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).