Delusional predictions and explanations

Parrott, Matthew

DOI:
10.1093/bjps/axz003

Document Version
Peer reviewed version

Citation for published version (Harvard):
Parrott, M 2019, 'Delusional predictions and explanations', *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz003

Delusional Predictions and Explanations

Matthew Parrott

Abstract: In both cognitive science and philosophy, many theorists have recently appealed to a predictive processing framework to offer explanations of why certain individuals form delusional beliefs. One aim of this essay will be to illustrate how one could plausibly develop a predictive processing account in different ways to account for the onset of different kinds of delusions. However, the second aim of this essay will be to discuss two significant limitations of the predictive processing framework. First, I shall draw on the structure of explanatory why-questions to argue that predictive processing theories can only partially explain the formation of delusional beliefs. Second, I shall argue that predictive processing theories cannot explain how implausible delusional hypotheses are generated. Yet understanding why an agent even generates a delusional hypothesis is a crucial step to understanding why she eventually comes to believe it. The final section of the essay presents three alternative ways in which the process of hypothesis generation might be functionally divergent in cases of delusional cognition.

1 Introduction

It is not clear what causes certain individuals to form delusional beliefs. Many computational psychiatrists, and some philosophers, have tried to account for the formation of delusional beliefs by appealing to the recently developed predictive processing framework. As we shall see, this is a very general theoretical framework for modelling mental processes (Clark [2016]; Hohwy [2013]).[1] Put simply, the central claim of predictive processing is that the brain works by encoding a model of the world that it uses to continuously make predictions about the sensory input it is receiving, and then it generates a kind of error signal anytime those predictions are violated. This prediction error signal indicates a divergence between the actual sensory input and what the brain predicted, and it functions as a kind of feedback to adjust the brain's predictive model of the external world (Corlett *et al.* [2016]). As Karl Friston describes it, 'the brain is an inference machine that actively predicts and explains its sensations.' (Friston [2010], p. 129; cf. Hohwy [2016]) On this approach, our beliefs about the external world just are our best predictions about the distal causes of our sensations, and we adopt new beliefs in order to explain unexpected or unpredicted aspects of the incoming sensory signal.

By extension, according to the predictive processing framework, delusional beliefs are adopted in order to explain the occurrence of some irregular or unexpected aspect of the incoming sensory signal (cf. Corlett *et al.* [2016]; Corlett *et al.* [2010]; Frith and Friston [2013];

[1] In the literature, this is called both 'predictive processing' and 'predictive coding'. In this essay, I shall use 'predictive processing' to refer to any theoretical approach or model that claims neural computations are carried out by means of a system updating a predictive model of the world in response to internally generated error signals. By contrast, I think of 'predictive coding' as a specific data-compression strategy (for some further discussion see Clark [2016], Chapter 1). The two terms therefore have different senses (or intensions) even if they are co-extensive (for instance, even if it is the case that a system minimizes prediction error just in case it can be accurately described in terms of a 'predictive coding' strategy).

Hohwy [2015]).[2] This idea fits nicely with experimental data suggesting that some kind of irregular experience is implicated in the onset of several different delusions (Coltheart *et al.* [2011]; Langdon and Bayne [2010]). It also illustrates how predictive processing is a modern descendent of the following influential idea from the history of cognitive neuropsychology.

Many years ago, Brendan Maher presented the so-called 'explanationist' doctrine that delusional beliefs are formed in order to explain unusual experiences:

> Strange events, felt to be significant, demand explanation…In brief then, a delusion is a hypothesis designed to explain unusual perceptual phenomena and developed through the operation of normal cognitive processes. (Maher [1974], p. 103; cf. Stone and Young [1997])

Predictive processing theories accept Maher's central idea that delusional beliefs are adopted because they explain 'strange' or 'unusual' phenomena, but they diverge from Maher on two important points. First, Maher thinks that the 'unusual perceptual phenomena' that need to be explained occur consciously, but predictive processing theorists tend to maintain that delusional

[2] I shall assume that one aspect of delusional cognition is the formation of a delusional belief. This is the standard way cognitive neuropsychiatry conceives of delusions. There are some philosophers who question this, and there is therefore an ongoing debate among philosophers about whether delusions are beliefs or some other type of mental state. I shall not address this debate, other than to say that my own view is that we ought to not think of a delusion as equivalent to a type of mental state, whether that state is 'doxastic', 'imaginative', or something 'in-between'. Rather, I think it is better to think of delusion as a pattern of cognition, part of which includes the adoption of a strange belief. Nevertheless, I also think there are good arguments in favour of a 'doxastic conception' of the mental state that one forms in cases of delusion (Bayne and Pacherie [2004]; Bortolotti [2009]).

beliefs are adopted in order to explain non-conscious sensory experiences.[3] Second, predictive processing theorists resist drawing the distinction that Maher presupposes between 'cognitive processes' and perceptual processes. Instead, they believe that there is a uniform type of neural processing occurring across the brain and so 'the same process that accounts for abnormal perception can also account for abnormal belief.'[4] (Corlett *et al.* [2016], p. 1148; Corlett and Fletcher [2015]; Fletcher and Frith [2009]). Nevertheless, these theorists also think that, in cases of delusion, there is some kind of disturbance or disruption in an agent's neural processing. This is the reason why an agent adopts a delusional belief in order to 'explain away' an unexpected aspect of the sensory signal (Corlett and Fletcher [2015]; Corlett *et al.* [2016]; Hohwy [2015]).

---

[3] This outlook is shared by theorists who are not attracted to predictive processing. For instance, Coltheart and colleagues think that, in many cases of delusion, 'something abnormal occurs of which a person is not conscious.' (Coltheart *et al.* [2010], p. 264)

[4] Partly because they reject the distinction between perception and cognition, many predictive processing theorists identify themselves as defending a 'one-factor' theory of delusion formation, in contrast to prominent 'two-factor' theories of delusion formation (for example, Coltheart *et al.* [2018]; Coltheart *et al.* [2011]). A 'one-factor' theorist aims to explain the onset of a delusional belief by appealing to only one cognitive or neurobiological impairment, or anomaly (one 'factor'). For instance, Corlett and colleagues claim that, on their predictive processing model, 'a single deficit' is able to explain the onset of a delusional belief (Corlett *et al.* [2010], p. 357) However, even if some accounts appeal to only a 'single deficit', it would be a mistake to think that any predictive processing account of delusional belief must be a one-factor theory. The predictive processing framework is consistent with the central tenet of a 'two-factor' approach, namely that two distinct impairments or anomalies are causally implicated in the onset of delusional belief. Indeed, there are theorists who have developed predictive processing models of delusional belief formation that appeal to two deficits or 'factors' (for instance, McKay [2012]), and, in section 2, I shall present predictive processing accounts for Capgras delusion and for anosognosia for hemiplegia that appeal to multiple impairments. We shall return to the explanatory power of two-factor theories in section 4.

Because the predictive processing framework conceptualises the formation of a delusional belief as an explanatory process, it can be regarded a modern version of explanationism.[5]

Interestingly, different theorists have proposed different accounts of precisely how predictive processing is impaired or disrupted in cases of delusion. For example, Philipp Corlett proposes that 'delusions form when [prediction error] is signalled inappropriately with high precision, such that it garners new and aberrant learning.' (Corlett [2018], p. 54; Feeney *et al.* [2017]; cf. Sterzer, *et al.* [2018]). This suggests that delusions are caused because of some kind of impairment or disturbance in the neurobiological mechanisms that generate error signals, or in the mechanisms that determine the 'precision' of those signals (assuming that these are distinct), or in both.[6] Alternatively, Jakob Hohwy has proposed that delusions may be caused by a kind of functional overreliance on prior expectations or predictions, which leads to beliefs being 'biased in favour of internal models of the world.' (Hohwy [2015], p. 307; Hohwy [2013]; cf. Corlett *et al.* [2010]). This suggests that delusions are caused by some kind of impairment or disturbance in

[5] Indeed, Sterzer and colleagues classify Maher as a 'precursor' to contemporary predictive processing theories (Sterzer *et al.* [2018]; cf. Corlett *et al.* [2016])

[6] Think of 'precision' as a measure of the reliability of a prediction error signal (mathematically precision is the inverse of the variance of the signal). From a functional point of view, tagging a signal as 'precise' is roughly equivalent to labelling it as 'reliable', and the converse is true for signals that are tagged as 'imprecise' (cf. Davies *et al.* [2018]). Within a predictive processing system, unreliable predictions are attenuated and therefore have comparatively less influence on the dynamics of the system. At a neurobiological level, precision is realized by dampening down the post-synaptic gain on the relevant error units (Hohwy [2013], Chapter 3). The notion of precision will be discussed further in the following section.

the neurobiological mechanisms that encode predictions, or in the mechanisms that estimate the reliability of those predictions, or in both.[7]

The fact that there is divergence among theorists as to how exactly predictive processing is impaired or disrupted in cases of delusion should not be taken to count against the predictive processing approach. Predictive processing was not designed to explain the onset of delusional beliefs. Rather, it was conceived as a general theoretical framework for thinking about how the brain functions. As such, the way the entire framework conceptualizes psychological and neurological processes should be theoretically useful for explaining the onset of delusional beliefs. But adopting this theoretical framework does not commit us to thinking that there must be a single impairment or abnormality responsible for every case of delusional belief. Rather, we should be open to the possibility that different, and multiple, irregularities in predictive processing systems could be implicated in the onset of different delusional beliefs.

