any of the phenomena represented by our perceptual systems: distributions of light intensity, oriented edges, fast-moving patterns of auditory stimuli. If we could not, neuroscience—and thus predictive processing—would be impossible. Further, such representation is not a matter of activating parts of the relevant sensory cortices "offline" (see Sect. 3.2).

[10] For example, Colombo and Wright (2016) have recently drawn on the systemic role of mesocorticolimbic dopaminergic systems to challenge the explanatory scope of predictive processing, and Klein (2016) argues that the framework cannot account for motivation. My challenge here is intended to complement these other critiques.

[11] Clark (2012, p. 762), for example, characterises "a hierarchy of increasingly abstract generative models" as "models capturing regularities across larger and larger temporal and spatial scales," and Hohwy (2012, p. 2) writes, "In general, low levels of the hierarchy predict basic sensory attributes and causal regularities at very fast, millisecond, time scales, and more complex regularities, at increasingly slower time scales are dealt with at higher levels." For more evidence and quotes, see Williams (forthcoming).

A blind person can think about light patches or binocular disparity. This strongly undermines any attempt to restrict conceptual thought to regions of an inferential hierarchy if such regions are supposed to represent a specific spatiotemporal range of phenomena. Further, this point doesn't apply to uniquely "high-level" reasoning of the sort found in deliberate intellectual or scientific enquiry: patients suffering from "delusional parasitosis" (Prakash et al. 2012) wrongly *believe* themselves to be infested with *tiny* parasites, insects, or bugs—a fact difficult to square with Fletcher and Frith's (2008) suggestion outlined in Sect. 3.2 that delusions arise in "higher" levels of the hierarchy, if this hierarchy is understood in terms of increasing spatiotemporal scale (see Williams forthcoming).

In response, one might note that what predictive processing is *strictly* committed to is that higher levels within the hierarchy identify the environmental variables that best predict the regularities (statistical patterns) identified at the level below. For example, in deep learning applications in artificial intelligence research, this process is a matter of recursively extracting *the sources of mutual information* among variables, where two variables are mutually informative if observation of one reduces uncertainty about the other (Goodfellow et al. 2017; Ryder forthcoming). Importantly, this account *subsumes* the view that representations higher up represent phenomena at larger spatiotemporal scale, but it is also consistent with representations of phenomena—for example, individuals and kinds—where it might not strictly make sense to speak of different spatiotemporal scale, but which are nevertheless genuinely predictive of patterns of sensory input. Further, advocates of predictive processing often stress that the hierarchy here is in fact better understood as a *sphere* or *web*—with proximal inputs across different modalities perturbing the outer edges—than a stepladder (Penny 2012a, b). This allows for *multimodal* or *amodal* representations to effectively predict the phenomena represented across different sensory modalities. For example, representing the presence of a dog predicts a multitude of visual, tactile, olfactory, auditory, and interoceptive sensations. If this is right, perhaps it is only hierarchical representation in *sensory cortices* that functions in terms of increasing spatiotemporal scale, and that as one moves into association cortex representations latch onto more "abstract" environmental variables.

This response confronts a dilemma, however: *either* it simply abandons the view that conceptual representation is meaningfully located at some region of a hierarchy in favour of a conventional understanding of concepts as an autonomous domain of amodal representations capable in principle of ranging over *any* phenomena (e.g. Pylyshyn and Fodor 2015; see below Sect. 4.2); *or* it offers some principled means of characterising the region of the hierarchy involved in non-perceptual domains. The challenges for this latter option are formidable: *prima facie*, at least, the fact that we can engage in highly abstract thought about phenomena at *any* level of spatiotemporal scale undermines any simple hierarchically iterated process of prediction error minimization as the whole truth about cognitive representation. Do my thoughts about electrons activate representations at a different position in "the hierarchy" to my thoughts about

Paris, the English football team's defensive strategy, or the hypothesised block universe? If so, by what principle?[12]

The problems here are easiest to see when one focuses on the second strand of the problem: we can think and reason about phenomena at any level of *abstraction*. Specifically, it is not just that we can think about concrete phenomena at any spatiotemporal scale; we can think about phenomena with *no* spatiotemporal scale—for example, *Bayesian inference*, the number *3*, representational *content*, *postmodernism*, and so on. Such phenomena play no causal role and thus are predictive of no regularities (although of course our thoughts about them might themselves play causal roles). Insofar as inter-level relations in the inferential hierarchy are determined by a recursive process of prediction error minimization, it is not obvious how representations of such abstract entities could feature *at all*.

