Consciousness and Cognition xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/concog

Review article

Hierarchical Bayesian models of delusion

Daniel Williams*

Faculty of Philosophy, Trinity Hall, University of Cambridge, United Kingdom
Cognition and Philosophy Laboratory, Monash University, Australia

ARTICLE INFO

Keywords:
Delusions
Psychosis
Bayesian brain hypothesis
Predictive processing
Predictive coding
Two-factor
Rationality
Bayesian just-so stories
Optimality
Argumentative theory of reasoning
Confirmation bias
Motivated reasoning
Backfire effect

1. Introduction

> "In the normally functioning brain information from different sources is combined in a *statistically optimal manner*. The mechanism for achieving this is well captured in a Bayesian framework".
>
> (Frith and Friston, 2013, p. 5)

> "In our model, hierarchy is key".
>
> (Corlett, Honey, & Fletcher, 2016, p. 1148)

What is the probability that you will become the Emperor of Antarctica? That you are the left foot of God? That you are the victim of a conspiracy perpetrated by the Pope and the CIA? Most people would assign an extremely low probability to such propositions, if they were to consider them at all. Famously, however, the mathematician and Nobel laureate John Nash believed all three to be true to be true at various points in his life (Capps, 2004; Coltheart, 2007, p.1057). Such convictions are paradigmatic examples of *delusional beliefs*. They come in a wide variety of forms and arise from a comparably diverse range of underlying causes—in Nash's case, his long battle with schizophrenia. Despite substantial disagreement concerning how best to define delusional beliefs—indeed, whether they should properly be characterised as a species of *belief* at all—there is a widespread consensus that they comprise a genuine psychological kind in need of explanation (Bortolotti, 2010; Coltheart, 2007; Gerrans, 2014).[1] Why do people form such delusions? And why do they retain them in the face of seemingly incontrovertible evidence against them?

* Address: Faculty of Philosophy, Trinity Hall, University of Cambridge, United Kingdom.
E-mail address: dw473@cam.ac.uk.

[1] I will assume a "doxastic" (i.e. belief-based) understanding of delusions throughout, because this is also assumed by advocates of the hierarchical Bayesian models that I focus on. (Although see Gerrans (2014) for an important challenge to this doxastic approach that draws on predictive coding).

https://doi.org/10.1016/j.concog.2018.03.003
Received 20 December 2017; Received in revised form 20 February 2018; Accepted 3 March 2018

Researchers in the emerging field of computational psychiatry have recently sought to answer these questions by appeal to dysfunctions in a process of hierarchical Bayesian inference alleged to underlie perception and belief fixation[2] in the healthy (i.e. neurotypical) population (Adams, Stephan, Brown, Frith, & Friston, 2013; Corlett, Taylor, Wang, Fletcher, & Krystal, 2010; Fletcher and Frith, 2009; Frith and Friston, 2013; Schmack et al., 2013). These hierarchical Bayesian models have been motivated in large part by *predictive processing* (also known as *hierarchical predictive coding*), an influential theory in cognitive and computational neuroscience that models the brain as a "probabilistic prediction machine" striving to minimize the mismatch between internally generated *predictions* of its sensory inputs and the sensory inputs themselves (see Clark, 2013, 2016; Friston, 2010; Friston, FitzGerald, Rigoli, Schwartenbeck, and Pezzulo, 2017a; Hohwy, 2013). As Griffin and Fletcher, (2017, p. 265) note, this

> "growing understanding of the brain as an organ of predictive inference has been central to establishing computational psychiatry as a framework for understanding how alterations in brain processes can drive the emergence of high-level psychiatric symptoms."

In this paper I argue that these hierarchical Bayesian models of delusion are significantly less promising than is widely believed. Specifically, I raise challenges for the two core theoretical components of such models that have not been sufficiently addressed—or for the most part even recognised—in the literature. First, the characteristic that is supposed to most sharply distinguish hierarchical Bayesian models from previous approaches to delusions is their abandonment of the traditional distinction between perception and cognition in favour of a unified inferential hierarchy with bi-directional message-passing. Standard ways of characterising this inferential hierarchy, however, are inconsistent with the range of phenomena that delusions can represent. Second, there is little evidence that belief fixation in the *healthy* population is Bayesian, and a seeming abundance of evidence that it is not. As such, attempts to model delusions in terms of dysfunctions in a process of Bayesian inference are of dubious theoretical value.

I structure the paper as follows. In Section 2 I provide a brief overview of hierarchical Bayesian models of cortical information processing, focusing on predictive processing. In Section 3 I explain how these hierarchical Bayesian models have been used to illuminate the formation and retention of delusional beliefs. Sections 4 and 5 then raise challenges for the two core theoretical components of such models: the concept of an inferential "hierarchy" (Section 4) and the commitment to a Bayesian account of belief fixation (Section 5). I conclude in Section 6 by summarising the foregoing argument and extracting a lesson for the increasingly influential field of computational psychiatry: that it would benefit from an abandonment of global theories of brain function and optimality models of cognition in favour of a much more substantial engagement with research from other fields, especially evolutionary biology and cognitive and social psychology.

2. Predictive coding and hierarchical Bayesian inference

2.1. The Bayesian brain

Bayes's theorem states that:

$$p(h/e) = p(e/h)p(h)/p(e)$$

Under a set of plausible assumptions (see Zellner, 1988), this theorem describes the optimal calculus for belief updating under conditions of uncertainty. Specifically, if e is a piece of evidence and h is a possible hypothesis for explaining this evidence, Bayes's theorem states that the probability of the hypothesis given the evidence p(h/e) is proportional to its *likelihood* p(e/h)—how well the hypothesis predicts the evidence—weighted by its *prior probability* p(h)—the probability of the hypothesis considered independently of the evidence. Bayes's rule then states that one should update one's beliefs in accordance with this formula.

Recent years have seen an "explosion in research applying Bayesian models to cognitive phenomena" (Chater, Oaksford, Hahn, & Heit, 2010, p. 811; see Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Oaksford and Chater, 2007). This "revolution" (Hahn, 2014) has been driven by at least two important factors: first, a growing recognition that across many psychological domains the chief problem that the brain confronts is inference and decision-making under *uncertainty* (Tenenbaum et al., 2011); second, a growing appreciation of the way in which Bayesian statistics and decision theory can be used to capture the solutions to such problems in mathematically precise and empirically illuminating ways (Williams, forthcoming a).

Although perceptual and sensorimotor psychology are the areas where Bayesian models have won the most widespread acceptance (Chater et al., 2010, p. 820), they have been applied to an enormous and seemingly ever-increasing range of cognitive phenomena: categorization, causal learning and inference, language processing, abstract reasoning, and more (Chater et al., 2010; Tenenbaum et al., 2011). In some cases, these models are advanced merely as normative "ideal observer" models intended to capture optimal performance in cognitive tasks without any pretence to descriptive adequacy. In this paper, however, I focus exclusively on Bayesian models intended as descriptive accounts of actual cognitive mechanisms. The success of such accounts has inspired the "Bayesian brain hypothesis" (Knill and Pouget, 2004), the thesis that some (or even all—see below) information processing in the brain conforms to Bayesian principles.

The Bayesian brain hypothesis confronts at least two significant challenges. First, *exact* Bayesian inference is slow and often computationally intractable. As such, there is extensive research in statistics and artificial intelligence focused on developing

[2] For convenience, I will use the term "belief fixation" to subsume the mechanisms underlying both belief *formation* and belief *evaluation* should they be different (see Fodor 1983).

algorithms for *approximate* Bayesian inference, the most prominent of which are sampling and variational methods (Penny, 2012). Second, researchers must explain *how* their chosen approximation algorithms get executed in the brain's neural networks. In other words, descriptive realism about Bayesian cognitive science requires researchers to identify plausible theories at both the "algorithmic" and "implementational" levels of Marr's (1980) three-tiered schema for computational explanation.

There is a large amount of research tackling these problems. Probably the most influential of such frameworks, however, is *predictive processing* (also known as *hierarchical predictive coding* (see Clark, 2013; Friston, 2005; Friston, 2017a; Hohwy, 2013; Rao and Ballard, 1999; Seth, 2015). An enormous amount has been written about predictive processing both in the scientific and philosophical literature, and there are excellent overviews elsewhere (Clark, 2016; Hohwy, 2013; Seth, 2015). Here I focus on just three things: first, how Bayesian inference can be formalised in terms of precision-weighted prediction error minimization; second, how this process of precision-weighted prediction error minimization can be made hierarchical; and third, how predictive coding can be exploited as a message-passing strategy for implementing this process of approximate Bayesian inference in the neocortex.

2.2. Predictive processing

First, then, if one assumes Gaussian distributions (i.e. density functions), one can formalise Bayesian inference in the following way. One compares the mean value *m* of one's prior distribution with the mean value *e* of the evidence to compute a *prediction error* (i.e. the distance between these two values). The prediction error thus varies inversely with the *likelihood* of the hypothesis: the better that *m* predicts the evidence, the less prediction error it generates. The amount by which one updates one's prior in response to this prediction error therefore provides the posterior in Bayesian inference. According to Bayes's theorem, this should be calculated relative to the uncertainty of one's priors and likelihoods. With Gaussian distributions, one can exploit the *precision* (the inverse of the variance) of the two distributions to calculate this property. Specifically, the more precise one's priors are relative to the evidence, the less one updates *m* in response to the prediction error it generates—and vice versa. The relative precision of the two distributions thus determines the *Bayesian learning rate*: the amount by which priors are updated as a function of prediction errors (Hohwy, 2017).

In a world in which the evidence is a linear function of its causes, recursively employing this simple process of Bayesian inference would result in the optimal minimization of prediction error and arbitrarily precise priors as one's predictions come to accurately reflect how things are (Hohwy, 2017). In the case of real-world perception, however, this would be untenable. The sensory signals received by the brain are generated by a dynamic and volatile environment with a rich, hierarchical causal structure (Friston, Rosch, Parr, Price, and Bowman, 2017b; Hohwy, 2013). To effectively minimize prediction error under such conditions, an inferential system must be responsive to this structure. Specifically, it must reliably *invert* the structured causal process by which sensory input is generated, deconvolving hidden causes at different spatiotemporal scales (Friston, 2008; Williams, 2017; Williams and Colling, 2017). Rather than matching a single hypothesis space H against the evidence, such a system thus builds a *hierarchy* of hypotheses ranging over different levels of spatiotemporal scale and abstraction, where hypotheses at one level L + 1 provide the priors for the level below L and the evidence against which prediction errors are computed for the level above L + 2 (Friston, 2008; Lee and Mumford, 2003; see Section 4 for more on this concept of a hierarchy). With a hierarchical model of this kind in place, a Bayesian system can exploit hypotheses at higher levels to flexibly adjust the learning rate in context-sensitive ways (Hohwy, 2017; Mathys et al., 2012, 2014). For example, such a system might expect highly precise visual evidence in broad daylight, and highly noisy evidence during a fog, and adjust its learning rate accordingly (Clark, 2016, p. 58).

So far I have briefly described how hierarchical Bayesian inference can be formalised in terms of a process of precision-weighted prediction error minimization. How might this process be implemented in neural circuitry? According to predictive processing, through *hierarchical predictive coding*.

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). In predictive processing, this encoding strategy captures the fundamental nature of message passing in the hierarchically structured neocortex (Bastos et al., 2012; Clark, 2016; Friston, 2005; Seth, 2015). Specifically, its proposal is that backwards synaptic connections from higher cortical (e.g. frontal or temporal) areas match predictions against activity at lower levels, and that only the unpredicted elements of the signal—the *prediction errors*—are fed forward. These prediction errors thus provide the brain with an *internally accessible* quantity to minimize. According to predictive processing, it is the brain's ceaseless efforts to minimize this quantity that both installs a rich, hierarchically structured generative model of the environment and then updates the parameters of this model in real time. At the "implementational level," an influential proposal in the literature is that predictions and prediction errors are carried by deep and superficial cortical pyramidal cells respectively, with precision-weighting occurring through alterations to the "postsynaptic gain" of such superficial pyramidal cells. In turn, the mechanism responsible for altering the postsynaptic gain is at least in part the action of neuromodulators such as dopamine, serotonin, and acetylcholine (a detail that will be important below) (Adams et al., 2013).

For some, this process of hierarchical predictive coding merely captures the information-processing strategies of *some* regions within the brain—for example, the visual cortex (Rao and Ballard, 1999). Others have extended it to a framework for understanding *all* cortical information processing (Bastos et al., 2012; Clark, 2016; Friston, 2005; Hohwy, 2013). Specifically, it has been proposed as a global theory of brain function, a putative explanation of "perception and action and everything mental in between" (Hohwy, 2013, p. 1). On this view—which bears a close theoretical relationship to the "free-energy principle" as described by Karl Friston (2010)—"the brain is an organ *for* prediction error minimization" (Hohwy, 2014, 259, my emphasis). That is, *all* neural functioning is orchestrated around the long-term minimization of prediction error (for excellent overviews, see Clark, 2016; Friston et al., 2017a; Hohwy, 2013; Seth, 2015).