The aim of this essay is to assess the prospects of the predictive processing framework for explaining why certain individuals form delusional beliefs. In the following section, I shall present the basics of predictive processing and a portion of the corroborating experimental evidence that supports adopting it as a general theoretical framework. In section 2, I shall present several different predictive processing accounts for different types of delusions. As we shall see, one of the strengths of the predictive processing framework is that it is flexible enough to causally explain different kinds of delusional beliefs in different ways. Nevertheless, the remainder of this essay will focus on bringing out two limitations of the predictive processing framework. In section 3, I shall argue that the explanatory power of predictive processing

[7] These two proposals are not exhaustive. For instance, a third suggestion would be that the emergence of delusions 'results from a decreased precision of priors' (Sterzer *et al.* [2018], p. 5). As a result, predictive processing would be driven more strongly than normal by the incoming sensory signal. Sterzer and colleagues propose that this causes psychotic symptoms such as delusion.

accounts depends crucially on what one takes to be the relevant explanatory contrast. I shall also argue that shifting away from the default contrast, shows that a predictive processing account can only be a partial explanation of the formation of a delusional belief. In section 4, I shall turn to the second limitation, which is that the exceptional implausibility of delusional beliefs makes them extremely poor candidate explanations for even highly unusual experiences. For this reason, we need to understand why an agent would even consider a delusional belief to be a *potential* explanation—in other words we need to understand how the mind generates explanatory hypotheses. Indeed, as will see, explaining why delusional agents consider implausible candidate explanations poses a serious challenge for any type of explanationist theory of delusion formation, including influential two-factor theories. (Coltheart *et al.* [2018]; Coltheart *et al.* [2011]; Davies *et al.* [2001]) In section 5, I shall present three proposals for how hypothesis generation might be functionally anomalous in cases of delusion. These three speculative proposals suggest plausible ways in which the neurocognitive processing underlying hypothesis generation might be impaired in cases of delusional cognition.

2 Basics of Predictive Processing

The central innovation of the predictive processing framework is the way it reconceptualises how the brain processes information. According to predictive processing, in cognition, perception, and action, the mind confronts just a single computational problem: how to best minimize error. Like traditional computational theories, the predictive processing framework views mental processes as performing information-processing tasks. However, unlike traditional theories, predictive processing theories maintain that the driving signals of neural computations are only error signals, signals that indicate something has gone wrong, rather than positive information signals. The reason that only error signals are processed is that the brain uses a model of the world 'to predict and fully 'explain away' the driving sensory signal, leaving only any residual

'prediction errors' to propagate information forward within the system.' (Clark [2013], p. 182; Clark [2016]; Hohwy [2013]).

We can think of the brain's predictive model as its best guess about what in the external world is causing its current sensory input. A good model generates a good prediction about the sensory signal, one which matches the actual pattern of neuronal responses caused by incoming stimulation. But, if the brain's predictive model is bad, then there is a mismatch between prediction and actual sensation. This produces an error signal, which has the important function of indicating that the brain's predictive model must be revised in some way. Thus, a prediction error signal causes the brain to refine its model of the world so that future predictions minimize or cancel out that signal (Rao and Ballard [1999]).

The previous description just is how an agent updates her beliefs about the external world, because, within this theoretical framework, the brain's model, at least at 'higher' levels, just is what an agent believes. As Friston describes it, the brain encodes a model of the world that 'can generate predictions, against which sensory samples are tested to update beliefs about their causes.' (Friston [2010], p. 129; Clark [2013]) So, in the predictive processing framework, all beliefs are adopted because they best explain an agent's current sensory experience (Corlett [2018]; Corlett *et al.* [2010]; Hohwy [2016]; Sterzer *et al.* [2018]).

The standard way predictive processing theorists model belief formation is to characterize it as Bayesian inference. An agent's beliefs are modelled mathematically as subjective probabilities, or degrees of confidence, that the agent assigns to various hypotheses, based on her background knowledge. When confronted with surprising evidence, these prior probabilities are adjusted in accordance with Bayes's theorem (Joyce [2003]). This is essentially a rule for determining how confidence in a hypothesis should be changed (i.e., what *posterior* probability

one should assign to a hypothesis) in response to a novel experience. A Bayesian model therefore tells us how an agent ought to update her beliefs in response to some new experience.[8]

Bayesian models are standardly appealed to by predictive processing theorists partly because there is reason to think that, by virtue of minimizing prediction error, the brain approximates optimal Bayesian inference (Clark [2016], Chapter 1; Hohwy [2016]; Hohwy [2015]; Knill and Pouget [2004]). For example, Hohwy writes:

> [Prediction error minimization] is essentially inference to the best explanation, cast in (empirical, variational) Bayesian terms. The winning hypothesis about the world is the one with the highest posterior probability, that is, the hypothesis that best explains away the sensory input, in a context-dependent fashion, under expectations of precision, and with long run, average fit taken into account. ([2016], p. 263)[9]

One crucial feature of the predictive processing framework that Hohwy mentions in this passage is precision. Precision is essentially a measurement of the reliability of a prediction error signal.

---

[8] Bayesian modelling raises many questions about whether the brain actually does implement Bayesian inference (cf. Williams [2018]). One well-known difficulty is that there is substantial evidence that, from a Bayesian perspective, human reasoning and decision-making is suboptimal (for example, Kahneman [2003]; Kahneman [2011]). Prima facie, this speaks against the idea that beliefs are updated by Bayesian inference. However, observed patterns of reasoning and decision-making can be accurately modelled in Bayesian terms by tweaking parameters of a mathematical model. But this raises further questions about the possibility of overfitting one's model to the observed data (Colombo and Series [2012]). Unfortunately, these are issues that cannot be explored further within the confines of this essay.

[9] It is controversial whether Bayesian inference really is compatible with inference to the best explanation. The basic reason to think it is not is that Bayesian confirmation need not track the hypothesis that best explains a piece of evidence. For an argument that Bayesian inference is nonetheless compatible with inference to the best explanation, see (Lipton [2004], Chapter 7).

To oversimply things, imagine that there is just a single neuron, the firing of which realizes a prediction error signal. Over an extended interval, the magnitude of that signal will vary. A highly variable neural signal is imprecise; whereas an invariant signal is precise. Ideally, the brain would update its predictive model of the world on the basis of only precise, or reliable signals. However, actual neural signals are embedded in environments with an abundance of additional neural activity, which introduces a lot of uncertainty and noise. Partly because of this, the brain cannot actually know the precision of any of its neural signals. So, to solve this problem, the brain learns to predict or estimate the precision of prediction error signals (for further discussion of precision estimation, see (Hohwy [2013], Chapter 3). In an optimally functioning system, signals that are estimated to be precise are given more weight (i.e., the brain increases the post-synaptic gain on the neural signal), and signals that are estimated to be imprecise are attenuated. Functionally, this allows the brain both to learn from reliable signals and to accommodate contextual parameters, such as environmental noise in the sensory signal. As we will see in the following section, disruptions in precision estimation figure prominently in leading predictive processing accounts of delusion formation.

According to the predictive processing framework, what we think about the world depends partly on how the world affects our sense organs, but it also depends on our predictions about what the world is like. As we have just seen, these predictions are constantly adjusted by means of a complex, dynamic process involving different parameters, such as the prior probabilities one assigns to candidate hypotheses and the estimated precision of a multitude of error signals. Nevertheless, when everything is functioning as it should, the predictive processing mechanisms embodied by the brain seem to generate fairly coherent beliefs about the world.

Several leading predictive processing theorists have proposed that delusional beliefs arise because of some kind of disturbance in the brain's predictive processing systems. Indeed, Hohwy claims that 'it does not take much disruption or suboptimal prediction error

minimization for the overall model of the world to take a wrong or even pathological turn.' ([2013], p. 225) Before exploring specific proposals, however, it is worth noting that several strands of empirical research support the claim that delusional beliefs are caused by some type of predictive processing disturbance.

First, there is evidence that dopamine underwrites prediction driven learning, and also evidence that striatal dopamine dysregulation contributes to psychosis (Corlett *et al.* 2006; Corlett *et al.* [2009]; Fletcher and Frith [2009]; Howes and Kapur [2009]; Murray *et al.* [2008]). Second, most predictive processing theorists speculate that there are two functionally distinct populations of neurons, one of which encodes predictions and the other of which encodes error signals (Clark [2016], Chapter 1; Friston [2005]). Relatedly, the post-synaptic gain of neurons is thought to encode precision. If these two assumptions are right, then anything that affects the post-synaptic gain of error units would plausibly affect the way brain processes prediction error. In addition to dopamine, the neural mechanisms that primarily affect post-synaptic gain are NMDA and GABA, and all three of these are abnormal in schizophrenia. As Adams and colleagues note, there is evidence of 'abnormal neuromodulation of superficial pyramidal cells', which are the cells thought to encode error signals. Third, abnormalities in these same neural mechanisms are associated with the psychosis-like effects of ketamine (Adams *et al.* [2013]; Corlett *et al.* [2007]; Corlett *et al.* [2016]). Fourth, there is neuroimaging data that shows aberrant activity in the right prefrontal cortex for psychosis and ketamine. This is a brain region correlated with prediction-dependent learning (Corlett *et al.* [2006]; Corlett and Fletcher [2015]). Finally, there is an abundance of behavioural evidence that delusion-prone individuals exhibit abnormalities in probabilistic reasoning. Specifically, they have a strong disposition to 'jump-to conclusions' insofar as they make inferences on the basis of less information than non-delusional subjects (Garety *et al.* [2005]; Garety and Freeman [1999]; So *et al.* [2012]). Bayesian models of this

'jumping-to-conclusions' phenomenon illustrate that a disturbance in predictive processing could plausibly predict these sorts of abnormalities in probabilistic reasoning.[10]

The empirical research I just mentioned lends credibility to the general idea that delusional beliefs are caused by some kind of disturbance or irregularity in predictive processing. However, this is not yet an explanation of delusion formation, only an explanation-sketch. If we wish to fill this sketch in, we need to know more about the nature of the specific disturbances involved in specific cases of delusion. Exactly what sort of predictive processing irregularity might cause someone to adopt a delusional belief, such as the belief that her mother is an alien imposter?

3 Forming Delusional Beliefs

The most popular proposal among predictive processing theorists is that delusional beliefs are formed because of a problem with precision estimation:

> The main conclusion is that a wide range of psychotic symptoms can be explained by a failure to represent the precision of beliefs about the world…The basic idea is that faulty inference leads to false concepts (delusions) or percepts (hallucinations) and that this failure is due to a misallocation of precision to hierarchical representations in the brain. (Adams *et al.* [2013], p. 1; cf. Feeney *et al.* [2017]; Sterzer *et al.* [2018])

[10] There is more empirical research that is relevant to the claim that a predictive processing disturbance is implicated in delusional belief formation, which I lack space to discuss fully in this essay. For good overviews of this work, see (Adams *et al.* [2013]; Fletcher and Frith [2009]; and Clark [2016], Chapter 7).