Finally, the third dimension of the challenge in this sub-section concerns not just our ability to think about this varied range of phenomena but our capacity to flexibly reason about phenomena in a way that combines representations from across levels of any conceivable hierarchy. To see the problem here, consider the cognitive architecture of predictive processing as outlined in Sect. 2: on this view, the Markov assumption codifying the conditional independence of non-adjacent levels of the hierarchy (see Sect. 2.2) ensures that the influence of representations at non-adjacent levels is always mediated by representations in the levels between them. The problem with this idea, however, is that conceptual thought and reasoning do not have this restriction: one can combine representations of phenomena at different levels of spatiotemporal scale and abstraction into structured beliefs and exploit these beliefs in reasoning in a manner that looks flatly inconsistent with an understanding of conceptual representation in terms of hierarchical graphical models. For example, we can think about how the laws of physics would have to change to make it possible for an elephant to move faster than the speed of light.

This point relates to Gareth Evans' (1982, p. 104) famous "generality constraint" on conceptual thought, according to which:

> "If a subject can be credited with the thought that a is F, then he must have the conceptual resources for entertaining the thought that a is G, for every property of being G of which he has a conception."

I have been using the term "generality" in a different sense to Evans—roughly, to indicate the range of phenomena about which we can think and reason, rather than

[12] One interpretation that might be attributed to Hohwy (2013) is that levels within the inferential hierarchy are determined not by *what* they represent—their content—but by the relative time-scale of the representations themselves (see, e.g. Hohwy's claim referenced above that "perceptions are…shorter-term expectations and concepts longer-term expectations" (2013, p. 72)). It is difficult to make sense of this suggestion, however, hence why I have not included it in the main text. First, although active perceptual states might be short-lived, the representational system from which they are generated is just as enduring as other representational systems. The putative difference must therefore concern the time-scale of *active* representations at different levels of the hierarchy. But insofar as active representations in different parts of the brain exhibit differential levels of persistence or invariance, this is itself explained by *what* they represent, and is highly context-sensitive. For example, deliberately focus your vision on an unchanging stimulus that allows for multiple interpretations, and let your interpretation change back and forth between them: the low-level perceptual state remains the same; your conceptual interpretation changes.

the way concepts can be combined to yield such thoughts. Nevertheless, the two ideas now relate to one another fairly directly: our capacity to think and flexibly reason about phenomena of any spatiotemporal scale and abstraction reflects our ability to combine concepts into structured thoughts irrespective of their contents. The problem with predictive processing is that hierarchical generative models of the sort it posits appear not to satisfy this constraint: the relation of different representations to one another is mediated entirely by their place in the specific *network structure*, sharply delimiting the allowable representational combinations.

To understand the nature of these deficiencies in detail requires us to turn to the second challenge of this section: the *rich compositionality* of thought, and predictive processing's current inability to capture this compositionality.

4.2 The rich compositionality of thought

The previous section focused on characteristics of thought that the predictive processing cognitive architecture cannot account for. In this section I focus instead on a widely accepted explanation of conspicuous characteristics of thought: its compositionality. As before, the argument of the section is straightforward: thought is richly compositional; extant views about cognitive representation within predictive processing are inconsistent with this fact; therefore, such views are mistaken.

Compositionality is a familiar principle of representational systems in which the representational properties of a set of atomic representations compose to yield the representational properties of molecular representations. For example, the content of "Ben loves Bob" is a compositional function of the grammar and meanings of its constituent expressions—"Ben," "Bob," and "loves"—and the way in which they are combined. If one combines a compositional representational system with recursive rules whereby the outputs of some functions provide the inputs to others, one thereby yields the infinitary character of representational systems such as natural languages.

The compositionality of the representational system underlying conceptual thought is widely considered a necessary explanation of two basic characteristics of cognition: its productivity and systematicity (Fodor 1975; Fodor and Pylyshyn 1988). "Productivity" names the infinitary character of thought. "Systematicity" names the systematic relations our thoughts bear to one another, such that an ability to think some thoughts is inherently tied to the ability to think others (Fodor and Pylyshyn 1988). For example, if one can think that Bob loves Ben, one can think that Ben loves Bob.