Although the foregoing overview has been skeletal, and has neglected many important facets of predictive processing, it will

nevertheless suffice for the aim of this paper—namely, to evaluate its capacity to explain the formation and retention of delusions, to which I now turn.

3. Hierarchical Bayesian models of delusion

A central aim of psychiatry is to explain psychopathologies and the symptoms they give rise to in terms of *dysfunctions* in the mechanisms underlying psychological processes in the healthy population (Murphy, 2006).[3] Given the widespread view that our best explanations of cognitive mechanisms treat them as computational mechanisms involved in information processing, the recent field of *computational psychiatry* exploits such computational models to explain the psychiatric and neurological disorders at the root of mental illness (Friston, Stephan, Montague, & Dolan, 2014; Montague, Dolan, Friston, & Dayan, 2012). As Teufel and Fletcher (2016, 2601) put it:

> "In the modern brain sciences, the most powerful models are computational in nature…The relatively novel approach [of computational psychiatry] harnesses these powerful computational models and applies them to psychiatric and neurological disorders."

Given this explanatory strategy, recent years have seen an outpouring of work that draws on the Bayesian brain hypothesis and predictive processing to illuminate psychopathologies such as autism (Lawson, Rees, & Friston, 2014), anxiety disorders (Seth, Suzuki, & Critchley, 2012), and depression (Chekroud, 2015). In this section I focus on recent efforts to explain the formation and retention of *delusional beliefs* in terms of dysfunctions in the process of hierarchical Bayesian inference introduced in Section 2. Because my aim is to identify structural problems with this approach, readers should look elsewhere for detailed overviews of specific applications (see, for example: Adams et al., 2013; Corlett et al., 2016). Specifically, I focus on just two things: first, the appeal to dysfunctions in *precision-weighting* as the core pathology underlying the emergence and retention of delusional beliefs; and second, the way in which such hierarchical Bayesian models are supposed to undermine explanations of delusions that distinguish perception from cognition.

First, however, it will be useful to begin with a brief review of traditional theoretical approaches to delusions.

3.1. Delusional beliefs and the two-factor framework

Traditional theoretical approaches to delusions have typically focused either on a dysfunction in the agent's perceptual systems or a deficit in her reasoning capacities.[4] In the first case, delusional beliefs are explained as a reasonable response to anomalous experiences (Maher, 1974). For example, if a patient can hear her own thoughts spoken aloud, an appropriate inference might be that other people can hear them as well (Fletcher and Frith, 2009, p. 49). In the second case, the malfunction is claimed to reside in the agent's reasoning capacities or the mechanisms underlying hypothesis evaluation more generally (Hemsley and Garety, 1986). Roughly, the agent's perceptual experiences are normal but the processes underlying the conclusions she draws from such experiences are damaged or otherwise malfunctioning.

Recently, Coltheart (2007) has argued that any tenable explanation of delusions must in fact appeal to deficits in *both* factors. Specifically, Coltheart identifies two deficits necessary for the production and retention of delusions (taken from Bortolotti, 2016):

1. The first is a neuropsychological impairment that presents the patient with new (false) data, and the delusional belief is one which if true would best explain these data.
2. The second is a neuropsychological impairment of a belief evaluation system, which prevents the patient from rejecting the newly formed belief even though there is much evidence against it.

Consider Capgras delusion, for example. In this condition, the individual forms the delusional belief that someone close to her (usually a spouse or a loved one) has been replaced by a visually indistinguishable (or nearly indistinguishable) imposter. An influential explanation of this delusion advanced by Ellis and Young (1990) contends that it arises from damage to the part of the brain responsible for delivering the suite of autonomic responses an agent usually receives when visually recognising a face. Under such conditions, the agent will visually recognise the face but fail to receive the usual autonomic signals associated with that face. As such, a possible explanation arises: *this person is not really my spouse. They are an imposter.*

As Coltheart (2007, 1048) points out, however, an appeal to this perceptual deficit is not sufficient to explain the delusional belief. First, there seem to be much *better* explanations available for the relevant evidence—for example: *I have a neurological disorder*. Second, there are people who have the relevant neurological impairment who do *not* form the delusion. As such, a second impairment or deficit is necessary: an impairment in the agent's "belief evaluation system" (Coltheart, 2007). Specifically, a deficit in this system—thought to arise from frontal right hemisphere damage—is responsible for the agent retaining the delusional belief in the face of seemingly incontrovertible evidence against it.

The two-factor framework has so far focused predominantly on *monothematic* delusions focused on one specific theme (e.g.

[3] Technically, a widespread assumption is that such dysfunctions must be *harmful*, a qualification that I will ignore throughout (see Murphy, 2006).

[4] By "traditional" here, I refer to work in the tradition of cognitive neuropsychology (Coltheart, 2007; Corlett and Fletcher, 2015), reflecting the orthodox distinction between perception and cognition in both cognitive psychology (Firestone and Scholl, 2015) and philosophy (Fodor, 1983).

Capgras delusion), although it has recently been tentatively extended to the case of polythematic delusions as they arise in conditions such as schizophrenia as well, in which the individual forms delusional beliefs about a range of themes only disparately connected to one another (Coltheart, 2013). In general, there is a broad consensus that the framework has been most successful in identifying damages or dysfunctions in the *first* factor. As even its proponents acknowledge, the specification of the *second* factor is extremely coarse-grained at the physiological level ("frontal right hemisphere damage") and "too underspecified" (Coltheart, 2017, p.3) at the algorithmic level (a deficit in a "belief evaluation system") to yield concrete testable theories or mechanistic models. Specifically, although there has been a good deal of speculation on what the second deficit amounts to (Coltheart, Menzies, & Sutton, 2010; Davies and Egan, 2013; McKay, 2012), this speculation has so far taken place at a high level of abstraction that is largely if not wholly unconstrained by algorithmic and implementational details in the brain.

For our purposes, what is crucial about the framework is the functional distinction[5] it draws between the mechanisms underlying perception and those underlying belief fixation. As Ross, McKay, Coltheart, and Langdon (2016, p. 47) put it:

> "All two-factor accounts are predicated on a conceptual distinction (and empirical dissociation) between perception and cognition: abnormal perception as the first factor and a cognitive belief evaluation deficit as the second factor."

It is this feature of the framework that is the chief target of hierarchical Bayesian models of delusion, to which I now turn.

3.2. Hierarchical Bayesian models of delusion

In the context of theoretical approaches to delusions, the most distinctive feature of hierarchical Bayesian models is their advocacy of a *single deficit* framework (Fletcher and Frith, 2009).[6] Rather than locating this deficit in either the agent's perceptual systems or reasoning capacities, however, proponents of hierarchical Bayesian models disavow the very distinction between perception and cognition as a relic of folk psychology with no place in mature cognitive neuroscience (Corlett and Fletcher, 2015; Frith and Friston, 2013). Specifically, they argue that speculation about a distinct "belief evaluation system" is neurobiologically implausible (Fletcher and Frith, 2009, p. 51) and that delusions can be accounted for in terms of a *single* dysfunction in the information-processing architecture outlined in Section 2.

> "It is possible to understand these symptoms [hallucinations and delusions] in terms of a disturbed hierarchical Bayesian framework, without recourse to separate consideration of experience and belief".
>
> (Fletcher and Frith, 2009, p. 48)

> "The unusual perceptual experiences and beliefs in psychosis can be explained *by one core atypicality*, namely a shift in the balance of Bayesian inference within a *hierarchically-organised* information processing system".
>
> (Teufel and Fletcher, 2016, p.5 , my emphasis)

What is the proposed nature of this atypicality? All advocates of hierarchical Bayesian models point to "a disturbance in error-dependent updating of inferences and beliefs about the world" (Fletcher and Frith, 2009, p. 48), although there is disagreement concerning what the relevant disturbance amounts to (Bortolotti and Miyazono, 2015, p. 642; Notredame, Pins, Deneve, & Jardri, 2014, pp. 9–10). For example, whereas some theories locate the disturbance in psychosis in excessively precise prior expectations (Friston, 2005; Chambon et al., 2011; Schmack et al., 2013), others locate it in excessively precise sensory evidence (Adams et al., 2013; Corlett et al., 2010; Frith and Friston, 2013). Because it has been by far the more influential, I will focus on the latter account here. Nevertheless, as I note below, my critique in what follows will apply to *any* approach to delusions that shares the broad architecture of hierarchical Bayesian models, and not just this specific proposal within that framework.

The core of this proposal is summarised by Adams et al. (2013, p. 1):

> "A wide range of psychotic symptoms can be explained by a failure to represent the precision of beliefs about the world… [D]elusional systems may be elaborated as a consequence of imbuing sensory evidence *with too much precision*… [T]he primary pathology here is quintessentially metacognitive in nature: in the sense that it rests on a belief…about a belief… Crucially, there is no necessary impairment in forming predictions or prediction errors – the problem lies in the way they are used to inform inference or hypotheses".
>
> (Adams et al., 2013, p. 1, my emphasis)

To understand this, recall the function of precision-estimation as introduced in Section 2. Such estimates of precision in effect determine the *learning rate* in Bayesian inference—that is, the extent to which priors (higher-level predictions) should be updated as a function of the prediction errors they generate. Highly precise prediction errors will engender sharper revisions up the processing hierarchy. A breakdown in the brain's capacity to reliably assign precision would thus amount to "failures of the very mechanisms whose task is to estimate the reliability of our own information sources" (Clark, 2016, p. 205).

[5] By "functional distinction" I mean to imply both that perception and cognition are subserved by different underlying systems or mechanisms, and that these systems are individuated (and so distinguished) by their *functional* properties (their effects, or the capacities they produce), not their physical properties (see Firestone and Scholl, 2015; Fodor, 1983).

[6] Importantly, much of this work has focused specifically on schizophrenia (e.g. Fletcher and Frith, 2009) rather than delusions *in general*, although proponents of hierarchical Bayesian models have sought to extend the framework to this more general explanandum (e.g. Adams et al., 2013; Corlett et al., 2010; Corlett and Fletcher, 2015).

A useful analogy here is with classical statistical inference (see Adams et al., 2013; Hohwy, 2013, pp. 64–66). Imagine comparing the mean of some data against the null hypothesis that the mean is zero. The difference between these two values provides a *prediction error*. This prediction error provides evidence against the null hypothesis. To quantify *how much* evidence, however, one must factor in the *precision* of the prediction error. If the data are highly variable (low precision), one should not reject the null hypothesis. The prediction error likely just reflects noise. By contrast, if the precision-weighted prediction error is sufficiently large, one should reject the null hypothesis. As this example illustrates, dramatic failures of inference can be engendered by failures of a kind of *second-order statistics*—that is, errors not in the process of comparing predictions against evidence, but in the process of weighting the resultant prediction errors according to their reliability. Failures in this latter calculation could very easily be the difference between a new drug being considered safe or harmful for the general population.

Advocates of predictive processing models of delusion appeal to this deficit as the primary pathology underlying the emergence of delusional beliefs in conditions such as schizophrenia (Adams et al., 2013; Corlett et al., 2016; Frith and Friston, 2013). Specifically, the proposal is that malfunctions in the process of precision-weighting result in the reliability of sensory evidence being over-estimated relative to "higher-level" prior beliefs. This results in persistent and highly weighted prediction errors signalling to the agent that her beliefs about the world are mistaken, engendering sharp revisions in her model of the world. Due to the bi-directional nature of message passing within the predictive processing architecture, however, these revised predictions are in turn fed back to influence the interpretation of incoming sensory evidence. Worse, because highly precise prediction errors require updates not just in inference but also *learning*—the construction of the very models from which those inferences arise—they can ultimately produce sweeping revisions in the agent's understanding of her world, including the scope of epistemic possibility.