The basic idea that delusional beliefs are the result of a 'misallocation' of precision has been enormously influential in recent computational psychiatry. Yet there are different ways in which precision could be 'misallocated'

Hohwy proposes that a general expectation of imprecise prediction error signals could explain the onset of delusional beliefs. As he describes it, 'an individual who constantly expects the sensory signal to be noisier than it really is will then tend to be caught in his or her own idiosyncratic interpretations of the input and will find it hard to rectify these interpretations'. (Hohwy [2013], p. 158; Hohwy [2015]) As we have seen, according to the predictive processing framework, error signals that are estimated to be imprecise are normally attenuated. As a result, the brain relies more heavily on its prior predictions about what the world is like – in other words the value assigned to one's prior probabilities is effectively inflated. So, if a delusional agent generally has a much greater expectation than normal of being in a noisy environment, she would rely much more heavily than normal on her internal predictive model of the world. From a dynamic perspective, the agent's model of the world would not be constrained by the incoming sensory signal and, at higher levels, this would result in beliefs that are not shaped by her experiences.

Hohwy further notes that discounting the value of error signals, or, equivalently, overvaluing confidence in one's prior predictions, would plausibly have additional knock-on effects. First, low-valued error signals are typically generated in cases where there is a very small discrepancy between an agent's predictions about the incoming signal and the actual signal. The low-value has a purpose, which is to indicate that one's predictive model must be fine-tuned in a minimal way, not completely scrapped. However, if an agent were expecting a very noisy environment, then low-valued error signals would tend to be treated as background noise. This would further inhibit empirical learning. Second, Hohwy suggests that if one's predictive processing system expects noise, a high-valued error signal 'may appear as more exceptional than

it would to other people because it occurs against an overall background of more subdued prediction error'. ([2013], p. 159). Hohwy suggests that such a signal would attract more attention; it would appear to be as something that absolutely must be explained away.

Someone with an anomalously high expectation of imprecision would continuously ignore sensory feedback. As a result, her predictive model of the world would gradually become less and less constrained by the way the world actually is. As Hohwy suggests, this could eventually lead to a highly 'idiosyncratic interpretation' of the world. This is precisely the sort of developmental trajectory we see in some cases of psychosis when an agent gradually comes to adopt an elaborate system of delusional beliefs (Bentall [2018]). Consider the following passage:

> The game took on huge proportions. Not only was I convinced that my home town was manipulating things, I also thought the nation as a whole was participating in fooling me…In Lake City, I thought adults were playing basketball in teams in the civic centre for me. Nationwide, I thought the President of the United States was going to visit me, so I was very nervous in my apartment, thinking that he would soon arrive. After a few days when he didn't come, I decided that he had changed his mind…Eventually, the game seemed to have gone just too far. I had needed to do something dramatic to stop it! The final straw was the morning when I woke up and felt the shadow of Satan was on my living room floor. Satan, I thought, was beginning to take advantage of the game. (Emmons *et al.* [1997])

In reading this passage, one gets a sense of someone developing a more elaborate delusional belief system over time, eventually resulting in a very bizarre or 'idiosyncratic' world view. In a predictive processing framework, the development of this kind of delusional belief system would, as Hohwy proposes, plausibly be the result of consistently devaluing sensory-based prediction error.

Undervaluing precision is not the only way it could be 'misallocated'. Another possibility, proposed by Andy Clark, is that delusional beliefs are adopted when the precision of error signals is overestimated. Clark remarks that 'sometimes dealing with ongoing, highly weighted

sensory prediction error may require brand new generative models gradually to be formed'. Clark calls this idea the key to a 'better understanding of the origins of hallucinations and delusion'. (Clark [2016], p. 79; cf. Corlett *et al.* [2016]; Corlett *et al.* [2010]). Highly precise error signals normally function to indicate that the brain's predictive model of the world is mistaken and needs revision. Recurring error signals with extremely high estimates of precision would indicate that the model is rather significantly mistaken and needs to be radically revised.[11] In this sense, Clark's suggestion is that a persistent wave of highly precise error signals would cause the brain to make highly significant revisions by virtue of adopting odd beliefs about the world. As he puts it, a predictive processing system would form 'increasingly bizarre hypotheses so as to accommodate the unrelenting waves of (apparently) reliable and salient yet persistently unexplained information'. ([2016], p. 206)

To illustrate how a distorted overestimation of precision might cause an odd delusional belief, consider Capgras delusion (Corlett [2018]; Fineberg and Corlett [2016]). This delusion involves the belief that a familiar person, such as one's mother, is really a qualitatively identical imposter. A number of studies have shown that Capgras delusion is associated with a deficit in an individual's autonomic nervous system, specifically visual presentations of familiar faces do not elicit autonomic arousal, as they do in non-delusional subjects (Bobes *et al.* [2016]; Brighetti *et al.* [2007]; Ellis *et al.* [2000], Hirstein and Ramachandran [1997]). Plausibly, the brain's internal model of the world predicts autonomic arousal to familiar faces, and so an absence of such a

[11] Even if overly precise prediction error signals occur initially in response to predictions about low levels of sensory stimulation, if those signals are estimated to be very precise, they will have repercussions for higher-level predictions. As Fletcher and Frith remark, 'prediction errors will be propagated even further up the system to ever-higher levels of abstraction'. ([2009], p. 55). One question for this proposal is why a hierarchically organised system wouldn't take 'ongoing, highly weighted sensory prediction error' to be an indication of a malfunction within the system.

response would generate a prediction error signal. So, the generation of an error signal in this case would not be dysfunctional, or at least not an additional impairment to whatever causes the lack of autonomic arousal. But what if this signal is over-estimated to be extremely precise? Clark's idea is that this would cause extreme revisions to an agent's model, in this case a discarding of the common-sense belief that this person (who looks just like my mother) really is my mother. This proposal is promising in part because we have evidence that some individuals with damage to ventromedial regions of the frontal cortex also experience faces without autonomic arousal, but do not form delusional beliefs (Tranel *et al.* [1995]). The proposal that Capgras subjects dramatically overestimate precision would allow us to explain this contrast.[12]

Both Clark's and Hohwy's proposals involve the idea that delusional beliefs are caused by problems with precision estimation. At one point, Hohwy hails this as a 'generic account' of delusion formation. ([2013], p. 159). This is an overstatement. Remaining within the predictive processing framework, it is plausible that some delusions arise because of other kinds of disturbances in predictive processing. If this is right, then there may be no fully 'generic account' of delusion formation. Rather, we should expect that the complex dynamics of a predictive processing system could be impaired or disrupted in different ways in different cases of delusion (Sterzer *et al.* [2018]). Before concluding this section, let me outline two further ways in which delusional beliefs could be caused by anomalies in predictive processing.

Some delusions might arise not because of any problem with precision estimation but because a system completely fails to generate an error signal when the brain's predictions fail to match the incoming sensory signal. This might be what happens in cases of anosognosia for hemiplegia. Most frequently, this is a condition in which a person denies the existence of a left-

[12] A Bayesian model of this sort of updating in response to overly precise prediction error signalling is straightforward. For two different ways of developing such a model, see (Coltheart *et al.* [2010] and McKay [2012]).

side motor impairment following a stroke that damages the right-hemisphere of the brain. (Davies *et al.* [2005], Fotopolou *et al.* [2008]). An individual with this condition will report, for example, that she can move her paralysed left arm, or indeed that she *is* moving her left arm, when she is asked to, even though her left arm is obviously paralysed (Berti *et al.* [1998]; Davies *et al.* [2005]).

In a predictive processing framework, an intention or motor command to move one's left arm generates predictions about the future position of one's arm and about the visual and proprioceptive sensory consequences of that movement (Blakemore *et al.* [2002]; Clark [2016], Chapter 4).[13] These predictions are compared with the actual sensory feedback from one's action and, in cases where there is a mismatch, an error signal is generated. Error signals in this domain are thought to facilitate motor control by indicating that the configuration of a bodily movement must be changed.

There is evidence that in anosognosia for hemiplegia the motor system continues to generate intentions or motor commands, and thereby also predictions about the sensory consequences of bodily movements (Fotopoulou *et al.* [2008]). However, because the left-side of the body is paralysed, there will be a discrepancy between those predictions and the actual sensory feedback. In an ordinary case, this would generate a prediction error signal. Thus, one plausible hypothesis for why an individual develops anosognosia in this case is that no error signal is generated (Davies *et al.* [forthcoming]; Frith and Fristion [2013]). For this reason, the agent's sense of what she is doing is fully determined by what she predicted she would do, namely move her left arm. This prediction is wrong, but the absence of an error signal means that the mistake is not detected.

[13] In a predictive processing framework, intentions and motor commands are not distinct from the predictions of sensory consequences of bodily movement (Clark [2016], Chapter 4)

In addition to anomalous absences of error signals, another way predictive processing could be irregular is if a system anomalously generates error signals (Corlett [2018]; Feeney *et al.* [2017]; Fineberg and Corlett [2016]). Corlett and colleagues report that positive symptoms of schizophrenia, including delusional beliefs, 'correlate with aberrant [prediction error] signals in the lateral prefrontal cortex'. (Corlett *et al.* [2016], p. 1151). They further propose that aberrant prediction error signalling would cause 'a compelling sense that one's existing model of the world was wrong, that something had changed', which would in turn lead to disruptions in causal learning and, eventually, to delusional beliefs. (Corlett *et al.* [2016], p. 1146; Corlett *et al.* [2007]). According to this proposal, rather than firing as a consequence of discrepancy between predictions and the actual stimuli, neurons that carry error signals might fire randomly, or as the result of some localized neurophysiological impairment.[14] Crucially, anomalously generated error signals could arise even in cases where an agent's predictive model is accurate. In addition to misidentification delusions like Capgras, something like this might be what happens in cases of thought insertion.