Nevertheless, to say that thought is systematic, productive, or compositional is not to say very much. The reason is that different representational systems have different compositional properties and thus generate different forms of productivity and systematicity. As such, my first premise is that thought is *richly* compositional, where by "richly" I mean the following: *at least as expressive as first-order logic*. Because predictive processing is committed to HPGMs whose structure is captured by probabilistic graphical models, it cannot capture this characteristic of thought. Specifically, graphical models are effectively restricted to the expressive power of propositional logic (Russell and Norvig 2010, p. 58; cf. also Goodman et al. 2015; Russell 2015).

To see this, it will be helpful to briefly review a general divide in the history of cognitive architectures between approaches that stress *symbols* and those that stress *statistics* (see Goodman et al. 2015; Tenenbaum et al. 2011). The former approach focuses on the compositional, rule-governed character of cognitive capacities, and is exemplified in the symbol-manipulating paradigm of classical cognitivism. Famously, whilst such architectures are well-suited to modelling abstract domain-general logical reasoning, they are traditionally confined to the realm of *certainty* and are widely held to be at odds with known facts about neurobiology, rendering them brittle and unrealistic models of cognition (Churchland 2012; Pearl 1988). By contrast, the latter cognitive architectures—exemplified by work on artificial neural networks, for example—are much more adept at dealing with uncertainty and make closer contact with neurobiology, but typically lack the structured and compositional representational abilities seemingly necessary to capture and explain higher-level cognition (Fodor and Pylyshyn 1988; Tenenbaum et al. 2011).

Clark (2015, 2016, p. 24, pp. 171–174) contends that the HPGMs in predictive processing can capture the attractions of *both* traditions without suffering the drawbacks of either, combining neurobiological realism and statistical inference with the rich, hierarchically structured representational capacities central to more traditional symbolic models (see Clark 2016, p. 24, pp. 171–174):

> The worries [advanced, e.g., by Fodor and Pylyshyn 1988] about structure are *directly addressed* because… prediction-driven learning, as it unfolds in these kinds of multilayer settings, tends to separate out interacting distal (or bodily) causes operating at varying scales of space and time (Clark 2016, p. 24).

To understand this, it will be helpful to return to the discussion of probabilistic graphical models in Sect. 2. First, such models *are* compositional. As with all statistical models, their compositional properties are determined by the possible assignments of values (and assignments of probabilities to these values) to their variables. These possible assignments correspond to the space of possible worlds such models can represent, with different assignments corresponding to different worlds (Kiefer and Hohwy 2017).

For this reason, probabilistic graphical models are at least weakly productive, in that such systems can represent phenomena they have never represented before. Further, the modularity of graphical models enabled through conditional independence relationships (see Sect. 2.2) ensures that new random variables can be added rather painlessly—that is, without resulting in a combinatorial explosion (Danks 2014, p. 44). Finally, as with statistical models more generally, such graphical models exhibit a kind of systematicity. In a supply/demand model, for example, an ability to represent that supply = X and demand = Y goes hand in hand with an ability to represent that supply = Y and demand = X.

The aspect of HPGMs in predictive processing that Clark thinks is most important, however, is their *hierarchical structure*. Specifically, this enables such networks to effectively deconvolve the structural elements of the environment, separating out functionally significant environmental causes in a way that radically amplifies their compositional abilities (Clark 2016, p. 24). As such, these HPGMS in effect capture compositional relationships *between environmental variables*, explicitly representing

how simple environmental elements are parts of and thus systematically related to the larger structures within which they participate. In other words, such models not only flexibly represent possible worlds but rich, *structured* possible worlds. As Clark (2015, p. 5) puts it:

> [Predictive processing] offers a biologically plausible means… of internally encoding and deploying richly structured bodies of information… [enabling] us, as agents, to lock onto worldly causes that are ever more recondite, capturing regularities visible only in patterns spread far in space and time….

Despite this, such hierarchical probabilistic graphical models are insufficient. They are not *richly* compositional. As noted, their expressive power is equivalent to propositional logic (Russell and Norvig 2010, p. 58). It is worth unpacking what this means.

Propositional logic provides a compositional representational system whose ontology comprises facts and whose operations are defined over atomic (i.e. unstructured) representations of such facts—namely, propositions. That is, the basic representational units are descriptions of world states, which both combine with other descriptions to yield molecular descriptions of world states and feature in formal argument patterns (like modus ponens). Propositional logic thereby produces *factored representations*, in which a represented state of the world comprises a vector of attribute values: the *attributes* (i.e. *variables*) are the propositions and the *values* are their truth-values (Russell and Norvig 2010, p. 58).