Frith and Friston (2013, p. 12) quote a psychologist detailing his own experiences with paranoid schizophrenia to illustrate this phenomenon: "I had to make sense, any sense, out of all these uncanny coincidences. I did it by radically changing my conception of reality" (Chadwick, 1993). They write,

> "In our terminology, these uncanny coincidences were false hypotheses engendered by prediction errors with inappropriately high precision… To explain them away Chadwick had to conclude that other people, including radio and television presenters, could see into his mind".
>
> (Frith and Friston, 2013, p. 12)

Like most theories of delusions, this proposal is highly schematic as it is current presented in the literature. Nevertheless, it boasts many well-advertised attractions. First, there is good reason to expect that a breakdown in the brain's capacity to reliably assign precision would be especially difficult to self-correct, which helps explain the strong intractability of delusional beliefs (Hohwy, 2013, p. 158–9). Second, it has long been known that dopamine disturbances play a major role in schizophrenia, and we saw above that dopaminergic activity is believed to be crucial to precision-weighting within predictive processing (Adams et al., 2013). Third, there is an impressive body of computational simulations of the effects of failures of precision-weighting in inference, and these seem to corroborate the idea that such failures would produce psychotic symptoms (Adams et al., 2013; Brown, Adams, Parees, Edwards, & Friston, 2013). Finally, there is a compelling body of neuro-imaging studies linking aberrant prediction error signalling to psychosis and delusion formation (Corlett et al., 2016; Teufel, Kingdon, Ingram, Wolpert, & Fletcher, 2010).

3.3. Hierarchical Bayesian models and the two-factor framework

Hierarchical Bayesian models of delusion are intended to compete with the two-factor framework introduced in S3.1 (Corlett et al., 2010; Corlett and Fletcher, 2015; Fletcher and Frith, 2009; Teufel and Fletcher, 2016).[7] I think that there are three important points of difference between the two frameworks, and it is important to spell out what they are.

First, and most obviously, whereas the two-factor framework posits *two* deficits necessary for the formation and retention of delusional beliefs, hierarchical Bayesian models of delusion posit just *one*—namely, the aberrant encoding of precision in a predictive coding hierarchy (although, as noted above, there is some disagreement on what exactly this dysfunction consists in).[8]

Second, advocates of hierarchical Bayesian models of delusion place a strong emphasis on the role of *bi-directional* message passing. Standard applications of the two-factor framework imply that message passing between perceptual and cognitive systems flows in *one* direction: perceptual information is passed onto systems underling belief fixation, which are charged with the responsibility of explaining this information. As such, perceptual systems are *informationally encapsulated* (Fodor, 1983): what the agent believes is responsive to what she perceives, but not vice versa. In contrast, advocates of hierarchical Bayesian models stress the influence of top-down predictions (i.e. priors) on the agent's perceptual experiences. Specifically, it is not just that abnormally high-precision prediction errors induce sharp revisions in the agent's "higher-level" beliefs about the world. In turn, these revisions then function as the *priors* with which the agent interprets (predicts) her own experience, leading to erroneous perceptual experiences that

[7] In a recent paper, Corlett et al. (2016, p. 1148) argue that hierarchical Bayesian models and the two-factor theory "are cast at different explanatory levels." Specifically, they argue that the two-factor theory "functions well at a descriptive level," but that perception and cognition are not distinct "at a deeper level," and that the two factors can be specified at two levels of the same hierarchy (Corlett et al., 2016, p. 1148). It is not clear how to understand this suggestion, however, as they still assert that "the same process that accounts for abnormal perception can also account for abnormal belief" (Corlett et al., 2016, p. 1148), they still advocate a single unified inferential hierarchy in place of a functional distinction between perception and cognition, and all three of the differences that I outline in this section still obtain.

[8] As an anonymous reviewer points out, it is important to stress that this single factor is nevertheless supposed to perturb all levels of the hierarchy, which yields a "multifactor psychopathology from a singular pathophysiology."

seem to confirm the very high-level beliefs that they are supposed to test—a process that Clark (2016, p. 81) describes as an "insulated self-confirming cycle" (see Denève and Jardri, 2016).

The third and most important difference, however, is not the appeal to a single deficit or the emphasis on bi-directional message passing as such, but the fact that hierarchical Bayesian models of delusion disavow the very idea that perceptual experiences and beliefs—and thus hallucinations and delusions—should be treated "as distinct entities" from the perspective of psychiatric explanation in the first place (Corlett and Fletcher, 2015, p. 97; see also Fletcher and Frith, 2009; Frith and Friston, 2013). Corlett and Fletcher (2015, p. 96), for example, write that positing two factors.

> "is only necessary in so far as there is a clear distinction between perception and inference: a distinction which is *not actually compatible with what is known about how the brain deals with the world*".
>
> (my emphasis)

They continue:

> "In a system that is arranged hierarchically, we may perhaps *choose* to refer to the inferences at the lower levels as perceptions and the inferences at the higher levels…as beliefs, but we suggest that it is important to consider that similar processing obtains at all levels of the hierarchy… [A]lthough it is possible—and sensible—at one level of analysis to distinguish beliefs from perceptions… at another level of analysis—the one that we think is *more useful—no distinction is called for*".
>
> (Corlett and Fletcher, 2015, pp. 96–7, my emphasis)

Importantly, this third difference is more radical than the first two. Specifically, one could accept that a single deficit underlies the formation and retention of delusional beliefs—for example, the dopamine disturbances at the core of predictive processing's account—whilst maintaining that this dysfunction nevertheless affects two functionally distinct computational systems. Likewise, one could also accept that an agent's beliefs can affect her perceptual experiences without disavowing the distinction between perception and belief. After all, there is bi-directional message passing between human beings, and this does not encourage us to deny that human beings are distinct systems.

The third difference goes much further than an emphasis on a single deficit and the possibility of cognitive penetration. Within the two-factor framework, it is assumed that the primary business of perceptual systems is to identify the best explanation of the agent's sensory inputs, which is then passed onto the agent's system of belief fixation. This latter system then combines this perceptual information with prior information and reasoning capacities in working out what to believe and how to act (Fodor, 1983). In hierarchical Bayesian models, by contrast, there is just *one* system at play—a *single* unified inferential hierarchy, with our intuitions about the differences between perception and belief merely tracking ill-defined regions of this hierarchy (Clark, 2013; Hohwy, 2013).

This is a radical break from traditional theorizing about delusions, not to mention much work in cognitive psychology (Firestone and Scholl, 2015). For this break to be *legitimate*, of course, the proposed unified inferential hierarchy must be able to accommodate the defining properties of delusional thought. Although advocates of hierarchical Bayesian models are confident that it *can*, in the next section I argue that this confidence is misplaced. First, however, it will be helpful to briefly summarise the lessons of this section, and outline the aims of Sections 4 and 5.

3.4. Summary and moving forward

Hierarchical Bayesian models of delusion are advanced as a substantial alternative to the influential two-factor framework. Specifically, such models disavow the distinction between perception and cognition in favour of a unified inferential hierarchy in which bi-directional message passing approximates Bayesian inference. Delusional beliefs are then claimed to arise from dysfunctions in this information-processing architecture. This approach is grounded in recent advances in computational neuroscience, and there is a compelling body of theoretical and evidential considerations that weigh in its favour. The upshot is an extremely impressive attempt to illuminate the mysterious process by which individuals lose their grip on reality—an attempt that warrants its status as one of the most promising and exciting products of the emerging field of computational psychiatry (Teufel and Fletcher, 2016).

Nevertheless, even the proponents of such models acknowledge that hierarchical Bayesian models of delusion confront challenges (Corlett et al., 2016). For example, how can a *single* dysfunction account for the asymmetry between individuals with Capgras delusion and non-delusional individuals who nevertheless share the same experience (Bortolotti and Miyazono, 2015, p. 642)? This challenge (see e.g. Coltheart, 2007) applies to monothematic delusions across the board, and so far seems to remain unanswered by advocates of hierarchical Bayesian models of delusion, suggesting that the application of such models might at best be restricted to polythematic delusions of the kind that feature in conditions such as schizophrenia (although see Gadsby and Williams, forthcoming).[9] In addition, further questions remain. How can such models account for the typically *social* contents of delusional beliefs (Corlett et al. 2016)? Does the underlying dysfunction consist in abnormally precise sensory evidence, abnormally precise priors,

[9] Corlett and Fletcher (2015, p.96) argue that positing two factors is "only necessary insofar as there is a clear distinction between perception and inference: a distinction which is not actually compatible with what is known about how the brain deals with the world." This objection to the two-factor framework is confused, however. First, the argument for positing two factors is that it accounts for certain dissociations—cases in which individuals share the same anomalous experience, but do not form the delusions—and has nothing to do with whether perception is inferential (Coltheart 2007; Davies and Egan 2013). Second, the distinction advocated in the two-factor framework is between perception and cognition, not perception and inference. One can think that perception is inferential in a computational sense without abandoning this distinction. In fact, that is the mainstream view in classical cognitive psychology (Firestone and Scholl, 2015; Fodor 1983). See footnote 7 for the possibility advanced recently by Corlett et al. (2016) that hierarchical Bayesian models can embrace two factors.

some combination of the two, or a temporal sequence by which one dysfunction mutates into the other (Notredame et al. 2014, pp. 9–10)?

Although answering these questions is important, in the next two sections I argue that hierarchical Bayesian models of delusion confront two much deeper challenges that must be answered if they are to genuinely illuminate the formation and retention of delusions. These challenges concern the very information-processing architecture that such models draw upon. Specifically, they concern the two core theoretical components of hierarchical Bayesian models: their *hierarchical* component, and their *Bayesian* component. It is no good trying to explain delusions in terms of dysfunctions in a Bayesian information-processing hierarchy if we have reason to suppose that human thought is neither hierarchical nor Bayesian—and, in Sections 4 and 5, I argue that we *do* have good reason to suppose that human thought is neither of these things.

Importantly, this means that the challenges that I raise in what follows concern *any* hierarchical Bayesian model of delusions, not just the specific theoretical proposal within that framework that I have outlined in this section. Specifically, these challenges will concern any approach to delusions with the following two characteristics:

(Hierarchical Bayesian Models of Delusion)

1. Hierarchy: Delusions arise at the higher or highest regions within a unified inferential hierarchy.
2. Bayesian: Information processing in this hierarchy implements approximate Bayesian inference.

4. The Inferential Hierarchy

Hierarchical Bayesian models of delusion disavow the distinction between perception and cognition in favour of a unified inferential hierarchy and contend that information processing in this hierarchy takes the form of approximate Bayesian inference. In this section and the next I argue that both components are deeply problematic.

In this section I focus on the first component—the claim that delusions are better explained by abandoning the distinction between perception and cognition in favour of a unified inferential hierarchy. I argue that extant ways of characterising this inferential hierarchy in the literature are inconsistent with the range of phenomena that delusions can represent.

4.1. Understanding the inferential hierarchy

The appeal to a unified inferential hierarchy in place of any substantive theoretical distinction between perception and cognition is one of the most distinctive and subversive features of hierarchical Bayesian models of delusion (Corlett and Fletcher, 2015; Fletcher and Frith, 2009; Teufel and Fletcher, 2016). For this radical departure from traditional theorizing to be legitimate, the concept of the inferential hierarchy must be able to accommodate the kinds of phenomena exhibited in delusional thought. To evaluate whether this condition is met, then, we must understand exactly how the inferential hierarchy is understood within hierarchical Bayesian models, and where delusions are supposed to feature within it.

First, a ubiquitous claim made within the literature is that what we intuitively think of as beliefs correspond to representations at "higher" levels of the inferential hierarchy (see Corlett et al., 2016; Frith and Friston, 2013):

> "There may well be a hierarchy of such inferencing devices in the brain, where lower levels of the hierarchy are more relevant to perception and upper levels are more relevant to beliefs".
>
> (Fletcher and Frith, 2009, p. 56)

As such, *delusional* beliefs are supposed to arise at the higher levels of the inferential hierarchy. In fact, Frith and Friston (2013, p. 10) go further and argue that it is the representations "at the *top* of this hierarchy that are particularly resistant to change" in schizophrenia (my emphasis).

To evaluate this claim, we must understand how the proposed inferential hierarchy is organised. A system is hierarchically structured when its parts can be arranged into levels that exist in a functional sense either below, above, or at the same level as other parts. For any given hierarchy, one must therefore be able to specify the principle that determines the nature of interlevel relations ("above," "below," "at the same level") and positions (e.g. "top," "bottom") within it. For example, command hierarchies are organised such that one level X exists above another level Y if and only if agents at level X are capable of commanding agents at level Y. A typical hierarchy of this kind will contain multiple levels, such that the influence of those at the top on the behaviour of those at the bottom is mediated by intermediate levels of subordinates.

In the case of the neocortex, it would clearly be mistaken to think of the proposed inferential hierarchy as anything resembling a stepladder (Clark, 2016, p. 145). Brains receive multiple streams of sensory input processed in parallel. The hierarchy is thus probably better conceptualised as a *sphere* with outer edges corresponding to different sensory modalities (Penny, 2012). As you reach the middle of this sphere (the "higher" levels of the hierarchy), representations become multimodal and amodal and thus capable of predicting activity across different sensory modalities. This is no stranger than military command hierarchies in which a general at the top is ultimately responsible for commanding the behaviour of individuals across distinct branches of the military (army, navy, air force, etc.), each of which has its own subordinate hierarchy of command.