Individuals who experience thought insertion report believing that they are consciously aware of thoughts that belong to another thinker. Here is a representative passage:

> I didn't hear these words as literal sounds, as though the houses were talking and I were hearing them; instead, the words just came into my head – they were ideas I was having. Yet I instinctively

[14] Note that the proposal presented by Corlett and colleagues that prediction error signals are generated irregularly is consistent with the previous proposals concerning problems with precision estimation. Indeed, in their work, Corlett and colleagues often suggest that error signals both arise 'inappropriately' and are assigned and 'anomalous degree of strength or precision' ([2016], p. 1146). Presumably, for this to be consistent with their description of the theory as a 'one-factor' theory, we are to suppose that a single impairment is responsible for both the inappropriate generation and the 'anomalous' precision. For a description of how this proposal could be used to explain the Capgras delusion, see (Corlett *et al.* [2016]; and Fineberg and Corlett [2016]).

> knew they were not my ideas. They belonged to the houses, and the houses had put them in my head. (Saks [2007], p. 27)

The belief that one's own thought belongs to a house is odd. But one reason why an individual might form such a belief is that her own episodes of conscious thinking are tagged with an anomalous prediction error signal (Parrott [2017]). When we are engaged in spontaneous thought or mind-wandering, it is very doubtful that the brain makes precise predictions about what we will think next. Nevertheless, a dysfunctional neurobiological mechanism could generate aberrant prediction error signals in a manner that codes conscious thoughts as highly surprising or unpredicted. This would give the impression that something about the conscious thought is wrong, something which the delusional belief that the thought is inserted might be adopted to explain.

We have now seen four different ways in which theorists have utilised the predictive processing framework to explain why an agent forms a delusional belief. It seems to me that a real strength of the framework is that it is flexible enough to be developed in these different ways to account for different types of delusional belief. What needs to happen now is that the details of these proposals must be developed more fully on a case by case basis and tested experimentally. Yet even though there remains work to be done, I hope to have illustrated how predictive processing offers a promising approach to understanding the formation of various delusional beliefs.

4 Explanatory Power

Now that we have seen some different ways of developing a predictive processing account of delusional belief formation, I would like to consider the explanatory power of these accounts. Several advocates of predictive processing are extremely optimistic about what the framework can explain. For instance, Clark assures us that the predictive processing framework has 'the resources required to illuminate the full spectrum of human thoughts, experiences, and actions'. ([2016], p. 203). I'm suspicious of this degree of optimism because I think there are important limitations on the explanatory power of predictive processing accounts. More specifically, I believe that accounts like those considered in the previous section can offer only partial explanations of the formation of delusional beliefs.

A causal explanation of some phenomenon can be thought of as answering a why-question. (Skow [2016], Chapter 1; Hempel [1965]). With respect to delusion formation, the general form of the question theorists aim to answer is this: why does an agent form a delusional belief? The goal of giving a causal explanation of the formation of delusional belief is to give a complete answer to this why-question. In the remainder of this section, I shall argue that contemporary predictive processing accounts do not give a complete answer to this question, only partial answers. To see this, we first need to consider the contrastive form of why-questions.

In asking why something happened, we typically have a contrast in mind. This is a point Peter Lipton illustrates in the following passage:

> We often pose our why-questions in contrastive form and it is not difficult to come up with examples where different people select different foils. When I asked my, then, 3-year old son why he threw his food on the floor, he told me that he was full. This may explain why he threw it on the floor rather than eating it, but I wanted to know why he threw it rather than leaving it on his plate. ([2004], p. 33)

Why-questions are motivated by our interests and Lipton thinks those interests impart a contrastive structure to the questions. In his words, we typically have a 'foil' in mind for the event or phenomenon that we are hoping to explain. We often make this contrastive foil explicit (e.g., why did you go to the cinema rather than stay home?) But in many contexts, we pose why-questions without any explicit mention of a contrast. However, even in those cases it is plausible that there is an implicit contrastive foil.[15] That is why Lipton could justifiably complain that his son did not really answer the question he was asked.

Lipton also notes that whenever we ask a why-question about a surprising phenomenon, the default contrast is the thing we were expecting. For instance, a doctor might ask why a healthy 6-year old has hypertension. The default way to understand this is not as asking why the child has hypertension rather than cancer, but as asking why she has hypertension rather than the healthy blood pressure we expect. In Lipton's view, having the default explanatory contrast set as the thing we expect 'focuses our inquiry on causes that will illuminate the reason our expectation went wrong'. (Lipton [2004], p. 47). Indeed, we often want to know the reason why our expectations failed because that 'directs our attention to the causes that we want to change'. ([2004], p. 47). The doctor wants to know why the child has hypertension in order to effectively treat the condition.

In many cases of delusion formation, there is a very natural belief that we strongly expect an agent to have. For example, we expect someone who is looking directly at her own mother to believe that the person she is looking at is her mother. Similarly, we expect someone who is consciously entertaining the thought that P to be believe that she is thinking that P, and we expect someone whose entire left-side is paralyzed to believe that she cannot move her left arm.

[15] Although Lipton suggests a general strategy for rending any why-question into a contrastive form, he himself wishes to remain agnostic about whether every why-question is implicitly contrastive (Lipton [2004], Chapter 3).

The fact that we have these default expectations is part of the reason that delusional beliefs seem so bizarre. It is not just that delusional agents believe something false, or unwarranted by their evidence--they believe something that is incompatible with the obviously true thing we expect them to believe.[16]

Let's call the belief we expect a delusional agent to have the 'obvious belief'. If why-questions have contrastive structure, and if Lipton is right about the default contrast, then, in asking why someone adopts a delusional belief, the implicit contrastive foil is the obvious belief. So, we can rephrase the explanatory question concerning delusion formation more explicitly as follows: why does an agent adopt a delusional belief rather than the obvious belief? I believe this is the question many cognitive neuropsychiatrists who study delusions are trying to answer. Why does someone believe that her mother is an imposter, rather than her mother? Why does someone believe that her conscious thought belongs to the houses, rather than to her? Why does someone believe that she can move her paralyzed arm, rather than that her arm is immobile?

The fact that why-questions have contrastive structure means that when we answer them there is a sense in which explain two phenomena. First, we explain what caused the surprising event. But, secondly, we also explain why the thing we expected did not occur. When these two events are incompatible, the explanations typically coincide. For example, whatever caused the 6-year old to develop hypertension also explains why she does not have healthy blood pressure. Thus, a complete explanation of why some surprising event happened is often equivalent to an explanation of why the thing we expected did not occur.

But the converse of this is not true. Even if we have a complete answer to the question of why something we expected failed to happen, we may still not understand what caused the

[16] The fact that the truth of the delusional belief is incompatible with the truth of the belief we expect the person to have does not mean that someone cannot hold both.

surprising event to occur. Suppose that I have always spent my summer holiday in Spain, but that this year I spend it in Greece. You might wonder why I went to Greece. The default contrast would be my taking a holiday in Spain; so, more explicitly, your question would be: why did I go to Greece rather than to Spain? Suppose my answer is that prices for holiday accommodation in Spain have become unaffordable. This explains why I did not go to Spain, but it does not fully answer your question because it does not explain why I went to Greece rather than to some other place, or rather than staying home. So, even on the assumption that going to Greece and going to Spain are incompatible, explaining why the expected event did not occur does not always explain why the surprising event happened. Rather, only if we presuppose that the two events are exhaustive does the former also fully explain the latter.

In any case where the expected and surprising outcomes are not exhaustive, we can shift the contrast of a why-question away from the default in order to illustrate how a putative explanation is only partial. Explicitly asking me why I went to Greece rather than to Italy highlights how my appeal to affordability does not completely answer your question. It illustrates that there must be some other cause in play, some cause that is the reason why I went to Greece rather than to Italy.

Predictive processing accounts seem like good explanations of why someone does not hold the obvious belief we expect them to have. Consider Clark's suggestion that an overestimation of precision causes substantial revisions to an agent's model of the world. If this model normally includes the obvious belief, then we can understand how a wave of extremely precise error signals could cause someone to discard the obvious belief. Similarly, consider Hohwy's proposal that expecting an extremely noisy sensory signal would cause an agent's internal model to become unconstrained by her environment. If the obvious belief is normally the result of empirical learning, then, since discounting the sensory signal would impede empirical learning, excessive expectations of noise could plausibly explain why someone lacks the

obvious belief. So, we can see, at least roughly, how these sorts of appeals to impairments in predictive processing could shed light on what causes an agent to lack an obvious belief.

But this only partially explains why an agent adopts a delusional belief. For one thing, in the cases of delusion that we have been considering, the delusional belief and the obvious belief are not exhaustive. Another option would be for the agent to withhold belief or suspend judgment. So, explaining why someone lacks the obvious belief does not explain why she believes something delusional rather than rather than nothing at all.[17] But even if we assume the agent has to believe *something,* a predictive processing account would only give us a partial explanation. As in the case of my summer holiday, this partiality can be made evident by explicitly shifting the contrast away from the default expectation. For instance, we might ask why someone believes that her mother is an alien imposter rather than that her mother has subtly altered her appearance, or rather than believing that she herself has sustained a brain injury, or rather than believing any of a number of other things about the person she is looking at. Leading predictive processing accounts do not address these sorts of contrastive why-questions. So even if a specific impairment in predictive processing could explain what causes the absence of an obvious belief, this would not tell us why a delusional belief is adopted instead.[18]

---

[17] The possibility of this contrast is often obscured in Bayesian models, which tend to think about agnosticism or suspension of judgment in terms of assigning some positive degree of subjective probability to a hypothesis, such as 0.5. This is not the place to discuss the merits of a Bayesian analysis of suspension of belief, but there are reasons to be suspicious of the analysis (Friedman [2013]; Sturgeon [2015]).

[18] In recent papers, Corlett and colleagues do attempt to address this issue by discussing the causes of the 'characteristic content' of delusional beliefs (Corlett *et al.* [2016]; Fineberg and Corlett [2016]). For instance, they consider the question of why 'delusions are highly personal…reflecting the fears and preoccupations of the individual, while at the same time, there are more generic and common aspects to them, and they draw broadly on the contents of that individual's culture and era.' (Corlett *et al.* [2016], p.

