Bayesian Psychiatry and the Social Focus of Delusions

1. Daniel Williams, Research Fellow, Corpus Christi College, dw473@cam.ac.uk

2. Marcella Montagnese, Department of Neuroimaging, Institute of Psychiatry, Psychology & Neuroscience, King's College London

Abstract. A large and growing body of research in computational psychiatry draws on Bayesian modelling to illuminate the dysfunctions and aberrations that underlie psychiatric disorders. After identifying the chief attractions of this research programme, we argue that its typical focus on abstract, domain-general inferential processes is likely to obscure many of the distinctive ways in which the human mind can break down and malfunction. We illustrate this by appeal to psychosis and the social phenomenology of delusions.

Keywords: Bayesian brain; predictive coding; predictive processing; evolutionary psychology; modularity; computational psychiatry; functional specialization; psychosis; social theory of delusion; paranoia

1. Introduction

It is common to think that psychiatric disorders are caused by dysfunctions in or disturbances to the neural mechanisms that underlie human psychology. If so, significant progress in our understanding of psychiatric disorders demands a model of how the healthy or typical brain functions.

In recent decades, a large and growing body of research in cognitive science has sought to model the brain as a statistical inference mechanism, constructing and refining probabilistic models and hypotheses about the world from the streams of noisy and ambiguous information it leaves on our sensory transducers (Doya et al. 2007; Knill and Pouget 2004). For example, theorists have drawn on Bayesian statistics to illuminate learning and inference across a wide variety of cognitive domains, including perception, motor control, intuitive theories, and more (see Doya et al. 2007). A prominent manifestation of this work has been in predictive coding, an influential theory that models the brain as a hierarchically structured prediction machine, comparing internally generated predictions of sensory information against the sensory information generated by the body and environment and striving to minimize the difference between the two (see Clark 2013; Friston 2005; Hohwy 2013; Rao and Ballard 1999).

Such ideas increasingly provide the framework for understanding healthy brain function that guides research in computational psychiatry (Friston et al. 2014; Teufel and Fletcher 2016). Specifically, researchers have sought to model a large range of psychiatric disorders by appeal to dysfunctions or aberrations in the neural mechanics of statistical inference and decision-making, including schizophrenia (Adams et al. 2013), autism (Lawson et al. 2014), Parkinson’s disease (O’Callaghan et al. 2017), anorexia (Gadsby and Hohwy 2019), addiction (Schwartenbeck et al. 2015), depression (Barrett et al. 2016), and more. As Griffin and Fletcher (2017, p.265) put it,

> “The growing understanding of the brain as an organ of predictive inference has been central to establishing computational psychiatry as a framework for understanding how alterations in brain processes can drive the emergence of high-level psychiatric symptoms” (Griffin and Fletcher 2017, p.265).

Some proponents of this approach are extremely optimistic about its explanatory reach. Carhart-Harris and Friston (2019, p.334), for example, argue that “most, *if not all*, expressions of mental illness can be traced to aberrations in the normal mechanics of hierarchical predictive coding” (our emphasis).

We have two principal aims in this article. First, we will identify and clarify some of the core theoretical attractions of what we call “Bayesian psychiatry” as a research programme. Second, we will argue that this research programme is often hindered by a focus on content-neutral, domain-general inferential processes that abstract away from much that is distinctive about human psychology. Drawing on psychosis and the social nature of clinical delusions to illustrate, we will argue that this focus likely blinds Bayesian psychiatry to many specific ways in which the human mind can break down and malfunction.

We structure the article as follows. In Sections 2 and 3 we introduce Bayesian psychiatry (S2) and outline applications of this research programme to understanding psychosis (S3). In Section 4 we draw on the distinctive social phenomenology of psychosis to argue that such applications seem inadequate and in Section 5 we suggest that combining Bayesian modelling with information about the functional specializations of the human brain might help to address this problem. We conclude in Section 6 by summarising our conclusions and highlighting important areas for future research.

2. Bayesian Psychiatry

Computational psychiatry seeks to build computational models of the dysfunctions and aberrations that underlie psychiatric disorders. It is built on two central ideas: first, psychiatry should strive to trace psychiatric disorders to dysfunctions in neural mechanisms; second, neural mechanisms are computational mechanisms, that is, mechanisms that extract and process information through transformations of and operations over information-encoding states and structures. Computational modelling of psychiatric disorders brings many theoretical benefits. For example, it provides an explanatorily illuminating link between neurobiological and psychological levels of description, it forces theories – and thus the predictions of theories – to be explicit and mathematically precise, and it grounds psychiatric explanation in independently well-established models of brain function in computational and cognitive neuroscience (see Friston et al. 2014; Teufel and Fletcher 2016).

Consonant with their broader influence in neuroscience, computational psychiatry has been dominated by neural network models, reinforcement learning models, and Bayesian models, the latter of which constitute our focus here. Bayes' theorem is an implication of probability theory that specifies the optimal procedure for redistributing the probabilities assigned to hypotheses in light of new information. Specifically, the aim of Bayesian inference is to calculate the posterior probability of a hypothesis conditional on novel evidence, p(hypothesis | evidence). Bayes' theorem states that this is proportional to how well the hypothesis predicts the evidence, i.e. the likelihood p(evidence | hypothesis), multiplied by the hypothesis's probability before encountering the evidence, i.e. the prior p(hypothesis). To calculate the posterior, this product is then divided by the categorical probability of the evidence, the marginal likelihood p(evidence), which is typically calculated as a sum (for discrete states) or integration (for continuous states) over the product of the priors and likelihoods for all possible hypotheses. Importantly, exact Bayesian inference of this sort is often infeasible when dealing with large or continuous hypothesis spaces. Thus, statistics and artificial intelligence have developed various algorithms for approximating Bayesian inference, the most influential of which are stochastic sampling approximations and deterministic variational approximations.

The growing importance of Bayesian inference and its approximations as a model of neural information processing can be traced to two principal factors. First, neuroscientists have increasingly recognised that inductive inference under profound uncertainty is a fundamental problem that the brain confronts. Bayesian inference provides the optimal method for solving this problem. Thus, it is argued that we should expect – perhaps on evolutionary grounds – that the brain implements some form of this solution. As Mathys et al. (2011, p.1) et al. put it,

> 'Since a Bayesian learner processes information optimally, it should have an evolutionary advantage over other types of agents, and one might therefore expect the human brain to have evolved such that it implements an ideal Bayesian learner'.

Second, experimental neuroscientists and cognitive psychologists have uncovered evidence across a wide variety of domains that human inference is approximately Bayes-optimal (see Clark 2013; Knill and Pouget 2004).

Both factors have motivated the Bayesian brain hypothesis, the hypothesis that information processing in the brain – or at least certain parts of the brain – is approximately Bayesian (Knill and Pouget 2004). However, this hypothesis is silent on how the brain implements approximate Bayesian inference. One of the most influential theories for addressing this issue is hierarchical predictive coding.