What, then, is the principle by which the proposed inferential hierarchy in the neocortex is organised? The most influential idea in the literature is that levels in the hierarchy are ordered according to *spatiotemporal scale* or just *temporal* scale:

> "…a hierarchy of increasingly abstract generative models. Basically, these will be models capturing regularities across larger and

larger temporal and spatial scales".

(Clark, 2012, p. 762)

"…higher levels of the hierarchy represent larger amounts of space and longer durations of time".

(George and Hawkins, 2009, p. 2)

"…the brain is organised hierarchically, according to the temporal structure and regularities in which events occur… [P]rogressively more extended time scales will map onto progressively higher regions of the cortical hierarchy"

(Harrison, Bestmann, Rosa, Penny, & Green, 2011, p. 1)

"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".

(Hohwy, 2012, p. 2)

"These hierarchies are ordered according to the temporal scales of representations, where the slowest time-scale is at the top… The basic idea is that temporal hierarchies in the environment are transcribed into anatomical hierarchies in the brain; high-level cortical areas encode slowly changing contextual states of the world, while low-level areas encode fast trajectories"

(Kiebel, Daunizeau, & Friston, 2008, p. 2)

On this view, one level X is above another level Y in the inferential hierarchy if and only if it represents phenomena at a larger spatiotemporal scale. This view thus captures the popular idea that going higher up cortical hierarchies corresponds to representations that are increasingly invariant in the face of changes in proximal sensory inputs (Hawkins & Blakeslee, 2005; Hohwy, 2013). For example, low levels of the inferential hierarchy predict fast-moving, highly "variant" environmental phenomena—basic colour and contrast information, edge-segments, and so on—whereas levels higher up predict phenomena that remain more invariant in the face of changes in proximal stimulation and perceptual conditions—for example, the categorisation of the objects in a visual scene that remains constant as you move around a room. Intuitively, this notion of variance captures temporal scale: highly variant phenomena change quickly—as a function of changing sensory input, at least—whereas invariant phenomena do not.

In principle, there are two ways in which this appeal to spatiotemporal scale could be understood that are not carefully distinguished in the literature.

On one view, the hierarchy is organised according to the *content* of different levels—according to *what* is represented. This interpretation is suggested, for example, by claims that "high-level cortical areas *encode* slowly changing contextual states of the world" (Kiebel et al., 2008, p. 2), that "higher levels of the hierarchy *represent* larger amounts of space and longer durations of time" (George and Hawkins, 2009, p. 2), and that a "hierarchy of cortical areas allows *information pertaining to regularities* at different spatial and temporal scales to settle into a mutually consistent whole" (Clark, 2013, p. 3) (all emphases my own).

On another reading, the hierarchy is organised according to properties of the representational vehicles themselves—specifically, according to the temporal scale of the *representations*.[10] This reading is suggested, for example, by the claim that "many aspects of brain function can be understood in terms of a hierarchy of temporal scales at which *representations* of the environment evolve" (Kiebel et al., 2008, p. 1), that "the key aspect of this generative model is that state transitions proceed at different rates at different levels of the hierarchy" (Friston et al., 2017b, p. 390), and that "*inferences* at the higher levels… [are] more abstract and immutable" (Corlett and Fletcher, 2015, p. 96) (all emphases my own).

This latter reading is difficult to make sense of, however (see Gadsby and Williams, forthcoming for a more thorough treatment). First, the temporal scale cannot refer to the relevant representational *systems*: the representational systems in primary sensory areas are just as enduring as any other representational systems in the brain. As such, it must refer to the time-length of what are sometimes called "active" or "occurrent" representations—namely, the representations actively implicated in the brain's ongoing explanation of its evolving sensory inputs. This is untenable, however. Although in cases of "online" perception higher-level representations in sensory cortices do remain more invariant in the face of changes in proximal stimulation, this reflects *what* they represent—that is, their *content*—and is not itself an autonomous principle for ordering hierarchies. To see this, consider that the time-length of active representations within different cortical areas is highly contingent on task and context. (For a simple illustration, focus your vision on an unchanging stimulus and let your mind wander).

For this reason, I will assume that the cortex's unified inferential hierarchy is organised in terms of the spatiotemporal scale of *what* is represented—in terms of the *contents* represented at different levels. On this view, what we intuitively think of as beliefs—whether delusional or otherwise—correspond to regions of an inferential hierarchy involved in representing phenomena at larger spatiotemporal scales.

4.2. Evaluating the inferential hierarchy

To state the proposal this clearly is to see that it cannot work. There are no restrictions on the range of phenomena we can think and form beliefs about. We can engage in highly abstract reasoning about small things, big things, slow things, and fast things (see Vance, 2015; Williams, forthcoming a). Further, we can form beliefs about phenomena that have *no* spatiotemporal scale, such as the number *42* or *Bayes' theorem*. Indeed, cognitive psychologists and philosophers have sometimes appealed to this characteristic of

[10] I thank Stephen Gadsby for making this interpretation of the hierarchy clear to me.

thought to *distinguish* the architecture of belief fixation from perceptual systems (Fodor, 1983). For example, it is plausible to think that the range of phenomena my *visual system* can represent is not unrestricted in this way: it can recover the three-dimensional spatial structure, shapes, and colours of the visual scene—namely, the information revealed in the specific kind of energetic stimulation to which one's retinal cells are responsive—and nothing more (Pylyshyn, 2003).

This "representational reach" (Williams, forthcoming b) of conceptual thought creates the following problem, however.[11] Where do our thoughts about tiny fast-moving entities fit within the inferential hierarchy? If the brain's inferential hierarchy is ordered such that increasingly higher levels correspond to representations of phenomena at increasingly greater spatiotemporal scales, they cannot exist high up the hierarchy. On the other hand, there is no plausible sense in which such highly abstract thoughts exist at lower levels either, if low levels of the hierarchy are functionally close to proximal sensory inputs. As such, there seems to be no place for them. Given that our capacity to engage in highly abstract thought about phenomena at *any* spatiotemporal scale or abstraction is a defining characteristic of human cognition, the inferential hierarchy thus seems to be incapable of accounting for this aspect of cognition (Williams, forthcoming a).

There are two points to add to this.

First, this lesson applies as much to delusional thought as to nonpathological thought. In the case of polythematic delusions exhibited in conditions such as schizophrenia, for example, the representational reach of delusional thought does not seem to be any more restricted than the reach of nonpathological thought. Just think of the three examples of John Nash's delusional beliefs with which I began this paper. Among other things, they concern the *left foot of God*, the *Emperor of Antarctica*, and a *conspiracy*. As with delusions more generally, it is unclear that such beliefs concern phenomena at *any* spatiotemporal scale. Where is the *left foot of God* predicted in one's inferential hierarchy? Would it be above, below, or at the same level as an inference concerning one's own death (the content of Cotard delusion)? If interlevel hierarchical relations are individuated in terms of spatiotemporal scale or levels of invariance, it is extremely difficult to even make sense of Frith and Friston's (2013) claim that delusions in schizophrenia exist "at the *top*" of the hierarchy. Do such delusions always represent the *biggest, slowest*, or most *invariant* phenomena?

To make this concrete, consider *delusional parasitosis*, in which individuals mistakenly believe themselves to be infested with *tiny* parasites (Prakash et al., 2012). Where in the inferential hierarchy does this belief exist? According to hierarchical Bayesian models, it should exist somewhere near the top—if not *the* top. Given that it concerns extremely small phenomena, however, standard characterisations of the inferential hierarchy imply that it should exist somewhere near the bottom. Again, there seems to be no place for it. One might respond that this belief is responsive to the anomalous experiences of individuals with this delusion, which *do* exist lower down in the inferential hierarchy. This fact just serves to illustrate an even deeper problem, however: delusional beliefs are often generated as an attempt to explain the contents of the agent's perceptual representation of the world (Coltheart, 2007). As such, we should expect many delusional beliefs to concern phenomena at the *same level of spatiotemporal grain* as the content of the individual's perceptual experience. If this is right, delusions evidently cannot always represent phenomena that are more invariant, bigger, or slower than the phenomena represented "lower" down in the inferential hierarchy. They often represent *the same thing*.

The second important point is this: this problem does not afflict the two-factor framework. Specifically, if one thinks that belief fixation is subserved by a distinct system, there is no reason why this system could not make use of an information-processing architecture capable of accounting for the representational reach of conceptual thought (Fodor, 1983). The problem comes from abandoning any such system in favour of a unified inferential hierarchy in which levels and interlevel relations are characterised in terms of the spatiotemporal scale of the phenomena that they represent.

4.3. Abstraction to the Rescue?

Given this, an obvious response is to reject the idea that the hierarchy should be understood exclusively in terms of representations of phenomena at increasing levels of spatiotemporal scale. For example, one might argue that increasing spatiotemporal scale only characterises the inferential hierarchy in sensory cortices, and that a different kind of hierarchy takes over in association cortex. This need not be post hoc. What predictive processing is *strictly* committed to is that higher levels of the inferential hierarchy identify the environmental variables that best *predict* the statistical patterns represented at the level below. Perhaps at first this means identifying phenomena at greater spatiotemporal scale, but that this process of identifying deeper predictive variables in the environment eventually requires tracking phenomena along a different dimension. For example, many advocates of hierarchical Bayesian models of delusion suggest that the inferential hierarchy is ordered by increasing levels of "abstraction" (see Corlett et al., 2016; Frith and Friston, 2013; Griffin and Fletcher, 2017, p. 267).[12]

As an example, Frith and Friston (2013) point to language processing in reading (see also Fletcher and Frith, 2009, p. 55). In this case, low levels estimate graphic shape components of which letters are composed. As you move up the hierarchy, you then traverse "representations of words and sentences, reaching meaning at the highest level" (Frith and Friston, 2013, p. 11). In this context, the concept of the hierarchy refers to something like relations of abstract composition (shapes compose letters compose words compose phrases etc.). Clark (2016, p. 24) seems to have the same in mind when he refers to the need for intelligent systems "to represent and process 'complex, articulated structures' (Hinton, 1990, 47) such as part-whole hierarchies: structures in which elements form wholes

[11] An anonymous reviewer asks me to define what I mean by "thought." I mean the suite of psychological capacities often associated with what cognitive neuroscientists call non-perceptual or "higher" cognition: reasoning, reflecting, deliberating, planning, etc.

[12] Another possibility raised by an anonymous reviewer is that the representational capacities I identify in this section are dependent on *language*. I explicitly consider this possibility in Williams (forthcoming a).

that can themselves be elements of one or more larger wholes."

I do not think that this suggestion can solve the underlying problem, however. Although the appeal to processes of abstraction is plausible for our broadly *perceptual grip* on the ambient environment—roughly, our extraction of nested compositional structure from the sensory signal—it is much less plausible for the domain of higher cognition. Recall the claim advanced by hierarchical Bayesian models: that perception-like states correspond to "lower" levels of the hierarchy, and belief-like states correspond to "higher levels" of the hierarchy. If the hierarchy is ordered in terms of abstract compositional relations, one confronts the immediate objection that we can *think* about *any* of the elements in such structures, not just the higher levels. The most straightforward way to appreciate this is to realise that we can engage in highly abstract reasoning about *any* of the properties represented "low down" in our perceptual systems: oriented lines, light patches, binocular disparity, and so on. Moreover, we saw above that this kind of phenomenon is likely to be *central* to cases of delusional thought, precisely because such beliefs often arise as an attempt to explain the contents of the agent's perceptual experience—and thus will likely concern phenomena at *the same spatiotemporal grain* as those experiences.

One might respond that there is a different notion of abstraction at play that can handle this problem. That is, one might argue that although there is *one* interpretation in which an individual's belief that she can hear her thoughts spoken aloud represents the same phenomenon as her hallucination of those thoughts, her beliefs are nevertheless still more *abstract*. For example, perhaps the notion of abstraction at play need not be characterised in terms of *content* as I have so far assumed—that is, in terms of the worldly items represented in the inferential hierarchy—but rather tracks the *way* in which worldly phenomena are represented in the hierarchy.

This suggestion is difficult to square with the information-processing architecture posited by predictive processing, however. If the inferential hierarchy is not ordered according to content, it must be ordered by some property of the relevant representational vehicles or forms of computation at different levels of the hierarchy. I argued above that popular appeals to the time-length of the activated representations at different levels is untenable. It is not obvious what other possibilities for ordering the hierarchy there are, however. It is a fundamental tenet of predictive processing that information processing at all levels of the hierarchy takes the same form (Corlett and Fletcher, 2015, p. 96; see also Clark, 2013; Corlett and Fletcher, 2015; Friston, 2005; Hohwy, 2013).

