DELUSIONAL PREDICTIONS AND EXPLANATIONS

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 a predictive processing account could plausibly be developed 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 generates a delusional hypothesis is a crucial step to understanding why she eventually comes to believe it. The final section of the essay outlines three ways in which the process of hypothesis generation might be functionally divergent in cases of delusional cognition.

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 making predictions about the external world, and then generating a kind of error signal anytime those predictions are violated. This signal then functions to adjust the brain's predictions. 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 current beliefs about

[1] In the literature, this is called both 'predictive processing' and 'predictive coding'. So, let me say that in this essay I shall use 'predictive processing' to refer to any theoretical approach or model that maintains that neural computations are carried out by means of a system which updates a predictive model 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, Ch. 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 any system minimizes prediction error just in case it can be accurately described in terms of a specific 'predictive coding' strategy).

the external world just are our best predictions about the distal causes of our current sensations, and we adopt new beliefs in order to explain unexpected or unpredicted aspects of the changing sensory signal.

By extension, according to the predictive processing framework, delusional beliefs are adopted in order to explain the occurrence of some highly irregular or unexpected aspect of the incoming sensory signal (cf. Frith and Friston, 2013).[2] This fits nicely with experimental data suggesting that some kind of irregular experience is implicated in the onset of several delusions (Coltheart, et. al., 2011; Langdon and Bayne, 2010). It also illustrates how predictive processing is a modern version of the following influential idea in 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)

[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 one forms in cases of delusion (e.g., Bayne and Pacherie 2004; Bortolotti 2009).

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 beliefs are adopted in order to explain non-conscious sensory experiences.[3] Second, Maher claims that the cognitive processes in delusional cognition are functioning normally. However, many contemporary theorists disagree with him in this respect because experimental evidence indicates that the occurrence of an irregular experience is not sufficient to produce a delusional belief. As Paul Fletcher and Chris Frith note, 'if a perceptual anomaly was sufficient to generate symptoms such as delusions of control, then the creation of such anomalies in otherwise healthy people should generate false beliefs. Several experiments show that this is not the case.' (2009, pg. 51). Thus, predictive processing theorists maintain that, in cases of delusion, there is some kind of disturbance or disruption to the cognitive or neural processing which underwrites belief formation.

The aim of this essay is to assess the prospects of a predictive processing explanation for why certain individuals form delusional beliefs. In the following section, I shall briefly present the basic explanation-sketch of delusion formation that is recommended by predictive processing theorists. I shall also present a portion of the corroborating evidence that supports adopting this theoretical framework. In section 2, I shall discuss several different ways in which one could fill in the basic explanation-sketch 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 slightly different ways. Nevertheless, the

[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)

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 a predictive processing account 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. Specifically, I shall argue that the exceptional implausibility of delusional beliefs makes them poor candidate explanations for even 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 hypotheses. As we will see, however, it is not clear that the predictive processing framework can illuminate this important aspect of human cognition. In section 5, I shall offer three proposals for how hypothesis generation might be functionally divergent in cases of delusion. One consequence of this discussion is that, depending on how hypothesis generation actually functions, it may be that multiple neurocognitive impairments are implicated in the formation of a delusional belief.

1. Predictive Processing: The Basics

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; cf. 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 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 predictions 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 (cf. 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 Karl 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; cf. Clark 2013) In the predictive processing framework, all beliefs are adopted because they best explain an agent's current sensory experience.

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 some surprising evidence, the prior probabilities assigned to a set of hypotheses are adjusted in accordance with Bayes's theorem (cf. Joyce, 2003). This is essentially a rule for determining how one's 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.[4]

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, Ch. 1; Hohwy, 2016; Hohwy, 2015; Knill and Pouget, 2004). For example, Jakob 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)[5]

[4] Bayesian modelling raises many questions about whether the brain actually does implement Bayesian inference (cf. Williams, forthcoming). One well-known difficulty is that there is substantial evidence that, from a Bayesian perspective, human reasoning and decision-making is suboptimal (e.g., 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 modelled in Bayesian terms by tweaking various parameters of a mathematical model. But this raises further questions about overfitting one's model to 'fit' the data (cf. Colombo and Series, 2012). Unfortunately, these issues cannot be explored further within the confines of this essay.

[5] 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, Ch. 7.

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. Sometimes error signals are going to be unreliable because they are generated in contexts where there is a lot of uncertainty or noise. For example, visual information is more reliable in daylight than in darkness, so error signals generated by visual stimuli in a dark room should be discounted. In order to not be misled by unreliable error signals, the brain learns to predict and estimate the precision of those signals (Hohwy, 2013). In an optimal predictive processing system, precise prediction errors are weighted accurately, and this allows the brain to both learn from reliable signals and 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 many predictive processing accounts of delusion formation.