Strictly speaking, predictive coding is an encoding strategy in which only the unpredicted elements of a signal are fed forward for further stages of information processing. In neuroscience, this encoding strategy was advanced as a model of visual processing by Rao and Ballard (1999), which proposes that cortical networks acquire and update probabilistic models of the causes of sensory signals through a process in which successive levels of cortical hierarchies attempt to minimize the error in their predictions of activity registered at the level below them. In recent work, however, "predictive coding" often refers to an extension and elaboration of this model of perceptual processing to encompass neural information processing more generally. Thus, Sterzer et al. (2017) write that,

> ‘Predictive coding conceives of *the brain* as a hierarchy whose goal is to maximize the evidence for its model of the world by comparing prior beliefs with sensory data, and using the resultant prediction errors (PEs) to update the model’ (our emphasis).

We will use the term “predictive processing” to refer to this more global theory of brain function (see, e.g., Clark 2013; Friston 2005; 2010; Hohwy 2013). There are two aspects of predictive processing that will be important in what follows. The first is a conception of neural information processing in terms of hierarchical precision-weighted prediction error minimization. “Prediction error” refers to the divergence between the brain’s predictions of incoming information and the information itself. “Precision” names the inverse of the variance of a probability density, and thus the degree of certainty or confidence associated with it. Precision-weighting therefore adjusts the degree to which predictions are updated in light of prediction errors as a function of the relative uncertainty associated with prior expectations and incoming evidence. Consonant with the hierarchical structure of the neocortex, this process of uncertainty-weighted prediction error minimization is thought to be iterated up an inferential hierarchy, with each successive level attempting to predict activity at the level below it.

The second component is the idea that prediction error minimization constitutes the overarching function, goal, or “imperative” of the brain (Friston 2010; Hohwy 2013). This radical view is often motivated by or grounded in the broader idea, central to the free energy principle (Friston 2010), that all self-organizing systems obey an overarching imperative to minimize surprise or maximize model evidence. On this view, action – typically described as active inference – is also modelled as a form of prediction error minimization, except that rather than updating predictions to match incoming sensory information, action involves intervening on the world to match sensory information to the brain’s expectations, the most fundamental of which are thought to be installed by evolution (see Hohwy 2013).

These broad ideas about brain function have played a central role in a large and growing body of research in computational psychiatry (see Friston et al. 2014; Teufel and Fletcher 2016). We will henceforth use the term "Bayesian psychiatry" to refer to this broad research programme. Setting aside the details and potential criticisms of specific theories and hypotheses within Bayesian psychiatry, we believe that it constitutes a promising framework for modelling psychopathologies for reasons over and above the more general theoretical attractions of computational psychiatry.

First, the Bayesian brain and related ideas were not developed to explain psychiatric disorders, but rather have compelling independent support as models of neural information processing (see Knill and Pouget 2004). Second, conceptualising the brain as an inferential organ provides an explanatorily illuminating link between the biological mechanisms that implement information processing and the intentionality and the role of misrepresentation essential to many psychiatric disorders (Friston et al. 2014). Third, Bayesian psychiatry draws attention to the important role of uncertainty and uncertainty management in psychiatric disorders (Hohwy 2013). Fourth, the emphasis on bi-directional hierarchical processing within theories such as predictive coding constitutes a promising framework for understanding the complex and often bi-directional interplay between percepts, beliefs, and more abstract self-narratives that are central to many psychiatric disorders (Sterzer et al. 2018). Finally, and most importantly, Bayesian psychiatry has undeniably been explanatorily fecund, generating myriad novel conceptualisations and surprising predictions (Teufel and Fletcher 2016).

In the next section, we will illustrate these attractions by appeal to influential predictive coding models of psychosis.

3. Psychosis and Bayesian Psychiatry

Psychosis is a complex and heterogeneous functional disorder that is generally understood as an impairment in reality testing. The term is used as an umbrella category for a cluster of symptoms that comprise hallucinations and delusions that can occur across many psychiatric, neurodevelopmental, and neurodegenerative disorders. Thus, psychotic symptoms have been widely researched in affective disorders such as bipolar disease (Shinn et al., 2012) and in neurodegenerative ones such as Parkinson's Disease (Fénelon et al., 2010) and dementia with Lewy bodies (Waters et al., 2014). The clinical manifestation of psychosis is varied and heterogenous across these nosological categories, as well as within each disorder. Here we will focus mostly on psychosis in schizophrenia, where much of the research within Bayesian psychiatry on psychosis has been focused.

3.1. Psychosis in Schizophrenia

Schizophrenia is a mental disorder affecting 0.3 to 0.7% of the population worldwide (DSM 5th ed., 2013). Patients diagnosed with schizophrenia can show a heterogeneity of symptoms, which are classified as either positive or negative, where positive symptoms include hallucinations and delusions and negative symptoms include a lack of useful goal-directed behaviours (Garofalo et al., 2017), anhedonia (i.e. a lack of anticipation and seeking of rewards), poverty of speech, and asociality (Frith, 2005). Research has shown that the aetiological roots of schizophrenia span from genetic risk factors (Tsuang, 2000) to social and environmental ones (Mortensen et al., 1999; see below Section 5), including complex interactions between them (Ursini et al., 2018). Further, it is important to distinguish changes in patients' psychopathology across time. For example, chronic patients with schizophrenia tend to have fixed delusions, whilst these tend to be less immovable in those at early stages of the disorder, such as in First-Episode Psychosis, where individuals often retain insight about the implausibility of delusional thoughts (see Sterzer et al., 2018). Even though the exact causes of schizophrenia are still not well understood, there is abundant evidence implicating different neurotransmitters (especially dopamine and glutamate) and multiple brain areas (see Gill and Grace 2016).

Hallucinations take different forms in schizophrenia, with heterogeneous manifestations across different sensory modalities, although auditory hallucinations are the most studied (Montagnese et al. 2020). These tend to revolve around hearing voices, either individual or in conversation, which often generate a running commentary on the individual's behaviour. Although the specific contents of delusions vary widely across individuals and cultures, they tend to cluster in a surprisingly small subset of themes, almost all of which concern the individual's standing in the social world (Bentall et al. 1991; Gold 2017). Here, we will focus largely on the most common form of delusions, persecutory delusions, an extreme form of paranoia which involves the unsubstantiated belief that an agent or group of agents wants to harm the delusional individual (Freeman 2016).

3.2. Predictive Coding and Psychosis

The most important precursors to predictive coding models of psychosis are those that posit a dysfunction in the integration of sensory experience, learned expectations, and higher-level explanations of such experiences (see Sterzer et al. 2018). For example, Maher (1974) famously proposed that delusions are best understood as reasonable responses to anomalous experiences caused by dysfunctions in or damages to perceptual mechanisms. Building on this research and on the aforementioned work implicating dopamine in psychosis, Kapur (2003) suggested that dopaminergic dysregulation in schizophrenia might disrupt the attribution of salience to stimuli. According to this influential aberrant salience hypothesis, seemingly irrelevant events and stimuli elicit excessive attributions of salience and delusions are understood as the individual's attempts to make sense of and explain such anomalous experiences.