Further, it is crucial that the relevant notion of "abstraction" not simply mask a tacit appeal to the fact that one representation is a *belief*, and one kind of representation is a *perceptual experience*. After all, advocates of traditional theoretical approaches to delusion will concede that there is a certain sense in which beliefs—including delusional beliefs—are more "abstract" than perceptual experiences. If hierarchical Bayesian models of delusion are to genuinely differentiate themselves from standard theoretical approaches, then, they must characterise the nature of the inferential hierarchy in a way that does not tacitly appeal to a distinct realm of belief fixation. That is, they must characterise the nature of the inferential hierarchy in such a way that it is still a *hierarchy*, rather than a set of hierarchically structured perceptual systems that interact with a domain of conceptual representation that can in principle range over *anything*. Without an explicit specification of the principle by which the alleged unified inferential hierarchy is ordered, I think that there is a significant risk that this constraint is violated—that nebulous gestures to the "higher" or "top" levels of the inferential hierarchy simply mask tacit appeals to a distinct system of belief fixation. If this is right, the concept of the hierarchy *as such* will not in fact be doing any genuine explanatory work. Given that the inferential hierarchy is supposed to be what most sharply distinguishes hierarchical Bayesian models from their predecessors—that "in our model, hierarchy is key" (Corlett et al., 2016, p. 1148)—this would be a damning indictment.

I do not claim that that this challenge *cannot* be answered. Specifically, it would be answered by offering an account of "abstraction" that avoids both horns of the foregoing dilemma—that is consistent with the representational reach of thought, and yet that preserves a genuine inferential hierarchy. I do not know of any such account, however. As such, until the challenge *is* answered, we should be sceptical of a model of delusion that rests so heavily on a theoretical posit that it leaves so underspecified.

4.4. Responses

I think that there are two likely responses to the foregoing argument.

First, one might point to the extensive evidence of hierarchical representation in the neocortex (Felleman and Van Essen, 1991). We have known since Hubel and Wiesel's (1962) pioneering work that hierarchical representation is central to information encoding in cortical networks, and this insight has only been vindicated and extended in recent decades. In addition, the importance of hierarchical feature extraction to information processing has recently been bolstered by the impressive engineering successes of multilayer ("deep learning") neural networks in artificial intelligence research (Goodfellow, Bengio, & Courville, 2017). Given this, one might argue that the existence of hierarchical representation is so well established that one can draw upon it for the purposes of scientific theorizing.

This objection is irrelevant, however. Such hierarchical Bayesian models do not claim just that there *is* hierarchical representation in the neocortex. That could not reasonably be denied, and I do not deny it. They claim that there is *only* hierarchical representation in the neocortex—that all cortical information processing is subsumed within a single unified inferential hierarchy. Of course, they are not alone in this view (e.g. Hawkins & Blakeslee, 2005). Nevertheless, this claim is dramatically underdetermined by the state of current neuroscience (Firestone and Scholl, 2015; Marcus, 2015). We do not have viable neuroscientific theories of high-level human cognition, and the foregoing considerations suggest that adequate models of phenomena like conceptual thought will need to move beyond hierarchical inference. In their overview of hierarchical predictive coding, Bastos et al. (2012, p. 697) assert that "current dogma holds that the cortex is hierarchically organised." I suggest that this dogma might be fruitfully abandoned when it comes to explaining the complex representational capacities exhibited in human thought (Williams, forthcoming a). As Marcus (2015, p. 205) writes,

> "Progress [in neuroscience] has been stymied in part… by too slavish a devotion to the one really good idea that neuroscience has had thus far, which is the idea that cascades of simple low level "feature detectors" … percolate upward in a hierarchy."

Second, one might argue that the distinctive feature of hierarchical Bayesian models of delusion is not their emphasis on an inferential hierarchy *as such*, but rather their commitment to a *single* deficit underlying the generation and maintenance of delusional beliefs and their commitment to the role of bi-directional message passing (see Section 3.3 above). Strictly, the foregoing argument does not directly address these components of hierarchical Bayesian models. As such, one might argue that the challenge I have raised in this section is irrelevant to the most distinctive and distinguishing feature of these models.

There is an element of truth in this response. Specifically, my criticism of the idea that all information processing in the brain is subsumed within a single unified inferential hierarchy does not itself directly undermine either the idea that a single disturbance underlies delusions—for example, an over-confidence in sensory evidence relative to prior beliefs (perhaps induced by dopaminergic malfunctioning)—or the emphasis on cognitive penetration at the core of hierarchical Bayesian models.

Nevertheless, these features of hierarchical Bayesian models exist within an information-processing architecture that makes a core appeal to the concept of an inferential hierarchy, and it is not obvious to what extent the other contributions of such models can be salvaged if one abandons that hierarchy. For example, the appeal to a deficit in precision estimation as the single dysfunction implicated in delusional beliefs is supposed to consist in over-weighting the precision of lower (sensory) levels of the hierarchy *relative* to higher levels, and the emphasis on bi-directional message passing is engendered by an information-processing architecture in which all levels are ultimately connected to each other (albeit via intermediary layers) in a unified hierarchy. Perhaps these features of the theoretical approach to delusions could be preserved *without* a commitment to a unified inferential hierarchy in place of a distinction between perception and belief. If so, my criticism in this section would only target *one* aspect of hierarchical Bayesian models of delusion. Given how central this aspect is to the presentation of such models, however, this would itself be a substantive discovery.

5. Is belief fixation Bayesian?

As noted above, a foundational methodological principle of psychiatry is to explain psychopathologies in terms of dysfunctions in the mechanisms underlying healthy psychological functioning. For example, Corlett et al. (2010, p. 346) begin their article defending hierarchical Bayesian models of delusion by advocating

> "a cognitive neuropsychiatric approach to delusions. That is, the starting point is to review what we understand about the healthy functioning of particular processes… before extrapolating to the disease case."

Hierarchical Bayesian models of delusion seek to explain delusions in terms of dysfunctions in a process of hierarchical Bayesian inference. As such, a central assumption of such models is that belief fixation in the *healthy* (i.e. neurotypical) population is Bayesian. On the widespread assumption that Bayesian inference is statistically optimal, the claim is therefore that the mechanisms underlying belief fixation in the healthy population are (approximately—see below) inferentially optimal. Importantly, this assumption is not just *implicit* in the approach:

> "In the normally functioning brain information from different sources is combined in a *statistically optimal manner*. The mechanism for achieving this is well captured in a Bayesian framework".
>
> (Frith and Friston, 2013, p. 5)

> "PP [predictive processing] aims to provide an account of how the brain *optimally infers* the causes of its noisy, unreliable, and ambiguous inputs".
>
> (Griffin and Fletcher, 2017, p. 282, my emphasis)

In this section I criticise this assumption. Before I advance this critique, however, I note three important clarifications.

First, I noted in Section 2 that nobody believes that the brain performs *exact* Bayesian inference. Instead, the claim is rather that the algorithms underlying belief fixation implement *approximate* Bayesian inference. This provides some safety in the face of putative counter-evidence. For example, in virtue of being approximation algorithms, the results of such algorithms will in certain contexts systematically deviate from the outcomes of exact Bayesian inference (Sanborn and Chater, 2016). Nevertheless, it evidently cannot provide unlimited safety. The whole point of approximation algorithms is that they *approximate* Bayesian inference. Deviations from exact Bayesian inference must therefore be *principled*.[13]

Second, there is obviously both enormous variability and noise in the healthy population. The fact that one's cousin Barry once violated Bayes optimality evidently does not undermine Bayesian models of cognition. More generally, if it were the case that deviations from Bayesian inference were distributed such that *on average* the mechanisms underlying belief fixation produced Bayes optimal inferences, it is not clear that this would be a problem for such Bayesian models. Such models would simply be *idealised*, but this would not distinguish them from any other model used in science.

Finally, advocates of hierarchical Bayesian models of delusion recognise that their optimality assumptions create at least a prima facie problem for accounting for *delusional* beliefs. For example, Fineberg and Corlett (2016, p. 75) write:

[13] An anonymous reviewer adds: "an important example of this is the free-energy principle, whose objective function is variational free energy that is used for approximate Bayesian inference."

> "This [hierarchical Bayesian] model of mind/brain function, *which is committed to veracity*, may seem at odds with the generation of psychotic symptoms like hallucinations and delusions. How can the complex and strongly held misbeliefs that characterise psychotic illness arise from a truth-seeking system?"

My worry here is deeper and more fundamental, however: how could the kinds of beliefs pervasive in the *healthy* population arise from statistically optimal "truth-seeking" inferential mechanisms? Among this set of nonpathological beliefs are extremely widespread superstitions about celestial father figures, ghosts, UFOs, conspiracy theories, and so on, which is not to mention the more mundane false beliefs that people often have—for example, the inclination of most people to think that they are better than average at most things (Hoorens, 1993). It is important to be clear about what advocates of hierarchical Bayesian models of delusion are claiming: that such convictions arise from approximately optimal inferential mechanisms inside people's heads.

What reason is there to accept this claim? This question can be decomposed into two parts. First, is there compelling *evidence* for a Bayesian view of belief fixation? Second, is there good theoretical reason to *expect* the mechanisms underlying belief fixation to be Bayesian? Of course, I cannot settle such questions here. Nevertheless, I think that one can make a strong case that the answer to both questions is *no*.

5.1. Evidence for Bayesian beliefs

First, then, advocates of hierarchical Bayesian models claim frequently and with confidence that the mechanisms underlying belief fixation in the healthy population are Bayesian. For example, we saw above Frith and Friston's (2013, p. 5) remark that "in the normally functioning brain information from different sources is combined in a *statistically optimal manner*." What evidence do they advance in defence of this surprising claim? Frith and Friston offer only a single reference: Ernst and Bank's (2002) influential demonstration that the integration of visual and haptic information can fruitfully be modelled in terms of Bayesian inference. Similarly, when Knill and Pouget (2004, p. 712) argue that "human observers behave as optimal Bayesian observers," they again draw heavily on Ernst and Bank's work, arguing that '[p]erhaps the most persuasive evidence for the Bayesian coding hypothesis comes from sensory cue integration." Likewise, a piece of evidence often advanced in favour of the Bayesian brain hypothesis is Weiss's, Simoncelli, and Adelson (2002) demonstration that many motion illusions and psychophysical results can be understood in terms of Bayes optimal motion perception (e.g. Clark, 2016, p. 40).

The problem with such examples, however, is that they do *not* concern the mechanisms underlying *belief fixation*. Instead, they concern relatively low-level sensorimotor processes. In fact, it is widely *conceded* among advocates of Bayesian models that the best evidence in their favour comes from low-level, largely unconscious mechanisms implicated in perception and motor control (Chater et al., 2010). Summarising such evidence, Clark (2016, p. 40) writes that "at least in the realm of low-level, basic, and adaptively crucial, computations, biological processing may quite closely approximate Bayes' optimality."

In the current context, however, such evidence is irrelevant. Delusions are beliefs. Even among advocates of hierarchical Bayesian models, we have seen that they are thought to arise and exist at the higher or even *highest* levels of the inferential "hierarchy." As such, evidence of Bayes optimality in the domain of sensorimotor processing does not directly bear on whether the mechanisms underlying *belief fixation* are Bayesian. It *would* bear on the issue if one thought that one could simply extrapolate from the case of sensorimotor processing to intuitively "higher" cognition. Any such extrapolation would beg the question in the current context, however.

Is there *any* evidence that higher-level cognitive processes—those implicated in the production of human person-level beliefs of the sort that feature in delusions—are approximately Bayes optimal? Any exhaustive summary or evaluation of the extant research attempting to answer this question in the affirmative is impossible here (for arguments *in defence* of Bayesian accounts of higher-level cognition, see: Chater et al., 2010; Griffiths et al., 2010; Oaksford and Chater, 2007, 2009; Sanborn and Chater, 2016; Tenenbaum et al., 2011). Nevertheless, the problems with such research are widely known (see Bowers and Davis, 2012; Glymour, 2011; Jones and Love, 2011; Marcus, 2009b; Marcus and Davis, 2013).

First, the methodology *underlying* much of Bayesian cognitive science is "rational analysis" (Anderson, 1990), which specifies the putative goal of a cognitive system and then derives the "optimal behaviour function" for solving such goals in the relevant environment, making "minimal assumptions about computational limitations" (Oaksford and Chater, 2009, p. 71; see Tenenbaum et al., 2011). Not only does this heavily bias such research in favour of Bayesian analysis, but these optimality assumptions are evolutionarily and biologically dubious (Marcus, 2009a, 2009b).