Predictive processing accounts sometimes give an impression of explanatory completeness by making certain presuppositions about an agent's prior probabilities. For instance, if we assume that an agent assigns a high subjective prior probability to the hypothesis that her mother is an alien imposter, then it will be easy enough to understand why an overly precise (possibly aberrant) error signal causes the agent to believe that her mother is an alien imposter. If this hypothesis is already something the agent thinks is pretty likely before the onset of any disturbances, then we can develop a straightforward Bayesian model to show how the agent would naturally come to adopt the imposter belief in response to an error signal (Coltheart *et al.* [2010]). Indeed, as Ryan McKay has shown, we could develop a Bayesian model of the onset of the Capgras delusion even on the assumption that a delusional agent would assign what he calls a 'vanishingly small' prior probability to the hypothesis that her mother is an alien imposter (McKay recommends 0.00027).[19] So, the assumptions a theorist would need to make about the distribution of an agent's prior probabilities need not be unreasonable or unrealistic. However, regardless of how intuitively plausible a prior probability distribution may be,

---

1149). But what they propose as a response to this question is clearly not cast in the language of predictive processing: 'their own prior knowledge and expectation will 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'. ([2016], p. 1149, cf. Fineberg and Corlett [2016]). In my view, this just restates the explanandum stated in the question and therefore does not really explain why the content of a delusional belief is adopted.

[19] If we assume that the prior probability that one's mother is an imposter is 'vanishingly small', then our formal Bayesian model will be empirically adequate only if we adjust its parameters to account for bias. McKay thinks that the Capgras delusion arises because of a bias 'toward explanatory adequacy' and he demonstrates how to adjust the parameters of a mathematical model to capture this ([2012], Appendices, pp. 350-352). Note, however, that even if McKay is right, our explanation would still be partial until we explain what causes the bias 'toward explanatory adequacy'.

predictive processing theories do not explain what causes the distribution which they rely upon in these sorts of computational models. So, even if we assume that the empirical adequacy of a model gives us good evidence for inferring an agent's priors, it remains true that whatever caused those priors is doing some of the explanatory work. The reason why an agent believes that the person she is looking at is an imposter might be partly because of a predictive processing disturbance, but it is surely also because the agent has assigned certain a degree of prior probability to the delusional hypothesis, or to the likelihood of her experience being caused by the delusional hypothesis (Parrott [2016]).

To say that a predictive processing account is only a partial explanation of why an agent adopts a delusional belief is no reason to scrap the framework. Partial explanation is better than none at all. Moreover, as we saw in the previous section, there are several promising suggestions that could causally explain why an agent fails to hold an obviously true belief. In this sense, predictive processing may help us understand part of what causes an agent to form a delusional belief. However, it is unclear whether a predictive processing account could fully answer the question of why someone forms a delusional belief. Such an account would need to say something substantive about the prior probability distributions found in cases of delusion. Perhaps this can be done within the confines of a predictive processing framework, but I suspect that a full understanding of why some individuals form delusional beliefs exceeds the limits of the framework.

5 Explanations and Implausibility

In the previous section, I argued that predictive processing theories can only partially explain the formation of a delusional belief. Although we can appeal to the predictive processing framework to sketch a plausible account of why an agent lacks an obviously true belief, the framework does not really address the question of why an agent believes something delusional instead. In this

section, I shall argue that a full answer to this question will also requires us to develop a much better understanding of how the brain generates hypotheses.

Recall the central explanationist principle underlying predictive processing, namely that a delusional belief is adopted in order to *explain* some irregular experience. One puzzle for explanationism generally is that many delusional beliefs look like exceptionally implausible explanations.[20] This is what Cordelia Fine and colleagues are getting at by describing delusional beliefs as explanatory 'nonstarters':

> [Delusions] explain the anomalous thought in a way that is so far-fetched as to strain the notion of explanation. The explanations produced by patients with delusions to account for their anomalous thoughts are not just incorrect; they are nonstarters. Appealing to the notion of explanation, therefore, does not clarify how the delusional belief comes about in the first place because the explanations of the delusional patients are nothing like explanations as we understand them. (Fine *et al.* [2005], p. 160; cf. Campbell [2001]; Davies *et al.* [2001]; Pacherie *et al.* [2006])

The description of a delusional hypothesis as an explanatory 'nonstarter' is meant to capture the sense that it is normally not even a viable candidate explanation for an observed phenomenon. As Fine and colleagues emphasize, it is unclear how 'nonstarters' come about in the first place. What could lead a system to even consider a nonstarter hypothesis as a possible explanation? How does a completely implausible hypothesis become a candidate for incorporation into an agent's predictive model of the world?

[20] The reader will notice that there are clear exceptions to this. Anosognosia for hemiplegia involves believing that one can move one's own left arm, which is obviously not outlandish. Similarly, delusions about infidelity are not extremely implausible. Thus, the problem being raised by Fine and colleagues does not arise for these cases of delusional belief.

Let's call the set of potential explanations for some experience the 'candidate set'. In the language of predictive processing, the hypotheses in an agent's candidate set are her prior probabilities—they are the hypotheses to which the agent assigns some positive degree of subjective probability. What we want to understand is not just how the brain assigns a specific distribution of probabilities to members of a candidate set, but also how the brain generates the members of that set. More specifically, as Fine and her colleagues suggest, we want to understand more clearly how a nonstarter delusional hypothesis comes to be a member of an agent's candidate set.

In response to this question, one might think that the Bayesian computational models employed by predictive processing theorists mean that every possible hypothesis is a member of an agent's candidate set. However, giving an explanation places significant demands on cognitive resources, and there are very good reasons to think our brains do not consider every possible hypothesis as a candidate explanation (Dougherty and Hunter [2003]; Navarro and Perfors [2011]; Norby [2015]; Thomas *et al.* [2008]; Weber *et al.* [1993]).[21] First, actually performing Bayesian inference on a very large set of candidate hypotheses would be computationally intractable (Rescorla [2015]). So, the brain would need to restrict the size of the candidate set in some way. Alternatively, the brain might be able to avoid the problem of intractability by not actually performing Bayesian inference. Instead, the brain might 'approximate' Bayesian inference by 'sampling' the probability distribution of a very large hypothesis space (Icard [2016];

[21] In addition to the considerations mentioned in the essay, there is some empirical evidence indicating that our brains process only a finite set of candidate hypotheses. For instance, it looks like the brain filters out contextually irrelevant alternatives before assigning subjective probabilities in certain decision-making tasks (Norby [2015]; Giguere and Love [2013]). And there is also evidence that the brain preferentially generates candidates that can easily be causally intervened upon (Buchsbaum *et al.* [2012]).

Sanborn and Chater [2016]). But even if some sort of sampling procedure would approximate Bayesian inference, there is a more basic reason why an agent's candidate set cannot consist of every possible hypothesis--the relevant hypothesis space is undefined. There simply is no well-defined set of 'all possible hypotheses', just as there is no well-defined set of 'all the contents that could possibly be believed'. In part this is because novel concepts generate novel hypotheses, but another difficulty stems from the fact that what we are able to believe seems to be partly determined by our external environment. So even if our brains could implement some type of Bayesian sampling, that would not obviate the need to understand how exactly candidate hypotheses are generated.

These sorts of considerations suggest that there is some mechanism that functions to generate the hypotheses that constitute an agent's candidate set. My claim is that, because the contents of delusional beliefs are extremely implausible, they are normally not even considered as candidate hypotheses for explaining an unusual experience. If this is right, then simply considering an implausible delusional hypothesis as a candidate explanation manifests a clear departure from ordinary cognition, which suggests that the mechanisms underlying hypothesis generation are impaired or disrupted in cases of delusion. This would explain why a delusional agent assigns some positive degree of subjective probability to an explanatory 'non-starter' hypothesis.

Someone attracted to the predictive processing framework might agree that hypothesis generation is functionally disrupted in cases of delusion, but they might speculate that the neurobiological mechanisms underlying hypothesis generation work by means of prediction error minimization.[22] If so, then we could still appeal to the predictive processing framework, and to

[22] This idea has prima facie credibility since the notion of 'prediction error' figures prominently in many contemporary approaches to unsupervised learning involving Bayesian networks. One influential idea that has emerged from this area of research is that artificial neural networks can learn structural

impairments in predictive processing, to explain how a 'non-starter' delusional hypothesis is generated. From a predictive processing standpoint, the question of how a particular 'non-starter' candidate hypothesis is generated is equivalent to the question of how a system selects a predictive model of the world that includes the 'non-starter'. Since any predictive model would be selected because it best minimizes overall prediction error, it is plausible to think that some kind of dysfunction in predictive processing could cause a system to select a highly irregular predictive model, which includes an explanatory 'non-starter'.

It should be clear that this suggestion would merely push the question about hypothesis generation back one step. There are several different ways mathematical models that aim to capture the processes involved in model selection, but all of them make simplifying assumptions about the model space over which a system learns (for example, we might assume an 'ordering' on variables (Friedman and Koller [2003]), or we might define a 'Markov Chain' over the space (Brooks *et al.* [2011]; Gilks and Roberts [1996]). For this reason, in realistic contexts where these simplifying assumptions do not hold, it is not straightforward how a model containing a 'non-starter' would come to be something that is available for selection by a system. So, even if aberrations in predictive processing could cause deviations in learning over a model space, this tells us nothing about how the space came to include a model containing a 'non-starter' delusional hypothesis. Nevertheless, the basic idea that some sort of disruption in predictive processing could cause a system to diverge from standard predictive models is plausible and I shall return to it in the following section.

In this section, I have been arguing that a complete explanation of delusional belief formation requires a better understanding of how the brain generates the candidate hypotheses

---

parameters of models by virtue of being exposed to a large enough sample of 'training data', for prediction error signals function as a kind of training signal. (Schmidhuber [2015])

over which an agent assigns some degree of subjective prior probability. Some theorists have suggested that the predictive processing framework can account for an agent's priors by virtue of the fact it conceives of them as 'empirical', or as priors that are estimated from sensory data. For example, Hohwy asserts that we can explain 'how these top-down priors are arrived at and how they are shaped over time,' by noting that they are 'guided by a particular kind of feedback signal stemming from processing of the incoming sensory signal'. ([2013], p. 34). Similarly, Friston and colleagues remark that in hierarchical Bayesian models, an agent's priors are constrained by virtue of being 'informed by empirical data'. ([2016], p. 413; cf. Friston [2005]; Friston [2010]) But even if this were plausible for ordinary cases of belief formation, it is very difficult to see how the notion of an 'empirical prior' could help us understand the origin of priors in cases of delusion. A distinguishing characteristic of delusional cognition is insensitivity to empirical evidence. It is very hard to see how, for example, a system would generate the hypothesis that one's mother is a qualitatively identical alien imposter in any way that is 'informed by empirical data', or 'guided by a particular kind of feedback'. So, appealing to 'empirical priors' does not really help us understand how an agent generates non-starter delusional candidates.