Another influential precursor comes from the model of control of intended action developed by Frith et al. (2000). Here one's sense of agency can be seen as emerging from the integration of different agency cues, including both internal (e.g. from processes serving motor control) and external cues (e.g. feedback from sensory systems), as well as prior information, where each kind of information is weighed by its reliability. To feel like the *agent* of one's actions, this model holds that agents must be able to reliably anticipate the sensory consequences of such actions. A failure in such prediction will thus render one's own behaviour surprising, thus suggesting an external cause. By extending this framework to psychosis more generally (Moore and Fletcher, 2012), positive symptoms can be seen as emerging from repeated confusion between external and internal origins of sensory data. Experimental evidence confirms this loss of normal attenuation of sensory feedback for motor action in patients with psychosis (Shergill et al, 2005; Blakemore et al., 2000).

Such ideas have laid the groundwork for the development of what Sterzer et al. (2017, p.634) call the "canonical predictive coding account of psychosis." According to this model, the emergence of psychosis can be explained in terms of a dysfunction in the interaction between and integration of top-down expectations and bottom-up information. As noted above, optimal prediction error minimization necessitates that prediction errors are effectively weighted by their precision or certainty. The canonical predictive coding accounts posits that this process of precision-weighting is disrupted in psychosis, such that sensory data is assigned too much precision relative to higher-level, more abstract expectations, leading to maladaptive statistical inference and learning and thus the development of inaccurate models of the world (see Adams et al. 2013; Clark 2016). Further, because of the bi-directional interaction between perceptual experiences and higher-level beliefs within predictive coding, inaccurate inferences at lower levels of the inferential hierarchy both influence and are influenced by maladaptive higher-level expectations, driving both hallucinations and delusions and a complex interplay between them (see Sterzer et al. 2017). Except when stated otherwise, reference to the predictive coding model of psychosis in what follows refers to this canonical model.

4. The Social Contents of Delusions

The canonical predictive coding model of psychosis has many well-advertised attractions (see Sterzer et al. 2018). For example, predictive coding comes with an implementational theory in which precision-weighting is regulated by the action of neuromodulators such as dopamine, and, as noted, there is substantial independent evidence that dopamine dysregulation plays a causal role in psychosis. Further, there is compelling neuro-imaging and behavioural evidence that individuals with psychosis do exhibit deficits in prediction error-driven learning and probabilistic reasoning. Finally, there are interesting simulations demonstrating that aberrations in precision-weighting generate effects similar to those observed in individuals with psychosis, including in psychological domains such as visual tracking distinct from psychosis itself (see Adams et al. 2013). Nevertheless, this theory also faces several objections and challenges (see Bell et al. 2019; Williams 2018; and Sterzer et al. 2018 for a review). Here, we focus on just one: namely, how to reconcile the hypothesis that psychosis results from a domain-general dysfunction of the sort posited by this theory with the apparent domain specificity of psychosis itself.

To see this problem, first consider how schematic the proposed account of psychosis is. Summarising this explanation, for example, Clark (2013, p.197) writes that

> "understanding the positive symptoms of schizophrenia requires understanding disturbances in the generation and weighting of prediction error… [M]alfunctions within that complex economy… yield wave upon wave of persistent and highly weighted "false errors" that then propagate all the way up the hierarchy forcing, in severe cases… extremely deep revisions in our model of the world. The improbable (telepathy, conspiracy, persecution, etc.) *then becomes the least surprising*…" (our emphasis).

However, this explanation leaves it opaque why the contents of common delusional themes such as persecution and conspiracy should constitute the least surprising hypotheses about the world in light of aberrant precision-weighting. Specifically, although dysfunctions in precision-weighting and prediction error-driven processing can explain why individuals process information in *abnormal* ways and thus form beliefs that appear implausible to those not suffering from the relevant dysfunction, an adequate explanation of delusions must explain why individuals form the highly specific delusional beliefs that they come to hold (Parrott 2019). That is, psychosis demands an explanation of the distinctive way in which psychotic experience is abnormal out of the vast space of possible ways in which it could deviate from normal perception and belief but does not.

Focusing specifically on delusions, the predictive coding model conforms to the standard view in the psychiatric literature that the explanandum should be characterised in a way that is content-neutral. Thus, the DSM-5 defines clinical delusions as "fixed beliefs that are not amenable to change in light of conflicting evidence" (American Psychiatric Association 2013, p.87). Setting aside the problem that this definition subsumes many widespread non-delusional (e.g. religious, ideological, self-serving) beliefs, it characterises delusions in a way that focuses on their purely formal characteristics, and specifically their irrationality. It therefore invites the view that delusions result from inferential or reasoning abnormalities (see Gold 2017). Further, because the definition is content-neutral, it strongly suggests that such abnormalities afflict domain-general inferential processes ranging over all possible contents of thought. This is a deep problem, however, because – as highlighted above – the distribution of delusional beliefs is not a random sample of all possible abnormal beliefs, but a highly specific subset, almost all of which concern the individual's standing in the social universe (see Bell et al. 2019; Gold and Gold 2015).

Further, it is not clear that delusional subjects do exhibit any significant domain-general inferential impairments or reasoning abnormalities (see Bell et al. 2019; Gold 2017). At best, the voluminous body of empirical research attempting to identify such impairments is inconclusive. Perhaps the most influential proposal in this area – often taken as support for the predictive coding model of psychosis (Adams et al. 2013) – is that delusional subjects suffer from a "jumping to conclusions" bias (Garety 1991). In the famous "beads task", for example, participants are told that there are two jars, A and B, with jar A containing 85% red beads and 15% black beads and jar B containing the reverse. On the basis of drawing beads from a jar, participants are asked to judge which of the jars the beads come from. The core finding is that individuals with psychosis tend to form a judgement on the basis of fewer beads than controls (Garety 1991). In addition, recent meta-analyses also indicate small-to-moderate effect sizes when it comes to other reasoning biases (McLean et al. 2017).

There are problems with this research, however. For example, often the alleged differences between delusional subjects and healthy individuals disappear when controlling for general cognitive function, which is known to be reduced in individuals with psychotic symptoms (Bell et al. 2019). In some meta-analyses, such as McLean et al's (2017), theorists do not control for possible confounds of this kind. Further, the domain-general reasoning differences between delusional subjects and healthy controls are typically small, especially when compared to the striking deviations from normality observed in psychosis. Thus, the relevant question is not whether delusional subjects exhibit domain-general differences in inference relative to neurotypical controls, but whether – and, if so, in what way – such differences are *causally responsible* for the formation and entrenchment of delusional beliefs. The relatively small differences in domain-general inference that have been discovered in the empirical literature suggest that such differences might be better understood as effects of other underlying factors associated with but not responsible for psychosis, or else factors that function as necessary but not sufficient causes of psychotic experience and delusions.