Second, because much of this work is so unconstrained by mechanistic considerations, there is a significant risk of "Bayesian just-so stories" (Bowers and Davis, 2012), whereby fiddling with Bayesian model parameters (priors, likelihoods, cost functions, etc.) allows researchers to fit Bayesian models to almost *any* behaviour post hoc (see also Marcus and Davis, 2013). As Glymour (2011, p. 200) quips, "I know of no Bayesian psychological prediction more precise than "We can model it.""

Third, much of this research has been accused of cherry-picking confirmatory evidence and ignoring the pervasive and well-documented evidence of suboptimal cognition and behaviour (Marcus and Davis, 2013). As Marcus and Davis (2013, p. 3) write,

> "[T]he probabilistic-cognition literature as a whole may disproportionately report successes….[I]n many domains, results that fit naturally with probabilistic techniques and claims of optimality are closely paralleled by equally compelling results that do not fit so squarely."

Finally, although much of the research within Bayesian cognitive science does not concern sensorimotor processing *as such*, it is not obvious that its focus is relevant to the formation and retention of delusional beliefs. For example, most of Tenenbaum et al's

(2011) focus is on fast, automatic information-processing mechanisms—parsing the syntactic structure of a sentence, running internal "physics simulations," categorisation judgements, and so on—that are not directly relevant to the fixation of person-level beliefs. Further, Sanborn and Chater's (2016) impressive research modelling apparent reasoning errors by appeal to the implications of *approximate* Bayesian inference focuses on a radically different kind of approximation algorithm to that which features in predictive processing (namely, sampling rather than variational methods) (see Penny, 2012), and their focus is on explicitly formulated logical and probabilistic reasoning tasks in which delusional individuals do not seem to be impaired anyway (Maher, 2001).

None of this is conclusive. Nevertheless, I think that a fair summary of the evidence in this context would be as follows: where there is strong evidence of Bayes optimal processing, it exists largely in the realm of sensorimotor processing (Clark, 2016, p. 40). As one moves up to more intuitively higher-level cognitive processing, the evidence becomes substantially weaker: Bayesian models are highly abstract, unconstrained by mechanistic details, and often methodologically dubious. Further, it is not clear that they are relevant to the mechanisms underlying belief fixation thought to malfunction in the case of delusional beliefs anyway.

5.2. Evidence against Bayesian beliefs

Of course, it is not sufficient to focus only on evidence *for* a Bayesian view of belief fixation. Is there any evidence *against* such a view? On this question, a cursory examination of previous decades in behavioural economics and social psychology reveals an apparent *abundance* of systematic deviations from Bayes optimality in human thought (e.g. Marcus, 2009a, 2009b; Kahneman, 2011). However, much of this literature is highly controversial, both methodologically and substantively, and we have already seen some reason to think that alleged reasoning errors can sometimes be *described* as an effect of approximate Bayesian inference (especially after the fact). Nevertheless, I think that one can identify three phenomena that are well documented and deeply problematic for a Bayesian view of belief fixation—either because they have nothing to do with Bayesian inference (see footnote 14), or because they flatly contradict norms of Bayesian inference.

First, all human beings seem to be afflicted by *confirmation bias*, the fact that evidence is sought or interpreted "in ways that are partial to existing beliefs, expectations, or a hypothesis in hand" (Nickerson, 1998, p. 175; Stanovich and West, 2007). In Bayesian terms, rather than updating prior beliefs upon receipt of new evidence, people instead systematically seek out new evidence—or are at least more receptive to evidence—that confirms their prior beliefs (Mercier and Sperber, 2017, p. 218).[14] Second, people seem to universally succumb to *motivated reasoning*, the tendency "to arrive at conclusions that they *want* to arrive at" when "accessing, constructing, and evaluating beliefs" (Kunda, 1990, p. 480, my emphasis).[15] That is, whereas confirmation bias filters for evidence that confirms existing beliefs, motivated reasoning involves much greater scrutiny towards beliefs that we don't like than the ones we do. For example, in a study that requires subjects to evaluate an article outlining the risks of caffeine for women, female caffeine drinkers were more likely to doubt the conclusion than women who were not caffeine drinkers, and men who had nothing at stake exhibited no such effect (Kunda, 1990; Marcus, 2009a, p. 57).

Perhaps most damningly, researchers have also found evidence of the so-called "backfire effect," and the effect of disconfirming evidence on the polarization of passionately held attitudes and beliefs more generally (Nyhan and Reifler, 2010; see Baron, 2008). The backfire effect consists of actively *increasing* one's confidence in a belief (or set of beliefs) when presented with contradictory evidence—the exact opposite of Bayesian inference. For example, Nyhan and Reifler (2010) famously found that presenting individuals with evidence that contradicted their political ideology or convictions resulted in *increased* confidence in such convictions.

These phenomena are implicated in the mechanisms underlying belief fixation in the *healthy population*, and they either have nothing to do with Bayesian inference or are flatly inconsistent with it. Perhaps most interestingly, advocates of hierarchical Bayesian models of delusion *themselves* seem to concede that human cognition is subject to such "biases." For example, in the very same article in which Fletcher and Frith (2009, p. 53) claim that "a hierarchical Bayesian system might be a basic principle for brain function," they also note the following:

> "Evidence suggests that *we are all* rather poor at letting our sensory experience update our beliefs, and that we are susceptible to prior beliefs and social constraints that greatly limit our ability to deal with evidence rationally… For the most part, *people do not depart from the beliefs of the herd*".
>
> (Fletcher and Frith, 2009, p. 52, my emphasis)

Likewise, to explain the specific contents of delusional beliefs, Corlett et al. (2016) advance the following suggestion:

> "Given that such inference relies on a person's best guess, then it follows that their own prior knowledge and expectation will

[14] One might argue that confirmation bias is irrelevant because Bayes' theorem applies only to the process of belief updating consequent upon receipt of new evidence, and not the process by which that evidence is chosen. But then a substantial component of what is distinctive about human belief fixation would have nothing to do with Bayesian inference, which is equally damning.

[15] An anonymous reviewer helpfully points out that the introduction of *goals* transforms the issue from simple inference to the kinds of phenomena modelled in Bayesian *decision theory*, which introduces a loss or reward function (and thus the idea of preferences). Under these contexts, "an optimal inference may not be consistent with prior preferences and, in relation to those preferences, could be considered suboptimal" (anonymous reviewer). There are three things to say here, however: first, such motivational factors do not currently feature in hierarchical Bayesian models of delusion (Fletcher and Frith, 2009; Corlett et al., 2010); second, the fact that one *can* model such phenomena as Bayesian in a post hoc manner ("just so") provides no evidence that they are Bayesian (see my response below to the issue of falsifiability); and third, it may be that some cognitive functions positively mandate updating procedures that are explicitly designed to produce false (not just suboptimal) beliefs and thus not conform to Bayesian updating (see the main text and Pinker, 2005; Kurzban, 2010), and it is not obvious that such processes of dynamic belief updating *can* be modelled in terms of *Bayesian* decision theory.

necessarily determine the content of the emergent belief. *And since their own expectations are, inter alia, socioculturally determined*, there will be a strong overlap between those of the person and the time and culture they inhabit".

(Corlett et al., 2016, p. 1149, my emphasis)

Superficially, at least, this assemblage of commitments is difficult to understand. It cannot *both* be the case that the mechanisms underlying belief fixation are Bayes optimal and that they are not Bayes optimal—for example, that they are socioculturally determined, or merely align with whatever "the herd" happens to believe. To make this combination of views consistent, one would have to believe that these non-Bayesian aspects of belief fixation are somehow irrelevant to the mechanisms underlying delusions (see below). Corlett et al. (2016), however, claim just the opposite: they explicitly appeal to these non-Bayesian elements to explain the social aspect of delusions.

To summarise, then, the evidence that the mechanisms underlying belief fixation are Bayesian is either weak or seemingly irrelevant, and the evidence that they are *not* is seemingly robust—so robust, in fact, that it is conceded by advocates of the Bayesian models themselves. Nevertheless, given the foregoing worries about the possibility of Bayesian just-so stories in this domain and the difficulties associated with adjudicating these claims, it is useful to turn to a deeper question: should we *expect* the mechanisms underlying belief fixation to be Bayesian?

5.3. Should we predict Bayesian Beliefs?

Again, despite the frequency with which advocates of hierarchical Bayesian models of delusion claim that we *should* expect the mechanisms underlying belief fixation to be Bayesian, there are very few explicit justifications of this expectation in the literature. The core idea seems to be that the *function* of the mechanisms underlying belief fixation in the healthy population is to enable agents to arrive at true beliefs about the world, or at least optimal inference under conditions of uncertainty. As such, given that Bayesian inference is the optimal method for performing this function, we should expect evolution to have designed mechanisms of belief fixation that operate according to Bayesian inference. After all, Bayesian brains would outcompete non-Bayesian brains.[16]

The most common objection in the literature to this kind of argument is to the inference from the *optimality* of Bayesian inference to an expectation that evolution would likely produce Bayesian brains. For example, several authors have argued that we should not *expect* evolutionary processes to construct optimal or even approximately optimal mechanisms. Instead, natural selection is a *satisficer*, not an optimizer, and evolutionary "design" (selective) processes often get stuck in local optima, not global maxima (Marcus, 2009a. 2009b). On this view, the brain is a *kluge*—a compromise of evolutionary history and idiosyncratic selective pressures, not a perfectly designed organ of adaptive success.

These objections are salutary. Nevertheless, I think that a much deeper problem concerns the *first* step of the argument. Advocates of predictive processing often write as if the main or only *function* of the brain is to infer the best explanation of its sensory inputs—to minimize (long-run, average) prediction error, or variational free energy (Adams et al., 2013; Corlett et al., 2016; Friston, 2010). Any such view is deeply suspect. Human beings are social primates whose cognitive, neurological, and bodily mechanisms are the upshot of selective pressures that constructed solutions to numerous different "design problems" in multiple idiosyncratic ancestral environments of selection (Dennett, 1995; Tooby and Cosmides, 2015). This suggests that the information-processing mechanisms inside our heads would serve a plurality of functions, not just one. Of course, advocates of hierarchical Bayesian models are not alone in assuming that the function of cognition is optimal inference. Summarising the traditional view of the mind in Western philosophy, for example, Fodor (2001, p. 68) writes that "the proper function of cognition is… the fixation of true beliefs." Nevertheless, I think that there are good reasons for being extremely sceptical of this view. As before, any exhaustive defence of this scepticism lies beyond the scope of the current paper. Instead, I will merely state two powerful considerations in its defence.

First, Mercier and Sperber (2011, 2017) have recently put forward a persuasive defence of the view that reasoning mechanisms in human beings did not evolve to facilitate individual epistemological functions—for example, statistically optimal inference. Instead, they evolved to facilitate *argument*. Specifically, Mercier and Sperber contend that reasoning mechanisms evolved to facilitate the *persuasion* of others and the *evaluation* of *their* arguments. They situate this argument in the context of the evolution of communication, and point out that the traditional theory of the function of reasoning mechanisms struggles to make sense of a whole range of conspicuous features—and failures—of human cognition. By contrast, an "argumentative theory of reasoning" of the sort that they endorse actively *predicts* phenomena such as confirmation bias and motivated reasoning: if a core function of reasoning is not to arrive at the truth, but to bring conspecifics round to one's point of view, these "biases" are positively adaptive. Further, their view of reasoning predicts that individuals would be substantially *better* at evaluating *other* people's beliefs and arguments than evaluating their own, which is exactly what the evidence suggests (see Mercier and Sperber, 2011 for a review).

If this is right, reasoning mechanisms in human beings might be adapted to a goal that the impartial mechanism of Bayesian updating would not serve. In addition, however, many researchers have long argued that some of our beliefs are not the upshot of inferential mechanisms *at all*, whether argumentative or otherwise. Rather, "beliefs have a *social* as well as an *inferential* function" (Pinker, 2005, p. 18, my emphasis; see also Tooby and Cosmides, 2015).[17] On this view, some of the mechanisms underlying belief fixation might have *no* inferential function. Instead, their function is to adapt the agent to her local tribe or community. Again, any

[16] I have heard this argument repeated numerous times in conversation, but never explicitly in print, although Anderson (1990) comes close, as do Oaksford and Chater (2009).