It is worth noticing that this difficulty of explaining how a delusional hypothesis enters an agent's candidate set presents an equally serious challenge for other forms of explanationism, including leading 'two-factor' theories of delusion formation (Coltheart *et al.* [2011]; Coltheart [2007]; Davies *et al.* [2001]). The central claim made by 'two-factor' theories is that delusional cognition arises because of two distinct impairments, one of which is some type of 'neuropsychological impairment in the patient that would generate some abnormal datum D, involving perceptual or affective processing, for which the patient will seek an explanation via abductive reasoning processes'. (Coltheart *et al.* [2011], p. 285; cf. Coltheart [2007]; Davies and Egan [2013]). Thus, although two-factor theorists disagree about the nature of the second impairment, orthodox two-factor accounts are explanationist in the sense that they maintain that an agent adopts a delusional belief in order to explain the occurrence of some type of irregular

experience. [23] As we have already seen, however, this style of explanationist account presupposes that some far-fetched non-starter hypothesis is a member of a delusional agent's candidate set, such that the agent might select it 'via abductive reasoning processes'. In this respect, existing two-factor theories also do not address the question of how an implausible delusional hypothesis enters an agent's candidate set.[24]

6 Hypothesis Generation

Quite a lot of work in cognitive science has been devoted to studying the processes and mechanisms involved in hypothesis selection, but we know comparatively much less about the processes and mechanisms involved in hypothesis generation. Despite this, in this section I would like to briefly present three possible ways in which hypothesis generation could be functionally anomalous in cases of delusion. As we shall see, it is plausible that that some kind of

[23]A number of different hypotheses have been suggested for the second factor. For instance, Coltheart and colleagues posit a deficit in belief-evaluation, (Garety and Freeman ([1999]) posit a data-gathering bias, and (Stone and Young [1997]) posit a bias in favour of observational adequacy over belief conservation. Ironing out the precise details of the proposed second-factor falls outside the boundaries of this essay. Interestingly, (Jaspers [1997]) is also a sort of two-factor theorist in the sense that he thinks that what he calls 'delusions proper' involve two anomalies: a 'delusional experience' and an 'alteration in the personality' (Jaspers [1997], pp. 98-105).

[24] It is plausible that implicit biases of various kinds might causally influence hypothesis generation and so some of the proposals for what constitutes a 'second factor' may figure in a full explanation of why a non-starter hypothesis enters an agent's candidate set. However, to determine whether this is the case, we need a much more developed causal model of the processes implicated in hypothesis generation.

predictive processing disturbance is causally implicated in these functional divergences. However, I shall argue that a complete explanation of why a system produces a delusional candidate hypothesis exceeds the resources of the predictive processing framework.

First, in comparison to non-delusional agents, it may be that a delusional agent over-generates candidate hypotheses. In other words, when confronted with a surprising experience, a delusional agent might generate a candidate set that contains more members than a non-delusional agent would in the very same context. This is exactly the sort of thing we should expect if hypothesis generation involved some kind of cognitive filter, for instance, something which functions to immediately rule out candidates that are incompatible with an agent's background knowledge (Parrott [2016], Perfors [2012]). If such a filter were to malfunction, then it would allow highly unusual hypotheses to enter the agent's candidate set. That would mean that a delusional hypothesis would enter an agent's candidate set if she were to think of it.

In a predictive processing framework, it would be natural to conceive of over-generation of candidates in terms of precision estimation. As we saw earlier, a system that generally expects imprecise error signals would develop a predictive model that is poorly constrained by the incoming sensory signal. If the incoming signal is the fundamental constraint on a predictive model, then a high expectation of imprecision would mean that little, if anything, functioned to constrain the model upon which the system settled. As a result, there would be little to no acquired background knowledge to exclude or filter out implausible hypotheses from the agent's candidate set.

It is worth noting that if a system over-generates hypotheses by virtue of failing to filter out or discard implausible ones, this could potentially help us understand why a delusional hypothesis is not excluded from a candidate set, but only if we presuppose that the hypothesis has been thought of. So, even if there is some reason to think that delusional subjects are

disposed to over-generate candidates, we still need to learn some more about how hypothesis generation functions in order to know how implausible hypotheses arise in the first place.[25]

A second way in which hypothesis generation may be functionally anomalous is that a delusional agent might under-generate candidates. Functionally, the basic idea is that a delusional agent would suspend the process of hypothesis generation more quickly than a non-delusional agent, which would lead to a comparatively impoverished candidate set.[26] Offhand, this might not seem problematic, but the size of a candidate set can significantly affect the assignment of subjective probability to its members (Sprenger *et al.* [2011]). For instance, the probability assigned to a specific hypothesis would be higher in a set of 10 candidates than it would be in a set of 20. The precise value of subjective probability would have clear consequences for how a system updates its predictive model of the world. For instance, if, due to under-generation, the prior probability assigned to a hypothesis were extremely high, then the agent may not revise or discard that hypothesis when presented with an error signal below a certain threshold. Thus, it may be that in non-delusional cognition implausible hypotheses, though briefly entertained, are

[25] Some recent relevant experimental can be found in (Corlett [2018]), which reports that in certain experimental contexts 'psychotic patients entertain a broader range of possible interpretations (rating multiple alternatives as excellent or good interpretations of a particular scenario), whereas healthy participants are more cautions and effectively narrow down the set of possible alternatives'. ([2018], p. 47)

[26] There is experimental data that seems relevant to the suggestion that delusional subjects under-generate candidates. Specifically, as we have already seen, there is a lot of evidence that delusional agents have a tendency to 'jump-to-conclusions', in the sense that they stop probabilistic reasoning tasks more quickly than non-delusional agents (Garety and Freeman [1999]). This indicates that they are disposed to set artificial time constraints, which could also lead them to under-generate candidate hypotheses. We also know that attentional resources are needed for hypothesis generation and that delusional subjects exhibit deficits in attention (Bell *et al.* [2006]).

immediately discarded in response to error signalling, but this fails to occur when the prior probabilities of impoverished candidate sets are comparatively inflated.

Alternatively, if an agent failed to encode many mundane hypotheses in her candidate set, then undergoing a highly irregular experience might easily cause her to generate implausible candidates. The idea would be that as soon all the members of an overly restricted candidate set are deemed inadequate and set aside, the system responsible for generating hypotheses would need to produce completely novel alternatives. In such a context, it is again plausible that a delusional hypothesis would enter an agent's candidate set if it is thought of.

Finally, there is a third way in which hypothesis generation could be functionally irregular in cases of delusion. We have seen that certain accounts postulate that delusions arise in response to a persistent and highly precise error signal. However, that might be enough to completely eliminate the agent's active hypothesis space.[27] The thought is that an extremely powerful error signal would completely fry the agent's candidate set. As a result, the hypothesis generation system would need to construct a completely novel candidate set from scratch, but the absence of priors would mean that any hypothesis the agent thinks of would become a member of the candidate set by default.

Each of these three proposals briefly illustrates a way in which the process of hypothesis generation could be functionally atypical in cases of delusion. In each case, anomalously functioning hypothesis generation would allow a 'non-starter' hypothesis to become a member of an agent's candidate set, which means it would then be subject to further computational processing. If a delusional agent's hypothesis generation system did function in one of these three ways, it is plausible that a 'nonstarter' could enter her candidate set, if she thinks of it. However, nothing I have said in this section sheds much light on why implausible hypotheses are

[27] This idea was first suggested to me by Chris Frith in conversation.

thought of in the first place.[28] It is obviously reasonable to speculate that socio-cultural factors and contextual parameters play crucial roles, but we do not have the slightest idea of how these things determine what a person thinks.

At this point I think we have reached a limit of the predictive processing framework. The theory was fabricated to illuminate the dynamics of a belief-formation system adjusting to various demands placed upon it by the incoming sensory signal. As such, it excels at illustrating how an agent adjusts her beliefs in response to a complex pattern of sensory stimuli. But not everything a person thinks is in response to sensory stimulation. Although much of the time the brain makes minor adjustments to its system of beliefs on the basis of sensory stimulation, there are times that one's belief system needs to adjust to a surprising idea, a theoretical innovation, a novel hunch, or an imaginative conclusion. In these cases, it is not clear that the predictive

[28] The three proposals that I make in this section all suggest that the functional differences in hypothesis generation involve some kind of neurocognitive impairment. This coheres with the approaches adopted by various theories in cognitive neuropsychiatry, including both predictive processing theories and two-factor theories. But, as one referee suggested, perhaps there is no general explanation of how a non-starter hypothesis enters an agent's candidate set. That is, once we give an explanation of why ordinary hypotheses are abandoned, it may be that idiosyncratic features of the agent or individual differences are causally responsible for the formation of an explanatory non-starter hypothesis. For example, it may be that, individuals with schizotypal personality traits, who are already strongly disposed to adopt strange beliefs, tend to generate hypotheses that are explanatory non-starters (Peters *et al.* [1999]). If this is right, then even if there is some neurocognitive impairment (e.g., a disruption in prediction error minimization) that causes mundane hypotheses to be discarded, there may be no additional neurocognitive impairment implicated in the formation of a non-starter delusional hypotheses (something like this is suggested in Hohwy [2015], Section 3). I think this is plausible but, unfortunately, I lack space in this essay to discuss this proposal any further.

processing framework can explain how the relevant thought arises without somehow presupposing that it has always been there. The same is true when it comes to delusional beliefs.

7 Conclusion

Within cognitive neuropsychiatry, explanations of the formation of delusional beliefs have been heavily influenced by the thought that an agent adopts a delusional belief in order to explain a highly irregular experience. Contemporary predictive processing theories fall within this tradition. I have argued that a virtue of the predictive processing framework, especially given the wide variety of delusional beliefs, is that it allows us to develop different theoretical explanations to account for the onset of different kinds of delusions, rather than relying on single causal variable in every single case. One of the aims of this essay has been to illustrate some of the ways in which theorists might go about filling in the general idea, recommended by the predictive processing framework, that delusional beliefs are the result of a disturbance in predictive processing.