Importantly, proponents of the predictive coding model of psychosis are aware of at least the first of these problems. Thus, Griffin and Fletcher (2017, p.272) refer to

> "the paradox of why, given that we are positing a very domain-general problem with weighting information by its reliability in Bayesian inference, delusions tend to be domain specific in their content, which usually "seem to concern the patient's place in the social universe" (Bentall et al. 1991, p.14)."

There are various responses available to proponents of the predictive coding approach. One strategy is to appeal to the contents of specific experiences. As noted above, an influential theory dating back to Maher (1974) is that delusions constitute attempts to explain – and thus derive their contents from – anomalous experiences. This seems highly applicable in many cases. For example, Capgras delusion has famously been connected to a dysfunction in which facial recognition is disconnected from interoceptive mechanisms in such a way that individuals cognitively recognise loved ones but fail to experience any of the typical autonomic (i.e. affective) cues that accompany such recognition (Langdon and Coltheart 2000). This violation of expectations cries out for explanation, thus generating the thought that perhaps the "loved one" is really an imposter. Similarly, influential precursors to predictive coding described above trace the hallucinated voices and illusions of control that are common in psychosis to dysfunctions in sensory predictive mechanisms that make the individual's own voice and actions seem surprising, thus suggesting an external cause.

Nevertheless, although it is extremely likely that anomalous experience plays an important causal role in delusion formation, there are two problems with locating delusional contents wholly in perceptual experiences. The first is that even in cases such as Capgras where one can identify a specific anomalous experience, there is still the question of why delusional subjects gravitate towards specific delusional *hypotheses*. As has been widely noted, for example, positing an imposter looks like an exceptionally implausible explanation in such cases (see Parrott 2019). Not only is the belief in tension both with many other beliefs that people hold in general (i.e. about the limits of disguise) and with the testimony of doctors and trusted love ones, but there appear to be many other, more plausible explanations of the relevant experience (e.g. "there's something wrong with me").

Second, the canonical predictive coding account of psychosis is supposed to apply in cases where there are no specific anomalous experiences over and above those generated by aberrant precision-weighting. One might respond that aberrant precision-weighting provides a computational-level description of – and can thus draw on the explanatory resources of – Kapur's (2003) influential "aberrant salience" model of psychosis described above, according to which dopaminergic dysfunction (here understood as aberrant precision-weighting) causes otherwise irrelevant stimuli and connections between stimuli to strike the agent as highly salient and thus in need of explanation. Once again, however, tracing delusions to a domain-general aberration in salience attribution predicts that delusional beliefs will range freely over all possible topics of attention (Gold and Gold 2015). Further, it is unclear why a hyper-attention to otherwise irrelevant low-level sensory stimuli (driven by highly weighted low-level sensory prediction errors) does not merely generate an immersion in the sensory world of the sort observed in autism. Indeed, as has been noted (Sterzer et al. 2017), the dominant predictive coding account of autism (Lawson et al. 2014) looks highly similar to the canonical predictive coding account of psychosis, which is a problem given the substantial dissimilarities in their associated symptoms.

Another suggestion is that social cognition is likely to be differentially impaired by a domain-general dysfunction in precision-weighting. For example, Griffin and Fletcher (2017, p.276) write that

> "social cues may be inherently more uncertain than non-social ones, because they rely on inferring intentions from ambiguous physical acts. Consequently, representations of the social world could be the first to break down when the system encounters a relatively minor impairment in uncertainty-weighting inference…"

Even if one accepts that social inference is more difficult than non-social inference, however, the social focus of delusions is not characterised by a general breakdown in social inference. For example, persecutory delusions are distinctive not just because they diverge from ordinary, non-delusional beliefs about the social world, but because of the malign and self-directed intentions that they attribute to other agents. Why should a paranoid stance towards the social world result from greater *uncertainty* or *difficulty* in social inference? One suggestion is that "aberrant predictive coding could render other people *unreliable*, to be treated with suspicion" (Griffin and Fletcher 2017, p.276; our emphasis). To quote Griffin and Fletcher (2017, p.276) again,

> "Just as reduced discriminability in PE [prediction error] signalling could lead to a consistent sense of unease or surprise, so too could reduced discriminability between social sources make everything (and everyone) seem uniformly unreliable, even suspicious."

Again, however, even granting that aberrant precision-weighting might make other people seem unreliable – and it is not clear why the substantial divergence between the individual's beliefs and other people's does not make her question her own reliability – unreliability need not entail suspicion or the attribution of malign intentions. Astrologists are unreliable, but we do not generally assume that they are part of a hidden plot to do us harm. Further, in the case of persecutory delusions, people's unreliability manifests itself primarily in disagreement over the veracity of the delusions, suggesting that the paranoia is the cause of the epistemic estrangement from other people, not the effect of such estrangement.

Finally, in recent work theorists have posited a novel kind of domain-general inferential difference as a potential driver of paranoia. In a set of fascinating experimental studies, Reed et al. (2020) demonstrate that paranoid individuals expect greater volatility relative to non-paranoid controls in a non-social learning task, and they show that this greater expectation of volatility can be reproduced in rats exposed to methamphetamine, a drug that is known to increase paranoia in humans. They (Reed et al. 2020) take this as "evidence of fundamental, domain-general learning differences in paranoid individuals" (p.1), and thus hypothesize "that aberrations to these domain-general learning mechanisms *underlie* paranoia" (p.2; our emphasis).

Granting the existence of such domain-general differences, however, the explanatory connection between a greater expectation of volatility and the specific focus and contents of paranoia and persecutory beliefs is opaque. Reed et al. (2020, p.2) write that "since excessive unexpected uncertainty is a signal of change, *it might drive the recategorization of allies as enemies*" (our emphasis). Why should higher levels of expected volatility drive the recategorization of allies as enemies rather than the reverse, however, or no change in their status at all? Reed et al. (2020, p.29) suggest that "when humans experience non-social volatility… they appeal to the influence of powerful enemies, even when those enemies' influence is not obviously linked to the volatility," but positing malevolent agency as the explanation of volatility without sufficient evidence constitutes an implausible – and so presumably unlikely – explanation of volatility. They also suggest that "with a well-defined persecutor in mind, a volatile world may be perceived to have less randomly distributed risk" (Reed et al 2020, p.29). It is not clear how connecting volatility to the seemingly unrelated actions of agents with hidden and inexplicitly malevolent intentions towards oneself – intentions which must be as volatile as the events they cause – is supposed to reduce uncertainty, however. Further, it is opaque why populating the world with malevolent agency directed towards oneself should be a desirable psychological outcome even if it did reduce uncertainty.

Importantly, our point here is not to deny that human beings might be biased towards suspicion and paranoia of the sort highlighted in these explanations. Our point is that these biases are independent of any proposed difference in domain-general statistical inference, and thus illicitly imported in from contingent assumptions about the human mind. These assumptions might be correct, but they reflect aspects of human psychology that are not themselves logical consequences of domain-general aberrations in statistical inference.