2. Forming Delusional Beliefs

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 mechanism embodied by the brain seems to generate fairly coherent beliefs about the world.

According to predictive processing theorists, delusional beliefs arise because of some kind of disturbance in brain's predictive processing system. 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) Several strands of empirical

research support the suggestion 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; 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 (Clark, 2016, Ch. 1, Friston 2005). Relatedly, the post-synaptic gain on neuronal populations 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 appear to be abnormal in schizophrenia. Specifically, 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 draw inferences on the basis of less information (Garety, et al. 2005; Garety and Freeman, 1999; So, et al. 2012). Bayesian models of this phenomenon illustrate how a

disturbance in predictive processing could plausibly predict these sorts of abnormalities in probabilistic reasoning.[6]

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

By far, 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)

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'

[6] There is more empirical work that suggests a predictive processing disturbance is implicated in delusional belief formation, which I lack space to mention in this essay. There are good overviews in Adams, et al. 2013; Fletcher and Frith, 2009; and Clark, 2016, Ch. 7.

According to the predictive processing framework, error signals that are unreliable or imprecise (i.e., those with a high variance) are normally attenuated. As a result, the brain relies more on prior predictions about what the world is like – in other words the value assigned to one's prior probabilities is effectively inflated. So, one way that precision could be 'misallocated' is if a delusional agent generally has a much greater expectation than normal of being in a noisy environment. This would mean that she would rely heavily on her internal predictive model of the world. From a more dynamic perspective, it means that the agent's model of the world would not be constrained by the incoming sensory signal, because the agent effectively devalues any mismatch between her predictions and the actual sensory signal. At higher levels, this would plausibly result in beliefs that are not shaped by one's experiences. As Hohwy claims, '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)

Discounting the value of error signals, or, equivalently, overvaluing confidence in one's prior predictions, could have additional knock-on effects on the functioning of a predictive processing system. First, low-valued error signals are typically generated in cases where there is a 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 noisy environment, low-valued error signals would tend to be treated as background noise, thereby inhibiting 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 further suggests that such a signal may attract more attention as something that must be explained away. Thus, there are several ways belief formation may become distorted by an expectation that error signals are imprecise.

Someone with an irregularly high expectation of imprecision will continuously ignore feedback from the external world. As a result, her model of the world would gradually become less and less constrained by the way the world is. Hohwy suggests that this eventually leads to an 'idiosyncratic interpretation' of the world. This sort of developmental trajectory may help us explain how an agent gradually comes to adopt the sort of elaborate system of delusional beliefs found in many cases of psychosis. 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 a person developing more elaborate delusional beliefs over time, eventually resulting in a bizarre and 'idiosyncratic' world view. In a predictive processing framework, the development of this sort of delusional belief system is plausibly the result of continuous devaluing of sensory-based prediction error.

Undervaluing precision is not the only way it could be 'misallocated'. Another possibility is that precision is overestimated. Andy 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). 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 high estimates of precision indicate that the brain's model is rather significantly mistaken and needs to be radically revised.[7] In this sense, Clark's suggestion is that a wave of highly precise error signals could cause the brain to make significant revisions by adopting odd beliefs about the world. As he puts it, a predictive processing system will 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 a delusional belief, let's consider Capgras delusion. 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 the Capgras delusion is associated with a deficit in an individual's autonomic nervous system, specifically visual presentations of familiar faced 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 response would generate a prediction error signal. But what happens if that signal is persistent and also estimated to be highly precise? Clark's suggestion is that this would cause significant revisions to an agent's model of the world, including a revision of the belief that this person (who looks 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,

[7] 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, 'as a result, prediction errors will be propagated even further up the system to ever-higher levels of abstraction.' (2009, p. 55)

but do not form delusional beliefs (Tranel, et. al., 1995). The proposal that Capgras subjects overestimate precision, allows us to explain this contrast.[8]

The two accounts we have just considered claim that delusional beliefs are caused by problems with precision estimation. Hohwy hails the idea as a 'generic account' of delusion formation. (2013, p. 159). I think 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 'generic account' of delusion formation. Rather, we should expect that the complex dynamics of a predictive processing system can be impaired or disrupted in different ways. Before concluding this section, let me briefly outline two further ways in which delusional beliefs could be caused by different kinds of predictive processing impairments.