Nevertheless, I have also argued that there are two important limitations to the predictive processing framework. First, the framework only partially explains why an agent adopts a delusional belief. Although predictive processing theories can offer a plausible account of why an agent fails to believe the obvious thing we expect her to believe, they shed little light on why a delusional belief is adopted instead, rather than something else, or rather than nothing at all. This explanatory gap is often obscured when an account presupposes, without explaining, a specific prior probability distribution that includes a delusional hypothesis.

That same presupposition also conceals the second limitation of the predictive processing framework. If, as I have claimed, implausible beliefs are not ordinarily candidate explanations for surprising phenomena, then it is not clear how delusional beliefs become

candidate explanations. Indeed, one of the things that has always been difficult to understand about delusions is why different agents come to take the same strange ideas so seriously, first as potential beliefs and then, eventually, as settled convictions. We therefore need a much better understanding of what mechanisms are responsible for a delusional hypothesis becoming a candidate explanation, and this requires a much more developed picture of hypothesis generation.

Much of what we think about the world is shaped by the experiences we have, but it is also significantly shaped by what we think is possible. The possibilities that we can envision partially determine how we explain surprising events and experiences. So, anything that alters how we think about what is possible, will have consequences on what we eventually come to believe. It therefore seems to me that our understanding of why certain individuals form delusional beliefs would be greatly advanced by an explanation of why individuals generate delusional candidate hypotheses. Although I have said that it is difficult to imagine how a predictive processing theorist could address this issue, we have seen that it is equally a challenge for alternative versions of explanationism, such as the two-factor framework. Thus, regardless of which theoretical approach we adopt, a fuller understanding of hypothesis generation is an important step toward developing a complete explanation of why certain individuals form delusional beliefs.

Acknowledgments

Some of the material in this essay was presented in earlier forms at a conference on predictive processing, hosted by the Cambridge Centre for Research in the Arts, Social Sciences and Humanities, at a workshop on delusions, hosted by Monash University, at the University of

Cambridge Philosophy of Science Seminar, and at the Mind Work in Progress Seminar at the University of Oxford. I am very grateful to members of the audience for their questions and comments on all of these occasions. I would also like to thank Tim Bayne, Martin Davies, Jakob Hohwy, Nicholas Shea, and Dan Williams for many helpful suggestions during our conversations on the topic of this essay. Finally, I am very grateful to the two anonymous referees who provided detailed and extremely thoughtful comments on previous drafts of this essay.

Matthew Parrott
Department of Philosophy
University of Birmingham
Edgbaston, Birmingham, UK
m.parrott@bham.ac.uk

References

Adams, R.A., Stephan, K., Brown, H., Frith, C. and Friston, K. [2013]: 'The Computational Anatomy of Psychosis', *Frontiers in Psychiatry*, **4**, pp. 1-26.

Bayne, T. and Pacherie, E. [2004]: 'Bottom-up or Top-Down: Campbell's Rationalist Account of Monothematic Delusions', *Philosophy, Psychiatry, & Psychology*, **11**, pp. 1-11.

Bell, V., Halligan, P., and Ellis, H. [2006]: 'Explaining Delusions: A Cognitive Perspective', *Trends in Cognitive Sciences*, **10**, pp.219-226.

Bentall, R. [2018]: 'Delusions and Other Beliefs' in L. Bortolotti (*ed*), *Delusions in Context,* London: Palgrave Macmillan.

Berti, E., Ladavas, A., Stracciari C., Giannarelli A., and Ossola, A. [1998]: 'Anosognosia for Motor Impairment and Dissociations with Patients' Evaluation of the Disorder: Theoretical Considerations', *Cognitive Neuropsychiatry*, **3**, pp.21-43.

Blakemore, S., Wolpert, D., and Frith, C. [2002]: 'Abnormalities in the Awareness of Action', *Trends in Cognitive Sciences*, **6**, pp. 237-242.

Bobes, M., Góngora, D., Valdes, A., Santos, Y., Acosta, Y., Garcia, Y., Lage, A., and Valdés-Sosa, M. [2016]: 'Testing the Connections within Face Processing Circuitry in Capgras Delusion with Diffusion Imaging Tractography', *NeuroImage: Clinical*, **11**, pp.30-40.

Bortolotti, L. [2009]: *Delusions and Other Irrational Beliefs*, Oxford: Oxford University Press.

Brighetti, G., Bonifacci, P., Borlimi, R. and Ottaviani, C. [2007]: "'Far From the Heart Far From the Eye": Evidence From the Capgras Delusion', *Cognitive Neuropsychiatry*, **12**, pp.189-197.

Brooks, S., Gelman, A., Jones, G. and Meng, X. (*eds*) [2011]: *Handbook of Markov Chain Monte Carlo*. London: Chapman & Hall, CRC Press.

Buchsbaum, D., Bridgers, S., Weisberg, D., and Gopnik, A. [2012]: 'The Power of Possibility: Causal Learning, Counterfactual Reasoning, and Pretend Play', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, pp.2202-2212.

Campbell, J. [2001]: 'Rationality, Meaning, and the Analysis of Delusion', *Philosophy, Psychiatry, & Psychology*, **8**, pp.89-100.

Clark, A. [2016]: *Surfing Uncertainty*. Oxford: Oxford University Press.

Clark, A. [2013]: 'Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science', *Behaviour and Brain Sciences*, **36**, pp. 181-204.

Colombo, M. and Series, P. [2012]: 'Bayes in the Brain – On Bayesian Modelling in Neuroscience', *The British Journal for the Philosophy of Science*, **63**, pp. 697-723.

Coltheart, M., Cox, R., Sowman, P., Morgan, H., Barnier, A., Langdon, R., Connaughton, E., Teichmann, L., Williams, N., and Polito, V. [2018]: 'Belief, Delusion, Hypnosis, and the Right Dorsolateral Prefrontal Cortex: A transcranial magnetic stimulation study'. *Cortex*, **101**, pp. 234-48.

Coltheart, M. [2007]: 'Cognivie Neuropsychiatry and Delusional Belief', *Quarterly Journal of Experimental Psychology*, **60**, pp. 1041-62.

Coltheart, M., Langdon, R., & McKay, R. [2011]: 'Delusional Belief', *Annual Review of Psychology*, **62**, pp. 271–298.

Coltheart, M., Menzies, P., & Sutton, J. [2010]: 'Abductive Inference and Delusional Belief', *Cognitive Neuropsychiatry*, **15**, pp. 261–287.

Corlett, P. [2018]: 'Delusions and Prediction Error', in L. Bortolotti (*ed*), *Delusions in Context*, London: Palgrave Macmillan.

Corlett, P., Honey, G. and Fletcher, P. [2016]: 'Prediction Error, Ketamine and Psychosis: An Updated Model', *Journal of Psychopharmacology*, **30**, pp.1145-1155.

Corlett, P. and Fletcher, P. [2015]: 'Delusions and Prediction Error: Clarifying the Roles of Behavioral and Brain Response', *Cognitive Neuropsychiatry*, **20**, pp. 95-105.

Corlett, P., Taylor, J., Wang, X., Fletcher, P. and Krystal, J. [2010]: 'Toward a Neurobiology of Delusions', *Progress in Neurobiology*, **92**, pp. 345-369.

Corlett, P., Frith, C. and Fletcher. P. [2009]: 'From Drugs to Deprivation: A Bayesian Framework for Understanding Models of Psychosis', *Psychopharmacology*, **206**, pp. 515-530.

Corlett, P., Honey, G. and Fletcher, P. [2007]: 'From Prediction Error to Psychosis: Ketamine as a Pharmacological Model of Delusions', *Journal of Psychopharmacology*, **21**, pp. 238–52.

Corlett, P., Honey, G., Aitken, M., Dickinson, A., Shanks, D., Absalom, A., Lee, M., Pomarol-Clotet, E., Murray, G., McKenna, P., and Robbins, T. [2006]: 'Frontal Responses During Learning Predict Vulnerability to the Psychotogenic Effects of Ketamine: Linking Cognition, Brain Activity, and Psychosis', *Archives of General Psychiatry*, **63**, pp.611-621.

Davies, D., Teufel, C. and Fletcher, P. [2018]: 'Anomalous Perceptions and Beliefs are Associated with Shifts Toward Different Types of Prior Knowledge in Perceptual Inference', *Schizophrenia Bulletin*, **44**: pp. 1245-1253.

Davies, M., Davies, A. and Coltheart, M. [2005]: 'Anosognosia and the Two-factor Theory of Delusions', *Mind & Language*, **20**, pp.209-236.

Davies, M., Coltheart, M., Langdon, R. and Breen, N. [2001]: 'Monothematic Delusions: Towards a Two-Factor Account', *Philosophy, Psychiatry, & Psychology*, **8**, pp.133-158.

Davies, M. and Egan. A [2013]: 'Delusion: Cognitive Approaches, Bayesian Inference, and Compartmentalization', in K. W. M. Fulford, M. Davies, R. Gipps, G. Graham, J. Sadler, G Stanghellini and T. Thornton (*eds*), *The Oxford Handbook of Philosophy of Psychiatry*, Oxford: Oxford University Press.

Davies, M., McGill, C. and Aimola Davies, A. [forthcoming]: 'Anosognosia for Motor Impairments as a Delusion: Anomalies of Experience and Belief Evaluation', In A. Mishara, P. Corlett, P. Fletcher, A. Kranjec and M. Schwartz (*eds*), *Phenomenological Neuropsychiatry: How Patient Experience Bridges Clinic with Clinical Neuroscience,* New York: Springer.

Dougherty, M. and Hunter, J. [2003]: 'Hypothesis Generation, Probability Judgment, and Individual Differences in Working Memory Capacity', *Acta Psychologica*, **113**, pp.263-282.

Ellis, H., Lewis, M., Moselhy, H. and Young, A. [2000]: 'Automatic Without Autonomic Responses to Familiar Faces: Differential Components of Covert Face Recognition in a Case of Capgras Delusion', *Cognitive Neuropsychiatry*, **5**, pp.255-269.