[17] An anonymous reviewer helpfully points out the excellent work within the Bayesian brain hypothesis on social cognition (e.g. Friston and Frith, 2015a, 2015b). This research still focuses exclusively on *inference*, however.

thorough defence of this claim lies beyond the scope of the current paper. Nevertheless, there are compelling reasons for taking it seriously. For example, it has the immediate advantage of *explaining* the data that Fletcher and Frith (2009, p. 52) themselves take as given—namely, that "for the most part, people do not depart from the beliefs of the herd." To take only one illustration, people's beliefs concerning matters of politically important scientific questions (e.g. anthropogenic climate change) are much better predicted by political affiliation than general scientific knowledge or intelligence, suggesting that the chief function of such beliefs is to provide a "sense of identity enabled by membership in a community defined by particular cultural commitments," not *truth* (Kahan, 2015, p. 1). On this view, running an approximately optimal statistical inference engine inside one's head as the exclusive filtering mechanism for beliefs might be positively *maladaptive*—at least in the ancestral environments in which such mechanisms evolved. Instead, beliefs serve a constellation of functions—they signal group membership, they adapt one's behaviour to local social norms and behaviours, "they reflect commitments of loyalty and solidarity to one's coalition" (Pinker, 2005, p. 18), they indicate one's social value (Kurzban, 2010), and so on—whose outcomes are wholly orthogonal to optimal inference.

Of course, these views are controversial. Nevertheless, they are also extremely influential, and they have received substantial defences in disciplines such as evolutionary, cognitive, and social psychology. Despite this, hierarchical Bayesian models of delusion simply take as given that the function of belief fixation is optimal inference. Given the paucity of *evidence* for the claim that Bayes optimal inference underlies belief fixation, and the apparent abundance of evidence against it, one should at least take seriously the idea that the mechanisms underlying belief were not *designed* to produce impartial models of the world—and would thus not be well-served by Bayesian brains.

5.4. Responses

I have argued that the evidence for a Bayesian account of belief fixation is weak, that there is substantial evidence that such mechanisms are not exclusively Bayesian, and that there is good evolutionary reason to *expect* that they would not be exclusively Bayesian. Given this, attempts to explain delusional beliefs "in terms of a disturbed hierarchical Bayesian framework" (Fletcher and Frith, 2009, p. 48) should be viewed with great suspicion. Such explanations violate the methodological principle at the heart of medical psychiatry—namely, to explain psychopathologies in terms of dysfunctions in the mechanisms that underlie *healthy* functioning—and they risk mistakenly exaggerating the extent to which delusional beliefs diverge from those that are present in the healthy population.

I think that there are three possible responses one might make to the argument of this section.

First, one might simply deny the substantive points that I have raised. For example, one might argue that there is good evidence that the mechanisms underlying belief fixation are Bayesian, that apparent counter-evidence can be explained away, and that we have good reason—evolutionary reason, for example—to expect human belief fixation to be Bayes' optimal.

Of course, I cannot rule this out conclusively—and, given the relatively primitive state of the cognitive sciences, it would be presumptuous to claim that we know with any certainty that a Bayesian view of belief fixation is mistaken. Nevertheless, I hope I have at least shown that this position faces an uphill battle—that it must contend with the paucity of evidence in its defence, and evolutionary considerations that seem to weigh heavily against it. Further, even if one cannot decisively rule out a Bayesian view of belief fixation, this does not mean that it should be *assumed* for the purposes of psychiatric theorizing. If one wants to explain delusional beliefs in terms of dysfunctions in a process of Bayesian inference, we should have *good reason* to accept a Bayesian view of belief fixation. If I am right, we do not have this good reason.

A second and more interesting response *concedes* that many of the mechanisms underlying belief fixation are not Bayesian, but argues that these non-Bayesian aspects of cognition are irrelevant to the explanation of delusional beliefs. On this view, even though not *all* information processing within the brain is Bayesian, *some* of it is, and this component is theoretically relevant to the explanation of delusions. This is an interesting suggestion. So far, however, I have not seen any concerted effort to defend it, or provide reasons or evidence in its defence. As such, I hope that advocates of hierarchical Bayesian models of delusion will actively explore its plausibility as a possible response to the objections that I have raised here.

A final response is provided by an anonymous reviewer, who writes:

> "The third and final response to the arguments of the previous section is that the notion of non-Bayesian belief fixation is itself a category error. Indeed, mathematically, the complete class theorem precludes any decision from being non-Bayesian. In other words, no matter how odd or untypical a decision, choice or inference, there is some set of prior beliefs that renders it Bayes optimal. The question then reduces to why the subject evinced these priors."

This response has proven popular in conversation with advocates of Bayesian models of delusion, and is reflected, for example, in recent articles by Schwartenbeck et al. (2015) and Parr, Rees, and Friston (2018). I think that it is deeply problematic. The fact that one can construct a model that *represents* any given decision, choice or inference as Bayes optimal does not entail that any given decision, choice or inference *is* Bayes optimal. One can model the behaviour of a simple look-up table as Bayes optimal (Maloney and Mamassian, 2009) and we can explicitly program systems *not* to update representations in accordance with Bayes' theorem (Danks, 2014, p. 189). Constructing Bayesian models of such non-Bayesian systems reflects the post hoc ingenuity of the modeller, not the nature of the systems being modelled. The infinite malleability of Bayesian modelling in the face of any possible evidence thus requires that we have some reason to endorse the Bayesian approach independent of its capacity to accommodate evidence. The standard reason given in the literature is that Bayesian updating is optimal or rational (Anderson, 1990; Oaksford and Chater, 2007). As Frith and Friston (2013, p. 5) write, the "mechanism for achieving [optimality] is well captured in a Bayesian framework" (Frith and Friston, 2013, p. 5). As such, widespread evidence of suboptimality, cognitive biases and psychological functions orthogonal to

optimal inference *do* provide evidence against a Bayesian view of cognition. Specifically, they undermine the reason for drawing on Bayesian models in the first place.

6. Conclusion

Hierarchical Bayesian models of delusion have recently enjoyed great popularity, and are often touted as one of the important success stories of the emerging field of computational psychiatry. I have argued that they are significantly less promising than is widely believed. Specifically, the two core theoretical components of such models—their hierarchical and Bayesian elements—confront substantial challenges that have not received sufficient attention in the literature. If there is more to thought than a single information-processing inferential hierarchy, and the mechanisms underlying belief fixation are not Bayesian, attempts to explain *delusional beliefs* in terms of dysfunctions in mechanisms underlying hierarchical Bayesian inference are unpromising.

If this is right, it is not just hierarchical Bayesian models of *delusion* that are in trouble, but hierarchical Bayesian models of the brain more generally—at least when the latter models are supposed to explain "perception and action and everything mental in between" (Hohwy, 2013, p. 1). This raises the obvious question: why have I targeted the explanation of delusional beliefs in this paper, and not just focused directly on hierarchical Bayesian models of neural information-processing more generally? There are two reasons for this.

First, although I have focused mostly on predictive processing, the objections that I have raised are in many places much more general. Specifically, the idea of explaining delusional beliefs by appeal to Bayesian inference is a trend that exists outside of predictive processing accounts (Coltheart et al., 2010; Davies and Egan, 2013; McKay, 2012), as is the appeal to an inferential hierarchy with bi-directional message-passing in place of a functional distinction between hallucinations and delusions (Denève and Jardri, 2016). As such, my critique does not just concern predictive processing, but these broader trends in psychiatry.

Second, the attempt to explain delusions is an important area where advocates of predictive processing have moved beyond sweeping claims about the capacity of prediction error minimization to explain *everything* to a concerted attempt to model a specific high-level cognitive phenomenon. As such, hierarchical Bayesian models of delusion provide an important *test case* for predictive processing. I have argued that it fails this test case—that the attempt to exploit predictive processing to explain delusional beliefs confronts serious problems of detail and clarity, alongside a lack of evidence and compelling justifications for core claims.

If there is an overarching lesson to be gleaned from the argument of this paper, it is this: if computational psychiatry is to fulfil its promise of illuminating the information-processing mechanisms underlying psychiatric disorders, it should consider abandoning *global* theories of brain function and *optimality* models of cognition in favour of a much more substantial engagement with research from other fields, especially cognitive and social psychology and evolutionary biology. This engagement, I believe, would reveal the brain's computational mechanisms to be subservient not to a single epistemological goal, but to a plurality of functions—often imperfectly realised—in the life of an exceedingly complex social primate.

Acknowledgements

This work was supported by the Arts and Humanities Research Council. I would like to thank Jakob Hohwy, Richard Holton, Marcella Montagnese, and Stephen Gadsby, as well as two anonymous referees, for helpful discussion and comments.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.concog.2018.03.003.

References

Adams, R., Stephan, K., Brown, H., Frith, C., & Friston, K. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry, 4*. http://dx.doi.org/10.3389/fpsyt.2013.00047.

Anderson, J. (1990). *The adaptive character of thought.* Hoboken: Taylor and Francis.

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

Baron, J. (2008). *Thinking and deciding* (4th ed.). Cambridge: Cambridge University Press.

Bortolotti, L. (2010). *Delusions and other irrational beliefs.* Oxford: Oxford University Press.

Bortolotti, Lisa, "Delusion", The Stanford Encyclopedia of Philosophy (Spring 2016 Edition), Edward N. Zalta (ed.), URL = < https://plato.stanford.edu/archives/spr2016/entries/delusion/ >.

Bortolotti, L., & Miyazono, K. (2015). Recent work on the nature and development of delusions. *Philosophy Compass, 10*(9), 636–645.

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

Brown, H., Adams, R., Parees, I., Edwards, M., & Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing, 14*(4), 411–427. http://dx.doi.org/10.1007/s10339-013-0571-3.

Capps, D. (2004). John Nash's delusional decade: A case of paranoid schizophrenia. *Pastoral Psychology, 52*(3), 193–218. http://dx.doi.org/10.1023/b:Pasp.0000010023.58529.95.

Chadwick, P. (1993). The stepladder to the impossible: A first hand phenomenological account of a schizoaffective psychotic crisis. *Journal of Mental Health, 2*(3), 239–250. http://dx.doi.org/10.3109/09638239309003769.

Chambon, V., Pacherie, E., Barbalat, G., Jacquet, P., Franck, N., & Farrer, C. (2011). Mentalizing under influence: abnormal dependence on prior expectations in patients with schizophrenia. *Brain, 134*(12), 3728–3741. http://dx.doi.org/10.1093/brain/awr306.