In one sense, the difference between propositional logic and probabilistic graphical models is immense. Most obviously, propositional logic cannot handle *uncertainty*: descriptions of the world are either true or false. By contrast, probabilistic graphical models define a probability distribution over world states. Relatedly, inference in such models is not deductive but statistical, capturing characteristics of inference like non-monotonic reasoning not possible in propositional logic (Pearl 1988). Nevertheless, despite these important differences, the ontology and thus expressive power of such representational systems is equivalent: namely, facts (Russell and Norvig 2010, p. 290).

The limitations of propositional logic and such factored representations more generally motivated the development of more expressive formal systems—most famously, first-order (i.e. predicate) logic (cf. Haaparanta 2009). In contrast to factored representations, the ontology of first-order logic comprises not just facts but *objects* and *relations*, thereby representing "the world as having *things* in it that are *related* to each other, not just variables with values" (Russell and Norvig 2010, p. 58). As such, first-order logic decomposes a propositional representation (e.g. Hume is a philosopher) into subjects (e.g. Hume), n-place predicates (e.g. is a philosopher), and quantifiers (e.g. there is an x such that x is Hume and x is a philosopher). The resulting gains in expressive power are famously enormous.

These expressive limitations of probabilistic graphical models have motivated many both in the tradition of Bayesian cognitive psychology (Goodman et al. 2015; Lake et al. 2015) and machine learning (Ghahramani 2015) to abandon such models as the data structures implicated in probabilistic inference in favour of *probabilistic programming languages*—that is, programming languages familiar from computer science that capture the expressive power of higher-order logics but nevertheless facilitate proba-

bilistic inference (Goodman 2013). In the domain of cognitive psychology that most concerns us, Goodman et al. (2015) explicitly argue that hierarchical graphical models are inadequate to capture the productivity of distinctively conceptual thought.

To see a concrete example of this, imagine trying to represent a simple choice model (i.e. theory of mind) with a probabilistic graphical model (Goodman et al. 2015). In *some* sense, one could capture the dependence relationships between actions, beliefs, and desires with a graph (i.e. actions are the children of beliefs and desires). The obvious problem with any such representation, however, is twofold: first, the domain of the relevant variables is *infinite*—that is, there is no *limit* on the kinds of things one can want, think, or do; second, a simple (causal) probabilistic dependency relationship does not capture the functional relationships that exist between mental states and actions in choice models—for example, some kind of utility maximization (Gerstenberg and Tenenbaum in press; Goodman et al. 2015). More generally, this example illustrates that graphical models will fail whenever the domain either requires a *richly compositional* representational system, or because the relationships between the elements in that domain cannot be captured with a simple graph-like structure. Crucially, hierarchical graphical models do not solve this problem: they introduce a computationally powerful kind of relationship between variables, but do not alter their basic expressive power (Gerstenberg and Tenenbaum in press; Goodman et al. 2015).

Insofar as predictive processing is committed to HPGMs whose structure is captured by probabilistic graphical models, then, it will be unable to account for any features of thought that are more expressive than propositional logic. Plausibly, this characterises *all* conceptual thought (Fodor 1975; Goodman et al. 2015). Even if one were to deny this, however, notice how low the bar is in the current context: all one requires is to show that *some* features of higher cognition are at least as expressive as first-order logic. This claim could not plausibly be denied.

4.3 Summary

I have argued that predictive processing is unable to capture either the generality of thought or its rich compositionality. Specifically, by restricting representations to specific locations within a network structure, it lacks the resources to accommodate either the radical *domain generality* of thought or the way in which concepts can be flexibly, systematically, and productively combined to yield such thoughts. I think it is plausible that these characteristics are not peripheral features of cognition but central to conceptual thought and reasoning in general (see Fodor 1975; Gerstenberg and Tenenbaum in press; Goodman et al. 2015; Jackendoff 2002). Nevertheless, all that the argument of this paper strictly requires is that such features are present in at least some instances of higher cognition.

Is there a way out? I conclude by identifying what I think are the two most plausible strategies for an advocate of predictive processing to pursue in light of the foregoing objections. In doing so, I hope to pave the way for constructive empirical and philosophical work in the future motivated by the critical arguments of this paper. Given space constraints, I simply identify the strategies, their attractions, and the steepest obstacles they confront, without elaborating on these points in detail.