Emmons, S., Geiser, C., Kaplan, K. and Harrow, M. [1997]: *Living with Schizophrenia*, Washington, DC: Accelerated Development.

Feeney, E., Groman, S., Taylor, J. and Corlett, P. [2017]: 'Explaining Delusions: Reducing Uncertainty Through Basic and Computational Neuroscience', *Schizophrenia Bulletin*, **43**, pp.263-272.

Fletcher, P. and Frith, C. [2009]: 'Perceiving is Believing: a Bayesian Approach to Explaining the Positive Symptoms of Schizophrenia', *Nature Reviews Neuroscience*, **10**, pp. 48-58.

Fine, C., Craigie, J. and Gold, I. [2005]: 'Damned if you do; Damned if you don't: The Impasse in Cognitive Accounts of the Capgras Delusion', *Philosophy, Psychiatry, & Psychology*, **12**, pp. 143-151.

Finberg, S. and Corlett, P. [2016]: 'The Doxastic Shear Pin: Delusions as Errors of Learning and Memory', *Cognitive Neuropsychiatry*, **21**, pp. 73-89.

Fotopoulou, A., Tsakiris, M., Haggard, P., Vagopoulou, A., Rudd, A. and Kopelman, M. [2008]: 'The Role of Motor Intention in Motor Awareness: An Experimental Study on Anosognosia for Hemiplegia', *Brain*, **131**, pp.3432-3442.

Friedman, J. [2013]: 'Rational Agnosticism and Degrees of Belief', *Oxford Studies in Epistemology*, **4**, pp. 57-82.

Friedman, N. and Koller, D. [2003]: 'Being Bayesian about Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks'*, Machine Learning*, **50**, pp. 95-125.

Friston, K., Litvak, V., Oswal, A., Razi, A., Stephan, K., van Wijk, B., Ziegler, G. and Zeidman, P. [2016]: 'Bayesian Model Reduction and Empirical Bayes for Group (DCM) Studies', *Neuroimage*, **128**, pp.413-431.

Friston, K. [2010]: 'The Free-Energy Principle: A Unified Brain Theory?', *Nature Reviews Neuroscience*, **11**, pp. 127-138.

Friston, K. [2005]: 'A Theory of Cortical Responses', *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **360**, pp. 815-836.

Frith, C. and Friston, K. [2013]: 'False Perceptions and False Beliefs: Understanding Schizophrenia', *Neurosciences and the Human Person: New Perspectives on Human Activities*, **121**, pp. 1-15.

Frith, C. [2012]: 'Explaining Delusions of Control: The Comparator Model 20 Years On', *Consciousness and Cognition*, **21**, pp.52-54.

Garety, P., Freeman, D., Jolley, S., Dunn, G., Bebbington, P., Fowler, D. and Dudley, R. [2005]: 'Reasoning, Emotions, and Delusional Conviction in Psychosis', *Journal of Abnormal Psychology,* **114**, pp. 373–384.

Giguère, G. and Love, B. [2013]: 'Limits in Decision Making Arise From Limits in Memory Retrieval', *Proceedings of the National Academy of Sciences*, **110**, pp.7613-7618.

Gilks, W. and Roberts, G. [1996]: 'Strategies for Improving MCMC' in W. Gilks, S. Richardson, and D. Spiegelhalter (*eds*), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall, CRC Press.

Hempel, C. [1965]: *Aspects of Scientific Explanation*, New York: Free Press.

Hirstein, W. and Ramachandran, V. [1997]: 'Capgras Syndrome: a Novel Probe for Understanding the Neural Representation of the Identity and Familiarity of Persons', *Proceedings of the Royal Society of London B: Biological Sciences*, **264**, pp.437-444.

Hohwy, J. [2016]: 'The Self-Evidencing Brain', *Nous*, **50**, pp. 259-285.

Hohwy, J. [2015]: 'Prediction Error Minimization, Mental and Developmental Disorder, and Statistical Theories of Consciousness', in R. Gennaro (ed), *Disturbed Consciousness: New Essays on Psychopathology and Theories of Consciousness*, Cambridge: MIT Press, pp. 293-324.

Hohwy, J. [2013]: *The Predictive Mind*, Oxford: Oxford University Press.

Howes, O. and Kapur, S. [2009]: 'The Dopamine Hypothesis of Schizophrenia: Version III – The Final Common Pathway', *Schizophrenia Bulletin*, **35**, pp.549-562.

Icard, T. [2016]: 'Subjective Probability as Sampling Propensity', *Review of Philosophy and Psychology*, **7**, pp.863-903.

Jaspers, K. [1997]: *General Psychopathology*, J. Hoening and M. Hamilton (*trans*), Baltimore: John Hopkins University Press.

Joyce, J. [2003]: 'Bayes' Theorem', in E. N. Zalta (ed), *Stanford Encyclopedia of Philosophy*, available at <plato.stanford.edu/entries/bayes-theorem/>.

Kahneman, D. [2011]: *Thinking, Fast and Slow*, New York: Farrar, Straus, and Giroux.

Kahneman , D. [2003]: 'A Perspective on Judgment and Choice: Mapping Bounded Rationality', *American Psychologist*, 58, pp. 697-720.

Knill, D. and Pouget, A. [2004]: 'The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation', *Trends in Neuroscience*, **27**, pp 712-719.

Langdon, R. and Bayne T. [2010]: 'Delusion and Confabulation: Mistakes of Perceiving, Remembering and Believing', *Cognitive Neuropsychiatry*, **15**, pp. 319-345.

Lipton. P. [2004]: *Inference to the Best Explanation*, London: Routledge, 2nd edition.

Maher, B. [1974]: 'Delusional Thinking and Perceptual Disorder', *Journal of Individual Psychology*, **30**, pp. 98-113.

McKay, R. [2012]: 'Delusional Inference', *Mind and Language,* **27**, pp. 330-55.

Murray, G., Corlett, P., Clark, L., Pessiglione, M., Blackwell, A., Honey, G., Jones, P., Bullmore, E., Robbins, T. and Fletcher, P. [2008]: 'How Dopamine Dysregulation Leads to Psychotic Symptoms? Abnormal Mesolimbic and Mesostriatal Prediction Error Signalling in Psychosis', *Molecular Psychiatry*, **13**, p.239.

Navarro, D. and Perfors, A. [2011]: 'Hypothesis Generation, Sparse Categories, and the Positive Test Strategy', *Psychological Review*, **118**, pp.120-134.

Norby, A. [2015]: 'Uncertainty Without All the Doubt', *Mind & Language*, **30**, pp.70-94.

Pacherie, E., Green, M. and Bayne, T. [2006]: 'Phenomenology and Delusions: Who Put the 'Alien' in Alien Control?', *Consciousness and Cognition*, **15**, pp.566-577.

Parrott, M. [2017]: 'Subjective Misidentification and Thought Insertion', *Mind and Language*, **32**, pp. 39-64.

Parrott, M. [2016]: 'Bayesian Models, Delusional Beliefs, and Epistemic Possibilities', *The British Journal for the Philosophy of Science*, **67**, pp. 271-296.

Perfors, A. [2012]: 'Bayesian Models of Cognition: What's Built in After All?' *Philosophy Compass*, **7**, pp. 127-138.

Peters, E., Joseph, S. and Garety, P. [1999]: 'Measurement of Delusional Ideation in the Normal Population: Introducing the PDI', *Schizophrenia Bulletin*, **25**, pp. 553-576.

Rao, R. and Ballard, D. [1999]: 'Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects', *Nature Reviews Neuroscience*, **2**, pp. 79-87.

Rescorla, M. [2015]: 'Bayesian Perceptual Psychology', in M. Matthen (ed), The *Oxford Handbook of Philosophy of Perception*, Oxford: Oxford University Press.

Saks, E. [2007]: *The Centre Cannot Hold*, New York: Hyperion.

Skow, B. [2016]: *Reasons Why*, Oxford: Oxford University Press.

Stone, T. and Young, A. [1997]: 'Delusions and Brain Injury: The Philosophy and Psychology of Belief', *Mind & Language*, **12**, pp. 327–364.

Sanborn, A. and Chater, N. [2016]: 'Bayesian Brains without Probabilities', *Trends in Cognitive Sciences*, **20**, pp. 883-893

Schmidhuber, J. [2015]: 'Deep Learning in Neural Networks: An Overview', *Neural Networks*, **61**, pp. 85-117.

So, S., Freeman, D., Dunn, G., Kapur, S., Kuipers, E., Bebbington, P. and Garety, P. [2012]: 'Jumping to Conclusions, a Lack of Belief Flexibility and Delusional Conviction in Psychosis: A Longitudinal Investigation of the Structure, Frequency, and Relatedness of Reasoning Biases', *Journal of Abnormal Psychology*, **121**, pp. 129–130.

Sprenger, A., Dougherty, M., Atkins, S., Franco-Watkins, A., Thomas, R., Lange, N. and Abbs, B. [2011]: 'Implications of Cognitive Load for Hypothesis Generation and Probability Judgment', *Frontiers in Psychology*, **2**, pp. 1-15.

Sterzer, P, Adams, R., Fletcher, P., Frith, C., Lawrie, S., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M. and Corlett, P. [2018]: 'The Predictive Coding Account of Psychosis', *Biological Psychiatry*, **84**, pp. 634-643.

Sturgeon, S. [2015]: 'The Tale of Bella and Creda', *Philosophers' Imprint*, **15**, pp. 1-9.

Thomas, R., Dougherty, M., Sprenger, A. and Harbison, J. [2008]: 'Diagnostic Hypothesis Generation and Human Judgment', *Psychological Review*, **115**, pp.155-185.

Tranel, D., Damasio, H. and Damasio, A. [1995]: 'Double Dissociation Between Overt and Covert Face Recognition', *Journal of Cognitive Neuroscience*, **7**, pp.425-432.

Weber, E., Böckenholt, U., Hilton, D. and Wallace, B. [1993]: 'Determinants of Diagnostic Hypothesis Generation: Effects of Information, Base Rates, and Experience', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **19**, pp.1151-1164.

Williams, D. [2018]: 'Hierarchical Bayesian Models of Delusion', *Consciousness and Cognition,* **61**, pp. 129-147.