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

Chekroud, A. (2015). Unifying treatments for depression: An application of the free energy principle. *Frontiers In Psychology, 6*. http://dx.doi.org/10.3389/fpsyg.2015.00153.
Clark, A. (2012). Dreaming the whole cat: Generative models, predictive processing, and the enactivist conception of perceptual experience. *Mind, 121*(483), 753–771. http://dx.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. http://dx.doi.org/10.1017/s0140525x12000477.
Clark, A. (2016). *Surfing uncertainty*. Oxford: Oxford University Press.
Coltheart, M. (2007). The 33rd Sir Frederick Bartlett lecture: Cognitive neuropsychiatry and delusional belief. *The Quarterly Journal of Experimental Psychology, 60*(8), 1041–1062 [RMR].
Coltheart, M. (2013). On the distinction between monothematic and polythematic delusions. *Mind & Language, 28*(1), 103–112. http://dx.doi.org/10.1111/mila.12011.
Coltheart, M. (2017). The assumptions of cognitive neuropsychology: Reflections on Caramazza (1984, 1986). *Cognitive Neuropsychology, 1–6*. http://dx.doi.org/10.1080/02643294.2017.1324950.
Coltheart, M., Menzies, P., & Sutton, J. (2010). Abductive inference and delusional belief. *Cognitive Neuropsychiatry, 15*(1–3), 261–287. http://dx.doi.org/10.1080/13546800903439120.
Corlett, P., & Fletcher, P. (2015). Delusions and prediction error: Clarifying the roles of behavioural and brain responses. *Cognitive Neuropsychiatry, 20*(2), 95–105. http://dx.doi.org/10.1080/13546805.2014.990625.
Corlett, P., Honey, G., & Fletcher, P. (2016). Prediction error, ketamine and psychosis: An updated model. *Journal of Psychopharmacology, 30*(11), 1145–1155. http://dx.doi.org/10.1177/0269881116650087.
Corlett, P., Taylor, J., Wang, X., Fletcher, P., & Krystal, J. (2010). Toward a neurobiology of delusions. *Progress in Neurobiology, 92*(3), 345–369. http://dx.doi.org/10.1016/j.pneurobio.2010.06.007.
Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Cambridge: The MIT Press.
Davies, M., & Egan, A. (2013). Delusion: Cognitive approaches, bayesian inference, and compartmentalization. In K. W. M. Fulford, M. Davies, R. G. T. Gipps, G. Graham, J. Sadler, G. Stanghellini, & T. Thornton (Eds.). *The Oxford handbook of philosophy of psychiatry*. Oxford: Oxford University Press.
Denève, S., & Jardri, R. (2016). Circular inference: Mistaken belief, misplaced trust. *Current Opinion in Behavioral Sciences, 11*, 40–48. http://dx.doi.org/10.1016/j.cobeha.2016.04.001.
Dennett, D. (1995). *Darwin's dangerous idea*. New York: Simon & Schuster.
Ellis, H., & Young, A. (1990). Accounting for delusional misidentifications. *The British Journal of Psychiatry, 157*(2), 239–248. http://dx.doi.org/10.1192/bjp.157.2.239.
Ernst, M., & Banks, M. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429–433. http://dx.doi.org/10.1038/415429a.
Felleman, D., & Van Essen, D. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*(1), 1–47. http://dx.doi.org/10.1093/cercor/1.1.1.
Fineberg, S., & Corlett, P. (2016). The doxastic shear pin: Delusions as errors of learning and memory. *Cognitive Neuropsychiatry, 21*(1), 73–89. http://dx.doi.org/10.1080/13546805.2015.1136206.
Firestone, C., & Scholl, B. (2015). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral And Brain Sciences, 39*. http://dx.doi.org/10.1017/s0140525x15000965.
Fletcher, P., & Frith, C. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience, 10*(1), 48–58. http://dx.doi.org/10.1038/nrn2536.
Fodor, J. (1983). *The modularity of mind*. Cambridge: The MIT Press.
Fodor, J. (2001). *The mind Doesn't work that way: The scope and limits of computational psychology*. Cambridge: The MIT Press.
Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of The Royal Society B: Biological Sciences, 360*(1456), 815–836. http://dx.doi.org/10.1098/rstb.2005.1622.
Friston, K. (2008). Hierarchical models in the brain. *PLoS Computational Biology, 4*(11), e1000211.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. http://dx.doi.org/10.1038/nrn2787.
Friston, K. J., & Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior, 68*, 129–143. http://dx.doi.org/10.1016/j.cortex.2015.03.025.
Friston, K. J., & Frith, C. (2015). A Duet for one. *Consciousness and Cognition, 36*, 390–405.
Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017a). Active inference: A process theory. *Neural Computation, 29*(1), 1–49. http://dx.doi.org/10.1162/neco_a_00912.
Friston, K., Stephan, K., Montague, R., & Dolan, R. (2014). Computational psychiatry: The brain as a phantastic organ. *The Lancet Psychiatry, 1*(2), 148–158. http://dx.doi.org/10.1016/s2215-0366(14)70275-5.
Friston, K., Rosch, R., Parr, T., Price, C., & Bowman, H. (2017b). Deep temporal models and active inference. *Neurosci Biobehav Rev. 77*, 388–402.
Frith, C. D., & Friston, K. J. (2013). False perceptions and false beliefs: Understanding schizophrenia. *Neurosciences and the Human Person: New Perspectives on Human Activities, 121*, 1–15 [RMR].
Gadsby, S., Williams, D. (forthcoming). Delusional thoughts and bayesian brains: Promises and pitfalls.
Gerrans, P. (2014). *The measure of madness*. Cambridge, Massachusetts: MIT Press.
George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLoS Comput. Biol. 5*, e1000532.
Glymour, C. (2011). Osiander's psychology. *Behavioral and Brain Sciences, 34*(04), 199–200. http://dx.doi.org/10.1017/s0140525x11000276.
Goodfellow, I., Bengio, Y., & Courville, A. (2017). *Deep learning*. Cambridge, Mass: The MIT Press.
Griffin, J., & Fletcher, P. (2017). Predictive processing, source monitoring, and psychosis. *Annual Review of Clinical Psychology, 13*(1), 265–289. http://dx.doi.org/10.1146/annurev-clinpsy-032816-045145.
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. http://dx.doi.org/10.1016/j.tics.2010.05.004.
Hahn, U. (2014). The Bayesian boom: Good thing or bad? *Frontiers in Psychology, 5*. http://dx.doi.org/10.3389/fpsyg.2014.00765.
Harrison, L., Bestmann, S., Rosa, M. J., Penny, W., & Green, G. G. R. (2011). Time scales of representation in the human brain: Weighing past information to predict future events. *Frontiers in Human Neuroscience, 5*, 37. http://dx.doi.org/10.3389/fnhum.2011.00037.
Hawkins, J., & Blakeslee, S. (2005). *On intelligence*. New York: Henry Holt and Company.
Hemsley, D., & Garety, P. (1986). The formation of maintenance of delusions: A Bayesian analysis. *The British Journal of Psychiatry, 149*(1), 51–56. http://dx.doi.org/10.1192/bjp.149.1.51.
Hinton, G. (1990). Mapping part-whole hierarchies into connectionist networks. *Artif. Intell. 46*(1–2), 47–75. http://dx.doi.org/10.1016/0004-3702(90)90004-j.
Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3*. http://dx.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. http://dx.doi.org/10.1111/nous.12062.
Hohwy, J. (2017). Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization. *Consciousness and Cognition, 47*, 75–85. http://dx.doi.org/10.1016/j.concog.2016.09.004.
Hoorens, V. (1993). Self-enhancement and superiority biases in social comparison. *European Review of Social Psychology, 4*(1), 113–139. http://dx.doi.org/10.1080/14792779343000040.
Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1), 106–154. http://dx.doi.org/10.1113/jphysiol.1962.sp006837.
Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences, 34*, 169–188. http://dx.doi.org/10.1017/S0140525X10003134.
Kahan, D. M. (2015). Climate-science communication and the measurement problem. *Political Psychology, 36*(S1), 1–43. http://dx.doi.org/10.1111/pops.12244.

Kahneman, D. (2011). *Thinking, fast and slow.* New York: Farrar, Straus and Giroux.
Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Computational Biology, 4*, e1000209. http://dx.doi.org/10.1371/journal.pcbi.1000209.
Knill, D., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719. http://dx.doi.org/10.1016/j.tins.2004.10.007.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. http://dx.doi.org/10.1037//0033-2909.108.3.480.
Kurzban, R. (2010). *Why everyone (else) is a hypocrite: Evolution and the modular mind.* Oxford: Princeton University Press.
Lawson, R., Rees, G., & Friston, K. (2014). An aberrant precision account of autism. *Frontiers in Human Neuroscience, 8*. http://dx.doi.org/10.3389/fnhum.2014.00302.
Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America, A, 20*, 1434–1448.
Maher, B. (1974). Delusional thinking and perceptual disorder. *Journal of Individual Psychology, 30*, 98–113.
Maher, B. (2001). Delusions. In P. B. Sutker, & H. E. Adams (Eds.), Comprehensive handbook of psychopathology. New York: Kluwer Academic/Plenum Publishers.
Maloney, L. T., & Mamassian, P. (2009). Bayesian decision theory as a model of visual perception: Testing Bayesian Transfer. *Visual Neuroscience, 26*, 147–155.
Marcus, G. (2009a). How does the mind work? Insights from biology. *Topics in Cognitive Science, 1*(1), 145–172. http://dx.doi.org/10.1111/j.1756-8765.2008.01007.x.
Marcus, G. (2009b). *Kluge.* London: Faber.
Marcus, G. (2015). The computational brain. In G. Marcus, & J. Freeman (Eds.). *The future of the brain: Essays by the World's leading neuroscientists* (pp. 205–219). Oxford: Princeton University Press.
Marcus, G., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science, 24*(12), 2351–2360. http://dx.doi.org/10.1177/0956797613495418.
Marr, D. (1980). *Vision.* New York: Freeman.
Mathys, C., Daunizeau, J., Iglesias, S., Diaconescu, A., Weber, L., Friston, K., et al. (2012). Computational modeling of perceptual inference: A hierarchical Bayesian approach that allows for individual and contextual differences in weighting of input. *International Journal Of Psychophysiology, 85*(3), 317–318. http://dx.doi.org/10.1016/j.ijpsycho.2012.06.077.
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, 8*. http://dx.doi.org/10.3389/fnhum.2014.00825.
Mckay, R. (2012). Delusional Inference. *Mind & Language, 27*(3), 330–355. http://dx.doi.org/10.1111/j.1468-0017.2012.01447.x.
Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral And Brain Sciences, 34*(02), 57–74. http://dx.doi.org/10.1017/s0140525x10000968.
Mercier, H., & Sperber, D. (2017). *The enigma of reason.* London: Penguin.
Montague, P., Dolan, R., Friston, K., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences, 16*(1), 72–80. http://dx.doi.org/10.1016/j.tics.2011.11.018.
Murphy, D. (2006). *Psychiatry in the scientific image.* Cambridge: MIT Press.
Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. http://dx.doi.org/10.1037//1089-2680.2.2.175.
Notredame, C., Pins, D., Deneve, S., & Jardri, R. (2014). What visual illusions teach us about schizophrenia. *Frontiers in Integrative Neuroscience, 8*. http://dx.doi.org/10.3389/fnint.2014.00063.
Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303–330. http://dx.doi.org/10.1007/s11109-010-9112-2.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality.* Oxford: Oxford University Press.
Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences, 32*(01), 69–84. http://dx.doi.org/10.1017/s0140525x09000284.
Parr, T., Rees, G., & Friston, K. (2018). Computational neuropsychology and Bayesian inference. *Frontiers in Human Neuroscience*. http://dx.doi.org/10.3389/fnhum.2018.00061.
Penny, W. (2012). Bayesian models of brain and behaviour. *ISRN Biomathematics,* 1–19. http://dx.doi.org/10.5402/2012/785791.
Pinker, S. (2005). So how does the mind work? *Mind and Language, 20*(1), 1–24. http://dx.doi.org/10.1111/j.0268-1064.2005.00274.x.
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. http://dx.doi.org/10.4103/0972-6748.110958.
Pylyshyn, Z. (2003). *Seeing and visualizing: It's not what you think.* Cambridge: The MIT Press.
Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87. http://dx.doi.org/10.1038/4580.
Ross, R., McKay, R., Coltheart, M., & Langdon, R. (2016). Perception, cognition, and delusion. *Behavioral And Brain Sciences, 39*. http://dx.doi.org/10.1017/s0140525x15002691.
Sanborn, A., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences, 20*(12), 883–893. http://dx.doi.org/10.1016/j.tics.2016.10.003.
Schmack, K., Gomez-Carrillo de Castro, A., Rothkirch, M., Sekutowicz, M., Rossler, H., Haynes, J., et al. (2013). Delusions and the role of beliefs in perceptual inference. *Journal of Neuroscience, 33*(34), 13701–13712. http://dx.doi.org/10.1523/jneurosci.1778-13.2013.
Schwartenbeck, P., FitzGerald, T. H. B., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., et al. (2015). Optimal inference with suboptimal models: Addiction and active Bayesian inference. *Medical Hypotheses, 84*(2), 109–117. http://dx.doi.org/10.1016/j.mehy.2014.12.007.
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. http://doi.org/10.15502/9783958570108.
Seth, A., Suzuki, K., & Critchley, H. (2012). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology, 2*. http://dx.doi.org/10.3389/fpsyg.2011.00395.
Stanovich, K., & West, R. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning, 13*(3), 225–247. http://dx.doi.org/10.1080/13546780600780796.
Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279–1285. http://dx.doi.org/10.1126/science.1192788.
Teufel, C., & Fletcher, P. (2016). The promises and pitfalls of applying computational models to neurological and psychiatric disorders. *Brain, 139*(10), 2600–2608. http://dx.doi.org/10.1093/brain/aww209.
Teufel, C., Kingdon, A., Ingram, J., Wolpert, D., & Fletcher, P. (2010). Deficits in sensory prediction are related to delusional ideation in healthy individuals. *Neuropsychologia, 48*(14), 4169–4172. http://dx.doi.org/10.1016/j.neuropsychologia.2010.10.024.
Tooby, J. & Cosmides, L. (2015). The theoretical foundations of evolutionary psychology. In Buss, D. M. (Ed.), The Handbook of Evolutionary Psychology, Second edition. Volume 1: Foundations. (pp. 3–87). Hoboken, NJ: John Wiley & Sons.
Vance, J. (2015). Review of the predictive mind. *Notre Dame Philosophical Reviews*.
Weiss, Y., Simoncelli, E., & Adelson, E. (2002). Motion illusions as optimal percepts. *Nature Neuroscience, 5*(6), 598–604. http://dx.doi.org/10.1038/nn0602-858.
Williams, D. (2017). Predictive processing and the representation wars. *Minds and Machines*. http://dx.doi.org/10.1007/s11023-017-9441-6.
Williams, D. (forthcoming a). Predictive Coding and Thought. Synthese.
Williams, D. (forthcoming b). Hierarchical Minds and the Perception/Cognition Divide.
Williams, D., & Colling, L. (2017). From symbols to icons: The return of resemblance in the cognitive neuroscience revolution. *Synthese*. http://dx.doi.org/10.1007/s11229-017-1578-6.
Zellner, A. (1988). Optimal information processing and Bayes's theorem. *The American Statistician, 42*(4), 278. http://dx.doi.org/10.2307/2685143.