5 Conclusion: two strategies for predictive processing

First, in Sect. 2 I argued that predictive processing has two parts: a vision of the brain as an organ for prediction error minimization, and a specific proposal concerning *how* the brain minimizes prediction error involving hierarchical predictive coding. The simplest strategy for an advocate of predictive processing to pursue in light of the foregoing objections is simply to either abandon this second part or restrict its domain of application—to dissociate the prediction error minimization framework from the specific "process theory" I have considered in this paper.

There are at least three considerations that recommend this strategy. First, I noted briefly in Sect. 2 that the motivation for conceptualising the brain as an organ for prediction error minimization has deep links to considerations from statistical physics and theoretical biology (cf. Friston 2010; Hohwy 2015), and the "process theory" I have focused on in this paper is plausibly just one (among many possible) suggestions for *how* the brain minimizes long-run prediction error. As such, the cognitive architecture I have addressed does not seem an essential characteristic of the prediction error minimization framework. One might even go further and argue that the most efficient means for organisms in our predicament to minimize prediction error would be to induce a representational system with the characteristics I have identified, although I will not pursue that suggestion here. Second, there is an enormous amount of excellent work that falls under the broad rubric of the Bayesian brain hypothesis and "prediction-and-generative-model-based architectures" (Clark 2016, p. 299) that is not specifically tied to the cognitive architecture I have addressed. Finally, and relatedly, I noted above (Sect. 4.2) a recent emergence of fascinating work in cognitive psychology and machine learning that focuses on *probabilistic programming languages* and is explicitly designed to answer the kinds of worries I have raised for predictive processing here within the confines of a Bayesian, generative-model-based approach to cognition (see Goodman et al. 2015). Given this, there is no shortage of excellent work an advocate of predictive processing might draw from were she to abandon her commitment to the cognitive architecture outlined in Sect. 2 as the whole truth about the brain.

Despite these attractions, I think this strategy pays a steep price: explanatory power. The fewer commitments that predictive processing has, the less it can genuinely explain. As noted in Sect. 2, there is plausibly *nothing* that could falsify the mere claim that the brain is Bayesian, and one might make a similar point with respect to the prediction error minimization framework. Short of substantive commitments concerning the actual information-processing architecture in the brain, it is difficult to see how such frameworks on their own could genuinely illuminate concrete psychological phenomena. At best, they would *constrain* the space of possible explanations and provide the theoretical resources for advancing specific explanatory models of such phenomena. In itself, this is no objection. Nevertheless, it does undermine the widespread view that predictive processing *itself* constitutes a "complete framework… for *explaining* perception, cognition, and action" (Seth 2015, p. 1, my emphasis).

A second strategy is more radical: it accepts the psychological phenomena I identified in S4 above, but associates them with the distinctive contribution of natural language and *public* combinatorial symbol systems more generally. This would relate

predictive processing to a long tradition in both psychology and philosophy (see Bermúdez 2005, Ch. 10) that attributes many of the idiosyncrasies of higher cognition in humans not to a qualitatively distinct neural architecture from other mammals, but to the *public* symbol systems that we are plausibly uniquely exposed to. Such a possibility is suggested by the latter chapters of Clark's (2016) recent monograph on predictive processing, in which he attributes the seemingly novel cognitive capacities we exhibit not just to a deeper and more richly structured neocortex, but also to "the influence of flexible structured symbolic language… and an almost obsessive drive to engage in shared cultural practices" (Clark 2016, p. 276; see Williams 2018). On this view, "human thought… inherit[s] such systematicity as it displays from the *grammatical structure of human language itself*" (Clark 2000, p. 77, my emphasis).

Again, this idea has numerous attractions. First, it is much less plausible that other mammals exhibit the psychological capacities I identified above, despite the seemingly profound anatomical and computational similarities between our neocortices. This encourages the idea that what differentiates human cognition from that of other mammals is largely a matter of the novel environments we inhabit—with its structured social practices and public symbol systems—rather than the basic kind of cortical computation (see Churchland 2012). Second, the characteristics I identified *do* appear to be strongly coupled to natural language: it is difficult to imagine how one could think about electrons or postmodernism, for example, without the conceptual scheme and language endowed by one's culture. Third, it is not difficult to find speculative accounts of why the evolutionary pressures on communication might have led to the development of a representational system in language qualitatively different in its character from the system that underlies neural computation (Bermúdez 2005, Ch. 10). As such, the characteristics of these representational systems might very well interact in ways to produce seemingly novel kinds of cognition. Finally, and perhaps most importantly, this strategy might provide an advocate of predictive processing with the resources to hold on to the basic cognitive architecture outlined in Sect. 2.