Some delusions might arise not because of any problem with precision estimation but because the system completely fails to generate an error signal when the brain's predictions fail to match the sensory signal. This may 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-side motor impairment following a stroke that damages the right-hemisphere of the brain. (Davies, et al. 2005, Fotopolou et al., 2008). An individual 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

[8] A Bayesian model of this sort of belief 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.

proprioceptive sensory consequences of that movement (Blakemore, et al., 2002; Clark 2016, Ch. 4).[9] 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 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 sort of discrepancy would generate a prediction error signal. Thus, one plausible hypothesis for why an individual might develop anosognosia is that no error signal is generated. 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 (Davies, et al. forthcoming, Frith and Friston, 2013). This prediction is wrong but the absence of an error signal means that the mistake is not detected.

In addition to anomalous absences of error signals, another way predictive processing could be impaired is if a system generates inappropriate error signals. For example, rather than firing as a consequence of a discrepancy between prediction and the actual stimulus, populations of neurons that carry error signals might fire randomly, or as the result of some localized neurophysiological impairment. In a predictive processing system, error signals indicate that something is wrong with the agent's predictive model, but inappropriate error signals could be

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

generated even in cases where that model is accurate. Something like this might be what happens in thought insertion (cf. Parrott, 2017).

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 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 by an anomalous prediction error signal. When we are engaged in spontaneous thought or mind-wandering, it is very doubtful that the brain makes precise predictions about what we will consciously think next.[10] Nevertheless, a dysfunctional neural 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 several different ways in which one could appeal to 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 different ways to account for different types of delusional belief. What needs to happen now is that the details of these proposals must be filled in on a case by case basis and tested experimentally. Yet even

[10] For this reason, Frith (2012) has expressed scepticism about being able to give a predictive processing account of thought insertion.

though there is more work to be done, I hope to have illustrated how predictive processing offers a promising approach to understanding the formation of various delusional beliefs.

3. 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 this sort of account. 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 the predictive processing framework. More specifically, I shall argue that a predictive processing account can offer only a partial explanation of the formation of delusional beliefs.

A causal explanation of some phenomenon can be thought of as answering a why-question. (Skow, Ch. 1; cf. Hempel, 1965). With respect to delusion formation, the general form of the relevant explanatory question 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 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.' (Lipton, 2004, pg. 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.[11] 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 question 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 expected. 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, pg. 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.

[11] 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 (2004, Ch. 3).

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. 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 incompatible with the obviously true thing we expect them to believe.[12]

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

[12] 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.

events are incompatible, the two 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 fail to understand what caused the 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 is my taking a holiday in Spain; so, more explicitly, your question is: why did I go to Greece rather than to Spain? Suppose my answer is that prices for accommodation in Spain have become too high for me to afford. 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 explain why the surprising event happened. Rather, unless we presuppose that the two events are exhaustive, the former only partially explains the latter.

In cases 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. Asking me explicitly why I went to Greece rather than to Italy highlights how my appeal to prices does not completely answer your question. It illustrates that there must be some other cause in play 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. For example, in a predictive processing system, if an overestimation of precision causes substantial revisions to an agent's model of the world,

and if this model normally includes the obvious belief, then it is clear how this kind of disruption to precision estimation would cause someone to discard that belief. Similarly, expecting a noisy sensory signal means that an agent's model would become unconstrained by her environment. If an obvious belief is normally the result of empirical learning, then discounting that signal would impede learning, thereby explaining why the agent lacks the obvious belief. So, we can see, at least roughly, how appealing to a disturbance in predictive processing can 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 all cases of delusion 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 an obvious belief does not explain why she believes something delusional rather than rather than nothing at all.[13] But even if we assume the agent has to believe *something,* a predictive processing account looks to be only 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. Predictive processing accounts have little to say about these sorts of contrastive why-questions. So, the idea that a predictive processing system

[13] The possibility of this contrast is often obscured by 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 suspicious of the analysis (cf. Friedman, 2013; Sturgeon, 2015).

can be impaired may present a plausible picture of what causes the absence of an obvious belief, but this does not tell us why a delusional belief is adopted instead.

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 probability to the hypothesis that her mother is an alien imposter, then it will be much easier to understand why an overly precise (possibly aberrant) error signal could cause the agent to believe that her mother is an alien. If this hypothesis is already something the agent thinks is fairly likely independently of any disturbances, then we can develop a straightforward Bayesian model to show how the agent would come to adopt the imposter belief in response to an error signal (cf. Coltheart et al. 2010; McKay 2012). But, predictive processing theories do not explain what causes the distribution of prior probabilities which they rely upon in this sort of computational model. So, even if we assume that the empirical adequacy of a model gives us good evidence for inferring an agent's prior probability distribution, it remains true that whatever caused that distribution 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 would surely also be 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 (cf. 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. As we saw in the previous section, there are several promising schemas that could causally explain why an agent does not 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 to me 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 irregular 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.