4.1. Summary

To summarise, the predictive coding account of psychosis is both attractive and problematic. Although there is compelling evidence that some form of dysfunction in uncertainty estimation plays a causal role in psychosis, it is difficult to reconcile such a domain-general explanation with the conspicuous domain specificity of psychotic symptoms, especially when it comes to delusions. Attempts to avoid this conclusion are either unconvincing or end up importing contingent assumptions about human psychology external to the model itself and beyond the scope of content-neutral, domain-general learning differences.

Crucially, this problem seems to stem directly from the emphasis on abstract, domain-general inferential processes within Bayesian psychiatry more generally. As noted above, this framework is often aligned with predictive processing, a global theory of brain function in which the brain is viewed as a general-purpose uncertainty management mechanism operating in the service of a single, overarching epistemic goal – namely, minimizing (long-term, average) prediction error or maximizing model evidence (see Friston 2010; Hohwy 2013). Thus, Adams et al's (2013, p.10) article outlining the canonical predictive coding account of psychosis involves "[s]tarting with the assumption that the brain is trying to maximize the evidence for its model of the world…" (our emphasis). Given this assumption, it is difficult to see how the account that they develop could locate psychosis in anything *but* a content-neutral, domain-general dysfunction in statistical inference. This assumption abstracts away from almost all of the distinctive functions, motives, interests, and concerns of the human mind, however. Thus, perhaps by integrating such contingent features of human psychology back into the framework and its starting assumptions, one might be able to address the explanatory gap described in this section. We turn to this possibility next.

5. A Bayesian Social Theory of Delusions

In recent years, a prominent social framework for understanding delusions has emerged (see, e.g. Bell et al. 2019; Gold 2017; Gold and Gold 2015; Raihani and Bell 2019). Although highly schematic, the unifying idea underlying this approach is that we should understand delusions not primarily in terms of domain-general inferential impairments but rather in terms of an evolved social psychology adapted to the recurring features, opportunities, and risks encountered in human social life. It is easy to see why this framework has been opposed to the predictive coding account – or, more generally, accounts – of psychosis (see, e.g. Bell et al. 2019). In this section, we briefly outline this framework and then argue that it can in fact be reconciled with Bayesian psychiatry once the latter's focus on abstract, domain-general inferential processes is replaced with a richer view of human psychology in which statistical inference mechanisms operate in the context of the distinctive and often idiosyncratic functions of the human mind.

5.1. The Social Approach to Delusions

The social approach to delusions is motivated by some of the facts outlined above: for example, that evidence of significant domain-general inferential differences between delusional subjects relative to neurotypical controls is weak, and that the actual delusional themes that occur cluster in a tiny region of the vast space of possible themes, with the overwhelming majority concerning the social world (Bell et al. 2019; Gold and Gold 2015). According to proponents of a social approach, these and other explananda suggest that delusions are better understood in terms of dysfunctions in psychological mechanisms specialized for the distinctive problems and opportunities of human social life. As Gold and Gold (2015, p.289) put it, "To understand delusions, one has to understand the history of human sociality." Thus, this approach takes its inspiration from an evolutionary framework for understanding human psychology, according to which the human mind is best understood not as a general-purpose statistical inference mechanism but as a mosaic of specialised mechanisms adapted to the distinctive features, opportunities, and risks of human life (see Del Giudice 2018).

Although human social dynamics exhibit massive variation across place and time, this variation is underpinned by certain core characteristics. Most fundamentally, human social life is characterised by a complex interplay between cooperation and competition at multiple scales, including both within and between groups. Success within such environments is thus dependent on substantial social support, protection, and interpersonal coordination in the service of shared goals, but such cooperation is always fragile given the diverse and often divergent interests of individuals and groups competing for dominance, prestige, and resources. Further, the difficulties of navigating such opportunities and risks are amplified by the suite of unique human traits that underpin cooperation and competition, including sophisticated communication abilities (along with the attendant risk of deliberate deception), flexible and reliable mindreading, and highly developed reasoning capacities that facilitate long-term plans and complex behavioural strategies.

How might such characteristics have selected for a psychological apparatus vulnerable to delusion? One proposal concerns the evolution of psychological mechanisms concerned with detecting and responding to social threats (see Gold and Gold 2015). "Social threat" here names a heterogeneous category of costs imposed by other agents and coalitions of agents, including those generated by outright violence, exploitation, betrayal, free riding on one's investments, and more. Such threats are ubiquitous and have likely constituted the most significant danger to individual survival and reproductive success throughout our ancestral past (Dunbar 1998). It is thus highly unlikely that the human mind has evolved to learn about the costs, cues, and sources of such threats wholly from experience. Such a blank slate would be quickly outcompeted by agents structured in advance of experience to detect, respond to, and actively learn about this recurring risk of human social life.

What characteristics would one expect from psychological mechanisms specialised for navigating social threats? First, one would expect them to err on the side of caution (see Gold and Gold 2015). That is, given the high – and potentially catastrophic – risk of social threats, false positives are likely to be less costly than false negatives. Further, this cost asymmetry is exacerbated by the fact that an absence of evidence of social threat does not imply evidence of its absence, especially given that the sources of such threats have the capacity and motivation to deliberately conceal their intentions from us. Thus, once a genuine suspicion of threat is activated, one would expect this suspicion to be difficult to assuage, and for agents to downgrade their level of trust in threat-related testimony. In these ways the structural characteristics of threat detection might have selected for a mild form of paranoia – or at least hypervigilance – even in properly functioning mechanisms (Raihani and Bell 2019).

Second, one would also expect the threshold for threat detection – and, by corollary, social trust – to be calibrated to the characteristics of the social environment that individuals encounter. That is, just as social threat detection mechanisms should motivate individuals to learn about and detect specific cues of potential threats, they should also modulate the threshold for threat detection in response to the more general statistical characteristics of the environment. There is considerable evidence for conditional adaptation of this kind, which involves adjustments to the structural development of mechanisms (including information-processing mechanisms) in response to environmental cues, especially during sensitive periods such as childhood (see Del Giudice 2018). Thus, early and/or recurrent exposure to social stressors and exploitation would be expected to lower the threshold for social threat detection, sometimes in ways that are extremely difficult to change.

Third, mechanisms of social threat detection need to combine both fast and automatic detection of threats – and attention to the potential cues of threats – posed by the immediate environmental context with a powerful motivation to ruminate and reflect on the possibility of more distant threats generated by complex, future-oriented and deliberately concealed intentions (see Gold and Gold 2015). That is, threat detection is not merely – or even mostly – a perceptual function but must draw on the resources of reasoning capacities capable of both integrating information from diverse sources and of exploring complex hypothetical risk scenarios and possibilities.