As with before, however, this strategy confronts significant challenges. First, it is not obvious that it is true. Of course, to what extent pre-linguistic cognition exhibits the generality and rich compositionality I have identified is a difficult and strongly contested empirical question, and I cannot settle it here. Two reasons for thinking it does, however, are: first, the seemingly rich structure of "intuitive theories" of domains like physics and psychology that appear to be pre-linguistic, central to commonsense reasoning, and yet richly compositional (Gerstenberg and Tenenbaum in press); and second, the fact that we can *learn* and *understand* natural languages—a fact that is difficult to understand without positing a representational system at least as expressive as language itself (Pylyshyn and Fodor 2015). I will not take a stand on such debates here. By contrast, this strategy must come down hard on one side of them.

The second and related challenge is that it is not obvious how the suggestion would help even if it were true. That is, even if one grants that the features of conceptual thought I have outlined in this paper are somehow intimately related to public representational systems, one would still have to explain *how* human thought might "inherit" these features from such systems. If public representations literally transform the brain's information-processing architecture, this strategy would collapse into the first strategy outlined above—namely, a (partial) abandonment of the idea that the

information-processing architecture outlined in Sect. 2 provides the whole truth about cognitive representation within the brain. If one wants to preserve that information-processing architecture, then, it seems that one would have to endorse some form of *externalism* about the vehicles of public representational systems like natural language.[13] This does not seem like an attractive option, however.

However these issues play out, I hope that I have done enough to show that they *must* play out if predictive processing is to have any hope of accounting for conceptual thought.

Acknowledgements This work was supported by the Arts and Humanities Research Council. I would like to thank Jakob Hohwy for helpful comments, discussion, and criticism, and the members of the Cognition & Philosophy Laboratory at Monash University, especially Andrew Corcoran, Stephen Gadsby, Julian Matthews, and Kelsey Palghat. I would also like to thank two anonymous reviewers for helpful comments and suggestions on an earlier version of this manuscript.

References

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

Bermúdez, J. (2005). *Philosophy of psychology: A contemporary introduction*. London: Routledge.

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology, 76,* 198–211. https://doi.org/10.1016/j.jmp.2015.11.003.

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

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

Churchland, P. (2012). *Plato's camera*. Cambridge, Mass: MIT Press.

Clark, A. (2000). *Mindware* (1st ed.). New York: Oxford University Press.

Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind, 121*(483), 753–771. https://doi.org/10.1093/mind/fzs106.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(03), 181–204. https://doi.org/10.1017/s0140525x12000477.

Clark, A. (2015). Predicting peace: The end of the representation wars—A reply to Michael Madary. In T. Metzinger & J. M. Windt (Eds.), *Open MIND: 7(R)*. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570979.

Clark, A. (2016). *Surfing uncertainty*. Oxford: Oxford University Press.

Colombo, M., & Wright, C. (2016). Explanatory pluralism: An unrewarding prediction error for free energy theorists. *Brain and Cognition, 112,* 3–12.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Cambridge: The MIT Press.

Dayan, P., Hinton, G., Neal, R., & Zemel, R. (1995). The helmholtz machine. *Neural Computation, 7*(5), 889–904. https://doi.org/10.1162/neco.1995.7.5.889.

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

Evans, G. (1982). *The varieties of reference*. Oxford: Oxford University Press.

[13] I thank an anonymous reviewer for raising this point.

Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2010.00215.
Firestone, C., & Scholl, B. (2015). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*. https://doi.org/10.1017/s0140525x15000965.
Fletcher, P., & Frith, C. (2008). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience, 10*(1), 48–58. https://doi.org/10.1038/nrn2536.
Fodor, J. (1975). *The language of thought*. Cambridge: Harvard University Press.
Fodor, J. (2011). Fire the press secretary. [Review of the book *Why everyone (else) is a hypocrite: Evolution and the modular mind*.] *London Review of Books*, *33*(9), 24–25. Retrieved from https://www.lrb.co.uk/v33/n09/jerry-fodor/fire-the-press-secretary.
Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5.
Friston, K. (2002). Beyond phrenology: What can neuroimaging tell us about distributed circuitry? *Annual Review of Neuroscience, 25*(1), 221–250. https://doi.org/10.1146/annurev.neuro.25.112701.142846.
Friston, K. (2003). Learning and inference in the brain. *Neural Networks, 16*(9), 1325–1352. https://doi.org/10.1016/j.neunet.2003.06.005.
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions Of The Royal Society B: Biological Sciences, 360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787.
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017a). Active inference: A process theory. *Neural Computation, 29*(1), 1–49. https://doi.org/10.1162/neco_a_00912.
Friston, K., Parr, T., & de Vries, B. (2017b). The graphical brain: Belief propagation and active inference. *Network Neuroscience*. https://doi.org/10.1162/netn_a_00018.
Geisler, W., & Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuroscience, 5*(6), 508–510. https://doi.org/10.1038/nn0602-508.
George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS Computational Biology, 5*(10), e1000532. https://doi.org/10.1371/journal.pcbi.1000532.
Gerstenberg, T., & Tenenbaum, J. B. (in press). Intuitive theories. In M. Waldman (Ed.), *Oxford handbook of causal reasoning*. Oxford University Press.
Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452–459. https://doi.org/10.1038/nature14541.
Glymour, C. (2002). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, Mass: MIT Press.
Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning*. Cambridge, Mass: The MIT Press.
Goodman, N. (2013). The principles and practice of probabilistic programming. *ACM SIGPLAN Notices, 48*(1), 399–402. https://doi.org/10.1145/2480359.2429117.
Goodman, N., Tenenbaum, J., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–654). Cambridge, MA: MIT Press.
Gopnik, A., & Wellman, H. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin, 138*(6), 1085–1108. https://doi.org/10.1037/a0028044.
Gordon, N., Koenig-Robert, R., Tsuchiya, N., van Boxtel, J., & Hohwy, J. (2017). Neural markers of predictive coding under perceptual uncertainty revealed with Hierarchical Frequency Tagging. *Elife*. https://doi.org/10.7554/elife.22749.
Griffiths, T., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004.
Haaparanta, L. (2009). *The development of modern logic*. Oxford: Oxford University Press.
Hawkins, J., & Blakeslee, S. (2005). *On intelligence*. New York: Henry Holt and Company.
Helmholtz, H. V. (1867). *Handbuch der physiologischen optik*. Leipzig: Leopold Voss.
Henson, R., & Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus, 20*(11), 1315–1326. https://doi.org/10.1002/hipo.20857.
Hinton, G. (2011). Machine learning for neuroscience. *Neural Systems and Circuits, 1*(1), 12. https://doi.org/10.1186/2042-1001-1-12.

Hobson, J., & Friston, K. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology, 98*(1), 82–98. https://doi.org/10.1016/j.pneurobio.2012.05.003.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2012.00096.

Hohwy, J. (2013). *The predictive mind*. Oxford: Oxford University Press.

Hohwy, J. (2014). The self-evidencing brain. *Noûs, 50*(2), 259–285. https://doi.org/10.1111/nous.12062.

Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.), *Open MIND: 19(T)*. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570016.

Hohwy, J. (2017). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness and Cognition, 47,* 75–85. https://doi.org/10.1016/j.concog.2016.09.004.

Huang, G. (2008). Is this a unified theory of the brain? *New Scientist, 2658,* 30–33.

Jackendoff, R. (2002). *Foundations of language*. New York, NY: Oxford University Press.

Jacobs, R., & Kruschke, J. (2010). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(1), 8–21. https://doi.org/10.1002/wcs.80.

Jones, M., & Love, B. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34*(04), 169–188. https://doi.org/10.1017/s0140525x10003134.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current Opinion in Neurobiology, 13*(2), 150–158. https://doi.org/10.1016/s0959-4388(03)00042-4.

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

Kiefer, A., & Hohwy, J. (in press). Representation in the prediction error minimization framework. In: J. Symons, P. Calvo, & S. Robins (Eds.), *Routledge handbook to the philosophy of psychology*. Routledge.

Kilner, J., Friston, K., & Frith, C. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing, 8*(3), 159–166. https://doi.org/10.1007/s10339-007-0170-2.

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

Lake, B., Salakhutdinov, R., & Tenenbaum, J. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050.