4. 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 requires us to develop a much better understanding of how the brain generates hypotheses. However, it is unclear whether this can be done within the predictive processing framework.

Recall that the central explanationist doctrine underlying predictive processing is that a delusional belief is adopted in order to *explain* some irregular experience. One puzzle for explanationism is that *prima facie* delusional beliefs look like extremely implausible explanations.[14] 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

[14] The reader will notice that anosognosia for hemiplegia is an exception--believing that one can move one's own left arm is not bizarre.

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, pg. 160; cf. Campbell 2001; Davies, et al. 2001; Pacherie et al. 2006)

The description of a delusional hypothesis as a '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 explanatory 'nonstarters' come about in the first place. What could lead a system to even consider a nonstarter hypothesis? How does a completely implausible hypothesis become part of an agent's predictive model of the world?

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 hypotheses to which the agent assigns some degree of subjective probability. So, 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 how a nonstarter delusional hypothesis comes to be a member of the candidate set.

One might think that Bayesian computational models suggest 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 reasons to think our brains do not consider every possible hypothesis as a potential explanation (cf. Dougherty, and Hunter, 2003; Navarro and Perfors, 2011; Norby, 2015; Thomas, et. al. 2008; Weber, et. al. 1993).[15] First,

[15] 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

actually performing Bayesian inference on a very large set of candidate hypotheses would be computationally intractable (cf. Rescorla, 2015). So, it seems that the brain must restrict the size of the candidate set in some way. Alternatively, the brain might be able to avoid the problem of computational intractability by not actually performing Bayesian inference. Instead, the brain might 'approximate' Bayesian inference by 'sampling' the probability distribution of a large hypothesis space (Icard, 2016; Sanborn and Chater, 2016). Yet even if some kind of sampling procedure would approximate Bayesian inference, a more basic reason why an agent's candidate set cannot consist of every possible hypothesis is that 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 a Bayesian sampling procedure, 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 in 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 mechanism underlying hypothesis generation is impaired or disrupted in cases of delusion. That would be the reason why a delusional agent

---

out contextually irrelevant alternatives before assigning subjective probabilities in certain decision-making tasks (Norby, 2015; cf. 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).

assigns some positive degree of subjective probability to a 'non-starter' hypothesis, which would then be subject to further predictive computational processing.

In the previous section, we saw that predictive processing accounts sometimes give the impression of offering a complete explanation by tacitly shifting some of the explanatory burden onto whatever causes an agent's priors. We can now see that, within a predictive processing framework, understanding how hypothesis generation functions also requires an explanation of an agent's priors. Some theorists have claimed that the predictive processing framework can meet this explanatory demand by virtue of the fact it conceives of an agent's priors as 'empirical', or as 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; 2010). However, even if this were plausible for ordinary cases of belief formation, it is very difficult to see how the notion of an 'empirical prior' can help us understand the origin of priors in cases of delusion. A distinguishing characteristic of delusional cognition is insensitivity to empirical evidence. It is hard to see how, for example, a system that generates a hypothesis that one's mother is an alien imposter is being 'informed by empirical data', or 'guided by a particular kind of feedback'. So, an appeal to 'empirical priors' does not really help us understand how a subject generates a delusional candidate hypothesis. In a way this is not surprising. 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.

5. Hypothesis Generation

Despite how little we know about it, in this section I shall briefly present three possible ways in which hypothesis generation could be functionally atypical in cases of delusion. I think it is plausible that a predictive processing disturbance could be implicated in some of these functional differences. However, as we shall see, a complete explanation of why a system produces a delusional candidate hypothesis appears to exceed the resources of the predictive processing framework.

First, it may be that, in comparison to non-delusional agents, a delusional agent over-generates candidate hypotheses. Thus, when confronted with an irregular experience, a delusional agent would generate a candidate set containing more members than a non-delusional agent would. This is exactly the sort of thing we might expect if hypothesis generation involved a kind of cognitive filter, which functioned to immediately rule out candidates that were incompatible with an agent's background knowledge (cf. 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 candidate hypotheses in terms of precision estimation. As we saw earlier, a system that generally expects imprecise error signals would have a predictive model that is poorly constrained by the incoming sensory signal. Since the incoming signal is the fundamental source of constraint on a predictive model, a high expectation of imprecision would mean that little, if anything, functioned to constrain the model. 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 would give us some sense of why a specific 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 more about how hypotheses generation functions in order to know how implausible hypotheses arise in the first place.