Finally, one consideration that has not – to the best of our knowledge, at least – been explored in the psychiatric literature is the fact that one's beliefs about social threats provide important information to other agents. Thus, beliefs about the likelihood of social exploitation might be influenced by social signalling pressures that adjust one's level of suspicion not just to the available evidence but to the deterrent effect of one's suspicion on others (see Williams 2019). In *The Godfather*, for example, Don Corleone exemplifies how deterrence can motivate a kind of strategic irrationality when he informs fellow mafia bosses of his willingness to jump to conclusions without evidence:

> "… I'm a superstitious man, and if some unlucky accident should befall [my son], if he should get shot in the head by a police officer, or if he should hang himself in his jail cell, or if he's struck by a bolt of lightning, then I'm going to blame some of the people in this room."

Such considerations help to clarify what is meant by "functional specialization." Gold and Gold (2015; see also Gold 2017) posit a "suspicion system" for detecting and responding to social threats, but this terminology carries the unfortunate connotation of a discrete self-contained cognitive module. As the foregoing suggests, adaptive threat detection requires mechanisms that integrate information from a variety of different sources, that are capable of substantial learning, and that modulate the activity of other psychological mechanisms involved in attention, deliberation, action, and so on. Such information-processing mechanisms and procedures are thus not self-contained and are certainly not realised in a discrete anatomical module at the macroscopic level of brain structures. Nevertheless, such mechanisms are still specialised insofar as their characteristics would not be appropriate for many other cognitive tasks, such as estimating the spatial layout of the environment, forecasting the weather, or parsing the syntactic structure of a sentence.

As noted, even properly functioning mechanisms of social threat detection might exhibit signs of paranoia. Now consider a dysfunction in such mechanisms, however. This dysfunction could make individuals less capable of detecting and responding to social threats, and thus extremely vulnerable to exploitation. Equally, however, it could make individuals overly sensitive to the possibility of social threat, driving their attention towards and ruminating on the possibility of such threats in a way that will appear wholly disconnected from objective evidence to other agents. At first, this hypersensitivity to social threat might be reined in by conscious reflection on the implausibility of paranoid thoughts. Over time, however, hyperactive threat detection might result in an accumulation of evidence that overcomes such rational defences and gives rise to entrenched persecutory beliefs, driving conscious reasoning away from challenging paranoid thoughts and towards integrating them with the rest of the individual's worldview.

This is the essence of the model of persecutory delusions advanced by Gold and Gold (2015; Gold 2017). As they note, it has myriad attractions. First, it explains why persecutory delusions have the specific theme that they do. Of course, contingent features about the relevant individual's time and cultural milieu will no doubt influence what kinds of social threats are salient to them, but this model has the advantage of explaining why social threat in general is such a common theme of delusional ideation. Second, it accounts for the relatively weak differences in domain-general inference found in the empirical literature. Although this model is consistent with such differences (see below), it suggests that they are not the sole driver of delusions. Third, it illuminates powerful correlations found between various forms of social adversity (e.g. trauma, abuse, exploitation) and the risk of clinical paranoia (see Raihani and Bell 2019). As noted above, conditional adaptation might have selected for a lower threshold for threat detection in response to such circumstances. Finally, there is some direct – albeit fairly limited – evidence that social threat detection is specifically impaired in conditions such as schizophrenia (see Gold and Gold 2015; Gold 2017).

Of course, as sketched here and as found in Gold and Gold's proposal, this model is highly schematic. For example, it may be that dysfunctional mechanisms underlying persecutory delusions do not track social threat as such but – at least in many cases – specific coalitional threats, which could account for why severe paranoia often involves misperceptions of coalitional boundaries and collective action (Raihani and Bell 2017). Further, persecutory delusions are obviously not the only kind of delusion. Nevertheless, this model illustrates a much more general approach to understanding delusions, one that explicitly connects the contents of delusional ideation and beliefs to the distinctive concerns, motives, and functions of the human mind, and the psychological mechanisms specialised for such distinctive characteristics (Bell et al. 2019; Del Giudice 2018; Gold and Gold 2015).

5.2. A Bayesian Social Approach to Delusions

This social approach to understanding delusions appears to conflict with the predictive coding model – or, more generally, models – of psychosis (see Bell et al 2019). Nevertheless, the apparent domain specificity of delusions need not be in tension with Bayesian modelling *as such*. Indeed, there are various ways in which Bayesian inference generally – and uncertainty-weighted prediction error minimization specifically – could accommodate domain specificity. Most obviously, domain-specific mechanisms might themselves make use of Bayesian computations that infer social threat from ambiguous cues. That is, even if psychological mechanisms are "function-specific, their algorithms needn't be" (Carruthers 2006, p.62). As Sperber (2019, p.36) puts it,

> "[T]he fact that the formal properties of a learning procedure are best specified without assigning to it any specific domain or goal does not entail that the use of such a procedure in an organism or a machine cannot be tied and adjusted to specific goals."

Such adjustment to specific goals or functions might take various forms. For example, it might involve domain-specific priors (Sperber 2019). Given the ubiquity and risks of social threats, it is highly like that humans have priors concerning the presence of such threats both in general and in specific contexts that need not be acquired wholly from experience. More subtly, a central issue for Bayesian inference concerns the hypothesis space itself (Parrott 2019). In principle, an infinite number of hypotheses could explain any given observation, and a real-life Bayesian inference machine cannot consider all of them. Thus, evolution might have endowed the human brain with constraints that narrow and structure the hypothesis space within which Bayesian takes place, including the procedures for generating candidate hypotheses. For example, people might instinctively consider threat-related hypotheses as explanations for events, especially those that strike the person as anomalous or distressing. Further, Bayesian decision theory provides a formal framework for explicitly modelling how asymmetries in the costs of false positives and false negatives modulate judgement and decision-making in different domains (see Williams 2020). Finally, all of these features of Bayesian mechanisms could be adjusted in accordance with conditional adaptation, such that individuals exposed to early social stressors and exploitation might have higher social threat-related priors, a greater motivation to generate social threat-related hypotheses, and a lower threshold for social threat detection.

Given such considerations, there is nothing in the social approach to delusions that is in tension with the idea that the computational architecture underlying delusions makes use of Bayesian inference or prediction error minimization. Indeed, one might view these frameworks as highly complementary, with the social approach proposing distinctive functions and dysfunctions that underlie delusional cognition at a conceptual level and the Bayesian approach generating hypotheses about how such phenomena are implemented in the brain's information-processing mechanisms.

Return to the canonical predictive coding model of psychosis, for example. At the core of this model is the idea that psychosis in schizophrenia is driven by aberrant uncertainty-estimation, with a bias towards assigning greater precision to lower-level sensory prediction errors relative to higher-level, more abstract expectations. As we have seen, there is compelling evidence for this proposal, but it struggles to account for the domain specificity of delusional ideation. Adams et al. (2013, pp.1-2), for example, propose that the failure of precision-weighting that they posit can be "understood intuitively by considering classical statistical inference," where "if we overestimate the precision of the data…. we expose ourselves to false positives." As noted above, however, positing such abstract, domain-general failures in statistical inference as the cause of psychosis fails to account for the highly specific focus of delusional ideation and belief. Now consider how such a domain-general difference in statistical inference might interact with the functionally specialised machinery for social threat detection outlined in the previous sub-section, however.