Lake, B., Ullman, T., Tenenbaum, J., & Gershman, S. (2016). Building machines That learn and think like people. *Behavioral and Brain Sciences*. https://doi.org/10.1017/s0140525x16001837.

Lee, T., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, 20*(7), 1434. https://doi.org/10.1364/josaa.20.001434.

Marcus, G., Marblestone, A., & Dean, T. (2014). The atoms of neural computation. *Science, 346*(6209), 551–552. https://doi.org/10.1126/science.1261661.

Marr, D. (1982). *Vision: A computational approach*. San Francisco, CA: Freeman & Co.

Mathys, C., Lomakina, E., Daunizeau, J., Iglesias, S., Brodersen, K., Friston, K., et al. (2014). Uncertainty in perception and the Hierarchical Gaussian Filter. *Frontiers in Human Neuroscience*. https://doi.org/10.3389/fnhum.2014.00825.

Mountcastle, V. (1978). An organizing principle for cerebral function: The unit model and the distributed system. In G. M. Edelman & V. B. Mountcastle (Eds.), *The mindful brain*. Cambridge MA: MIT Press.

Newell, A., & Simon, H. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126. https://doi.org/10.1145/360018.360022.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco: Elsevier Science.

Pearl, J. (2000). *Causality*. Cambridge: Cambridge University Press.

Penny, W. (2012a). Bayesian models of brain and behaviour. *ISRN Biomathematics, 2012,* 1–19. https://doi.org/10.5402/2012/785791.

Penny, W. (2012b). Bayesian models of brain and behaviour. *ISRN Biomathematics, 2012,* 1–19. https://doi.org/10.5402/2012/785791.

Pezzulo, G. (2017). Tracing the roots of cognition in predictive processing. In T. Metzinger & W. Wiese (Eds.), *Philosophy and predictive processing: 20*. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958573215.

Prakash, J., Shashikumar, R., Bhat, P., Srivastava, K., Nath, S., & Rajendran, A. (2012). Delusional parasitosis: Worms of the mind. *Industrial Psychiatry Journal, 21*(1), 72. https://doi.org/10.4103/0972-6748.110958.

Pylyshyn, Z., & Fodor, J. (2015). *Minds without meanings: An essay on the content of concepts*. Cambridge: The MIT Press.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, *60(1)*: 20–43. JSTOR 2181906. https://doi.org/10.2307/2181906. Reprinted in his 1953 *from a logical point of view*. Harvard University Press.

Rao, R. P., & Ballard, D. H. (1998). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience, 2,* 79–87.

Rescorla, M. (2013). Bayesian perceptual psychology. In M. Matthen (Ed.), *Oxford handbook of the philosophy of perception*. Oxford: Oxford University Press.

Rolls, E. (2016). *Cerebral cortex*. Oxford: Oxford University Press.

Roskies, A., & Wood, C. (2017). Catching the prediction wave in brain science. *Analysis*. https://doi.org/10.1093/analys/anx083.

Russell, S. (2015). Recent developments in unifying logic and probability. *Communications of the ACM, 58*(7), 88–97. https://doi.org/10.1145/2699411.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). London: Pearson.

Ryder, D. (forthcoming) *Models in the brain*.

Seth, A. K. (2015). The cybernetic bayesian brain—From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. M. Windt (Eds.), *Open MIND: 35(T)*. Frankfurt am Main: MIND Group. https://doi.org/10.15502/9783958570108.

Sloman, S. (2005). *Causal models*. New York: Oxford University Press.

Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788.

Thagard, P. (2011). Cognitive architectures. In K. Frankish & W. Ramsay (Eds.), *The Cambridge handbook of cognitive science*. Cambridge: Cambridge University Press.

Vance, J. (2015). Review of the predictive mind. *Notre Dame Philosophical Reviews*.

Weilnhammer, V., Sterzer, P., Hesselmann, G., & Schmack, K. (2017). A predictive-coding account of multistable perception. *Journal of Vision, 17*(10), 580. https://doi.org/10.1167/17.10.580.

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

Williams, D. (2018). Pragmatism and the predictive mind. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-017-9556-5.

Williams, D. (forthcoming) Hierarchical Bayesian models of delusion. *Consciousness and Cognition*.

Williams, D., & Colling, L. (2017). From symbols to icons: The return of resemblance in the cognitive neuroscience revolution. *Synthese*. https://doi.org/10.1007/s11229-017-1578-6.