A second way in which hypothesis generation may be functionally divergent is that a delusional agent might under-generate candidates. Functionally, the basic idea is that a delusional subject would quickly suspend the process of hypothesis generation, which would lead to a comparatively impoverished candidate set. 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 will be higher in a set of 10 candidates than it will 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.

There is some experimental data that seems relevant to the suggestion that delusional subjects under-generate candidates. Specifically, as we have seen, a lot of evidence indicates that delusional agents have a tendency to 'jump-to-conclusions', in the sense that they stop experimental probabilistic reasoning tasks more quickly than non-delusional agents. This indicates that they are disposed to set artificial time constraints, which plausibly might 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). So, there is some evidence that a delusional agent might under-generate candidate explanations.

What consequences might this have? If an agent fails to encode many mundane hypotheses in her candidate set, then undergoing a highly irregular experience could more easily cause her to generate implausible candidate explanations. The idea would be that as soon all the members of an overly restricted candidate set are deemed inadequate, 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 the candidate set if it is thought of.

Finally, there is a third way in which hypothesis generation could be functionally disturbed in cases of delusion. We have seen that some accounts postulate that delusions arise in response to an unusually precise error signal. However, if an agent were to experience an exceptionally high, exceptionally precise error signal, that might be enough to completely eliminate the agent's active hypothesis space.[16] The thought is that a very powerful error signal would fry the candidate set with which the brain is operating. As a result, the system would need to construct a completely novel candidate set from scratch. However, the complete 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. Thus, if a delusional agent's hypothesis generation system functions in one of these ways, it is plausible that a 'nonstarter' would enter her candidate set if she thinks of it. However, nothing I have said sheds much light on how implausible hypotheses are thought of in the first place. It is reasonable to speculate that socio-cultural factors and contextual parameters play

[16] This idea was first suggested to me by NN in conversation.

crucial roles, but we do not have the slightest idea of how these things determine what a person thinks.[17]

Regardless, 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 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.

[17] The idea that the process of hypothesis generation is functioning irregularly in cases of delusion might be caused by a specific neurocognitive impairment. If that is right, and if an impairment in perceptual processing causes the occurrence of the unusual experience that delusional subjects seek to explain, then there at least two deficits implicated in the formation of a delusional belief. This might appear to speak in favour of so-called 'two-factor' theories of delusion formation (Davies, et. al., 2001). However, it is consistent with everything I have said in this essay that more than two neurocognitive impairments are causally responsible for the onset of a delusional belief. It is also consistent to think the functional irregularity in hypothesis generation is not an impairment but is within the range of normal cognitive functioning.

6. 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 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 a predictive processing 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 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, it is equally a challenge for an alternative theoretical framework. Regardless of which theoretical approach we adopt, it seems to me that a fuller understanding of hypothesis generation is an important step toward developing a complete explanation of why certain individuals form delusional beliefs.[18]

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.

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 D 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.

[18] Acknowledgements

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.

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

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

Corlett, et. al. 2010: 'Toward a Neurobiology of Delusions' *Progress in Neurobiology*, *92*, pp. 345-369.

Corlett et. al. 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, 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., 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*. Taylor & Francis.

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. et. al. 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

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.

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-

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.*

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., 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.

Hempel, C.G., 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. 2015: 'Prediction Error Minimization, Mental and Developmental Disorder, and Statistical Theories of Consciousness', *Disturbed consciousness: New essays on psychopathology and theories of consciousness*, pp. 293-324.

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

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.

Joyce, J. 2003: 'Bayes' Theorem'*, Stanford Encyclopedia of Philosophy*

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*, 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. A. 1974: 'Delusional Thinking and Perceptual Disorder'. *Journal of Individual Psychology*, *30*, 98.

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, p.120.

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

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', *The Oxford Handbook of Philosophy of Perception*

Saks, E. 2007: *The Centre Cannot Hold.*

Skow, B. 2017: *Reasons Why*. Oxford: Oxford University Press.

Stone, T., & Young, A. W. 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

So, S., Freeman, D., Dunn, G., Kapur, S., Kuipers, E., Bebbington, P., and Garety, P. A. 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, p.129.

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, p.155.

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, p.1151.

Williams, D. forthcoming: 'Hierarchical Bayesian Models of Delusion', *Consciousness and Cognition*.