First, we have already seen that such machinery is likely biased towards false positives independent of any aberration in precision-weighting, perhaps especially so in individuals previously exposed to social stressors. Thus, an additional – and perhaps initially domain-general – bias towards false positives might have a disproportionate effect on social threat processing, with threat-related cues coming to seem even more salient and thus capturing the individual's attention.

Further, as noted, it is similarly plausible that people will have an inherent bias to generate and search for hypotheses positing social threats when confronted with anomalous experiences in general. Further, the tendency to generate such hypotheses is likely to be amplified given the anxiety known to be associated with paranoia and psychosis (see Freeman 2007). For example, Pezzulo (2014) has argued that interoceptive cues of anxiety (e.g. an increased heart rate and galvanic skin response) provide evidence that can bias Bayesian updating towards paranoid inferences that might seem deeply implausible to those without the relevant interoceptive evidence, just as the paranoid hypotheses that occur to us after watching a horror film at night might seem comically implausible to us when we awake the next morning.[1]

[1] The role of anxiety in biasing individuals towards paranoid hypotheses is central to Freeman's (2007) threat anticipation model of paranoia and persecutory delusions.

Once the possibility of social threat is seriously entertained as a consequence of one or both of these factors, the considerations about threat detection described above – for example, the difficulties in finding evidence of the absence of threat, the risks of wilful deception, and the potential source of threats in complex and concealed plans – will motivate individuals to differentially search out, attend to, and ruminate on threat-related information and possibilities, in addition to being on greater guard against the possibility of wilful deception. In this way the motivated search for threat-related information and possibilities might interact with a general oversensitivity to low-level sensory prediction errors to provide additional evidence that fuels the paranoia.

Whilst at first such paranoid forms of informational-sampling and hypothesis generation might be reined in by more global, integrative systems of reflection, over time the apparent accumulation of evidence might overpower such defences, changing the focus of conscious reasoning away from a reasonable scepticism and towards the development of explanations that rationalise the evidence of social threat.[2] Thus, this might explain the transition from a prodromal phase in schizophrenia in which individuals retain insight concerning the implausibility of their paranoia towards to the entrenchment of more fixed persecutory beliefs.

Finally, as the estimated risk of social threat and potential exploitation increases, the motivation for increasing the confidence in one's paranoid thoughts might be further incentivised by the deterrent effects of such paranoia on others. Here the willingness to identify persecutors and conspiracies in a world that has become increasingly distressing serves a protective function, signalling to others a hypervigilance for potential exploitation. Although such conspicuous paranoia might serve this protective function well, however, it

[2] Note that this might also occur due to more direct damage to those regions of the brain that subserve higher-level belief integration and evaluation (see Langdon and Coltheart 2000).

will also further alienate the individual from others and erode social trust, thereby reinforcing the paranoia and its evidential basis further.

Our repeated use of the word “might” in these suggestions should be emphasised. That is, we do not intend these extremely schematic and highly speculative suggestions as a serious model of clinical paranoia and the onset of persecutory delusions. Instead, we have advanced them to illustrate how augmenting a Bayesian approach to understanding psychosis with the content-rich, domain-specific biases and concerns of the human mind helps to broaden the hypothesis space for this approach, thus providing a greater range of potential explanations when it comes to accounting for some of the distinctive features of psychosis and delusional ideation.

6. Conclusion

We are convinced of the explanatory power and fecundity of the Bayesian brain and predictive coding when it comes to modelling the information-processing dysfunctions and aberrations that underlie psychiatric disorders. Nevertheless, we also believe that the emphasis within much of Bayesian psychiatry on highly abstract, domain-general inferential processes likely blinds it to many distinctive features of human psychology and psychopathology. The human brain is not a general-purpose blank slate employing statistical algorithms in the service of dispassionate inference, but the control centre of a unique primate that evolved to navigate a distinct world of opportunities and risks. This control centre might make extensive use of sophisticated statistical learning and inference, but such strategies must be understood in the context of the distinctive features, functions, and interests of the human mind. We have sought to illustrate this lesson by appeal to a highly influential predictive coding model of psychosis, which – we have argued – is currently unable to capture the specific contents of delusional ideation precisely because of its exclusive focus on aberrations in content-neutral, domain-general statistical inference. As noted, we are aware how schematic and speculative our proposals have been for integrating this application of Bayesian psychiatry with a richer, evolutionary framework for understanding human psychology. Nevertheless, we hope that this article motivates more extensive, detailed investigations into this subject in the future.

REFERENCES

Adams R.A., Stephan K.E., Brown H.R., Frith C.D., Friston K.J. (2013) The computational anatomy of psychosis. *Front Psychiatry*. 4:47.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). (DSM-5). Washington, DC: American Psychiatric Publishing.

Barrett, L. F., Quigley, K. S., & Hamilton, P. (2016). An active inference theory of allostasis and interoception in depression. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1708), 20160011.

Bell, V., Raihani, N., & Wilkinson, S. (2019). De-Rationalising Delusions.

Bentall, R. P., Kaney, S., & Dewey, M. E. (1991). Paranoia and social reasoning: an attribution theory analysis. *British Journal of Clinical Psychology*, *30*(1), 13-23.

Blakemore, S. J., Smith, J., Steel, R., Johnstone, E. C., & Frith, C. D. (2000). The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: evidence for a breakdown in self-monitoring. *Psychological medicine*, *30*(5), 1131-1139.

Carhart-Harris, R. L., & Friston, K. J. (2019). REBUS and the anarchic brain: toward a unified model of the brain action of psychedelics. *Pharmacological reviews*, *71*(3), 316-344.

Carruthers, P. (2006). *The architecture of the mind*. Oxford: Oxford University Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181-204.

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

Del Giudice, M. (2018). *Evolutionary psychopathology: A unified approach*. Oxford University Press.

Doya, K., Ishii, S., Pouget, A., & Rao, R. P. (Eds.). (2007). *Bayesian brain: Probabilistic approaches to neural coding*. London: MIT press.

Dunbar, R. I. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, *6*(5), 178-190.

Fénelon, G., Soulas, T., Zenasni, F., & de Langavant, L. (2010). The changing face of Parkinson's disease-associated psychosis: A cross-sectional study based on the new NINDS-NIMH criteria. *Movement Disorders*, *25*(6), 763-766.

Freeman, D. (2007). Suspicious minds: the psychology of persecutory delusions. *Clinical psychology review*, *27*(4), 425-457.

Freeman, D. (2016). Persecutory delusions: a cognitive perspective on understanding and treatment. *The Lancet Psychiatry*, *3*(7), 685-692.

Frith, C. (2005). *The cognitive neuropsychology of schizophrenia* (1st ed.). London: Psychology Press.

Frith, C. D., Blakemore, S. J., & Wolpert, D. M. (2000). Explaining the symptoms of schizophrenia: abnormalities in the awareness of action. *Brain Research Reviews*, *31*(2-3), 357-363.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, *360*(1456), 815-836.

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, *11*(2), 127-138.

Friston, K. J., Stephan, K. E., Montague, R., & Dolan, R. J. (2014). Computational psychiatry: the brain as a phantastic organ. *The Lancet Psychiatry*, *1*(2), 148-158.

Gadsby, S., & Hohwy, J. (2019). Why use predictive processing to explain psychopathology? The case of anorexia nervosa.

Garety, P. (1991). Reasoning and Delusions. *The British Journal of Psychiatry*, 159(S14), 14–18.

Gold, J., & Gold, I. (2015). *Suspicious Minds*. Riverside: Free Press.

Gold, I. (2017). *Outline of a theory of delusion: Irrationality and pathological belief.* In T.-W. Hung & T. J. Lane (Eds.), *Rationality: Constraints and contexts*. Elsevier Academic Press, pp.95-119.m

Griffin, J. D., & Fletcher, P. C. (2017). Predictive processing, source monitoring, and psychosis. *Annual review of clinical psychology*, *13*, 265-289.

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

Kapur S. (2003) Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am J Psychiatry*.160:13–23.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences*, *27*(12), 712-719.

Langdon, R., & Coltheart, M. (2000). The Cognitive Neuropsychology of Delusions. Mind & Language, 15(1), 184–218. https://doi.org/10.1111/1468-0017.00129

Lawson, R. P., Rees, G., & Friston, K. J. (2014). An aberrant precision account of autism. *Frontiers in human neuroscience*, *8*, 302.

Maher B.A. (1974) Delusional thinking and perceptual disorder. *J Indiv Psychol*. 30:98–113

Mathys, C., Daunizeau, J., Friston, K. J., & Stephan, K. E. (2011). A Bayesian foundation for individual learning under uncertainty. *Frontiers in human neuroscience*, *5*, 39.

McLean, B. F., Mattiske, J. K., & Balzan, R. P. (2017). Association of the Jumping to Conclusions and Evidence Integration Biases With Delusions in Psychosis: A Detailed Meta-analysis. *Schizophrenia Bulletin*, 43(2), 344–354. https://doi.org/10.1093/schbul/sbw056

Montagnese, M., Leptourgos, P., Fernyhough, C., Waters, F., Larøi, F., & Jardri, R. et al. (2020). A Review of Multimodal Hallucinations: Categorization, Assessment, Theoretical Perspectives, and Clinical Recommendations. *Schizophrenia Bulletin*. doi: 10.1093/schbul/sbaa101

Moore, J. W., & Fletcher, P. C. (2012). Sense of agency in health and disease: a review of cue integration approaches. *Consciousness and cognition*, *21*(1), 59-68.

Mortensen, P., Pedersen, C., Westergaard, T., Wohlfahrt, J., Ewald, H., & Mors, O. et al. (1999). Effects of Family History and Place and Season of Birth on the Risk of Schizophrenia. *New England Journal Of Medicine*, *340*(8), 603-608.

O'Callaghan, C., Hall, J. M., Tomassini, A., Muller, A. J., Walpola, I. C., Moustafa, A. A., ... & Lewis, S. J. (2017). Visual hallucinations are characterized by impaired sensory evidence accumulation: insights from hierarchical drift diffusion modeling in Parkinson's disease. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *2*(8), 680-688.

Parrott, M. (2019). Delusional predictions and explanations. *The British Journal for the Philosophy of Science*.

Pezzulo, G. (2014). Why do you fear the bogeyman? An embodied predictive coding model of perceptual inference. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(3), 902-911.

Raihani, N. J., & Bell, V. (2019). An evolutionary perspective on paranoia. *Nature human behaviour*, *3*(2), 114-121.

Rao, R. and Ballard, D. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), pp.79-87.

Reed, E. J., Uddenberg, S., Suthaharan, P., Mathys, C. D., Taylor, J. R., Groman, S. M., & Corlett, P. R. (2020). Paranoia as a deficit in non-social belief updating. *Elife*, *9*, e56345.

Schulz, S., Green, M., & Nelson, K. (2016). *Schizophrenia and psychotic spectrum disorders*. Oxford: Oxford University Press.

Schwartenbeck, P., FitzGerald, T. H., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M., & Friston, K. (2015). Optimal inference with suboptimal models: addiction and active Bayesian inference. *Medical hypotheses*, *84*(2), 109-117.

Shergill, S. S., Samson, G., Bays, P. M., Frith, C. D., & Wolpert, D. M. (2005). Evidence for sensory prediction deficits in schizophrenia. *American Journal of Psychiatry*, *162*(12), 2384-2386.

Shinn, A., Pfaff, D., Young, S., Lewandowski, K., Cohen, B., & Öngür, D. (2012). Auditory hallucinations in a cross-diagnostic sample of psychotic disorder patients: a descriptive, cross-sectional study. *Comprehensive Psychiatry*, *53*(6), 718-726.

Sterzer, P., Adams, R., Fletcher, P., Frith, C., Lawrie, S., & Muckli, L. et al. (2018). The Predictive Coding Account of Psychosis. *Biological Psychiatry*, *84*(9), 634-643.

Strauss, G., Waltz, J., & Gold, J. (2014). A Review of Reward Processing and Motivational Impairment in Schizophrenia. *Schizophrenia Bulletin*, *40*(Suppl 2), S107-S116.

Teufel, C., & Fletcher, P. C. (2016). The promises and pitfalls of applying computational models to neurological and psychiatric disorders. *Brain*, *139*(10), 2600-2608.

Tsuang, M. (2000). Schizophrenia: genes and environment. *Biological Psychiatry*, *47*(3), 210-220.

Ursini, G., Punzi, G., Chen, Q., Marenco, S., Robinson, J., & Porcelli, A. et al. (2018). Convergence of placenta biology and genetic risk for schizophrenia. *Nature Medicine*, *24*(6), 792-801.

Waters, F., Collerton, D., ffytche, D., Jardri, R., Pins, D., & Dudley, R. et al. (2014). Visual Hallucinations in the Psychosis Spectrum and Comparative Information From Neurodegenerative Disorders and Eye Disease. *Schizophrenia Bulletin*, *40*(Suppl_4), S233-S245.

Williams, D. (2018). Hierarchical Bayesian models of delusion. *Consciousness and Cognition, 61*, 129–147. doi:10.1016/j.concog.2018.03.003.

Williams, D. (2019). Socially adaptive belief. *Mind & Language.* 1– 22. https://doi.org/10.1111/mila.12294

Williams, D. (2020). Epistemic Irrationality in the Bayesian Brain. *The British Journal for the Philosophy of Science*, axz044, https://doi.org/10.1093/bjps/axz044