Counterfactual cognition and psychosis: Adding complexity to predictive processing accounts

Sofiia Rappe[1,2] and Sam Wilkinson[3]

[1] Faculty of Philosophy, Ludwig-Maximilians-Universität München, Germany
[2] Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Germany
[3] Department of Sociology, Philosophy, and Anthropology, University of Exeter, UK

Sofiia Rappe (corresponding author)
ORCID ID: 0000-0003-3343-7025
Email address: Sofiia.Rappe@campus.lmu.de
Phone: +4917674533826
Address of affiliation:
Faculty of Philosophy, Philosophy of Science, and the Study of Religion
Ludwig-Maximilians-Universität München
Ludwigstraße 31
80539 Munich, Germany

Sam Wilkinson
ORCID ID: **0000-0001-7414-7505**
Email address: S.Wilkinson@exeter.ac.uk
Address of affiliation:
Sociology, Philosophy and Anthropology
University of Exeter
Amory Building
Rennes Drive
Exeter
EX4 4RJ, United Kingdom

Disclosure of potential conflicts of interest: No potential conflict of interest was reported by the authors.

Counterfactual cognition and psychosis: Adding complexity to predictive processing accounts

Over the last decade or so, several researchers have considered the predictive processing framework (PPF) to be a useful perspective from which to shed some much-needed light on the mechanisms behind psychosis. Most approaches to psychosis within PPF come down to the idea of the "atypical" brain generating inaccurate hypotheses that the "typical" brain does not generate, either due to a systematic top-down processing bias or more general precision weighting breakdown. Strong at explaining common individual symptoms of psychosis, such approaches face some issues when we look at a more general clinical picture. In this paper, we propose an update on the current accounts of psychosis based on the realization that a neurotypical brain constantly generates non-actual, de-coupled, counterfactual hypotheses as part of healthy cognition. We suggest that what is going on in psychosis, at least in some cases, is not so much a generation of erroneous hypotheses, but rather an inability to correctly use the counterfactual ones. This updated view casts "accurate" cognition as more fragile and delicate, but also closes the gap between psychosis and typical cognition.

Keywords: psychosis; counterfactuals; delusions; hallucinations; predictive processing; reality monitoring

Article word count: 6778 Words

1. Introduction

Psychosis is a puzzling phenomenon. It involves having inaccurate and often strange beliefs and perceptual experiences. It is not clear, and certainly not from a pre-theoretic, un-scientific perspective, where this disconnection from consensus reality comes from, why it happens, what causes it. Over the last decade or so, several theorists have considered the predictive processing framework (PPF)[1] to be a useful perspective from which to shed some much-needed light on the mechanisms behind psychosis. We fully share this optimism. The PPF has a lot in its favour. It provides clear sources of potential problems for the functioning of a cognitive system (predictions, priors, and prediction errors with associated precision weightings). It also shows promise in tying neurobiology and neurochemistry (especially the role of neurotransmitters like dopamine, see, e.g., Corlett, Frith & Fletcher, 2009) to computational aspects of cognition, as well as dovetailing nicely with certain phenomenological features of experience (e.g., Ratcliffe, 2013, 2017).

Our aim in this paper is to update existing accounts of psychosis by helping ourselves to one recent innovation concerning neurotypical cognition, namely, the realisation that healthy, daily cognition is suffused with counterfactual hypotheses, and to apply it to thinking about psychosis in the PPF. To state our claim plainly: whereas standard accounts take psychosis to involve the "atypical" brain generating inaccurate hypotheses that the neurotypical brain does not generate, we explore the idea that the neurotypical brain is actually constantly generating inaccurate, de-coupled, counterfactual hypotheses, and that what is going on in psychosis (at least sometimes) is

[1] For an accessible introduction on predictive processing see Wiese and Metzinger (2018). For more detailed treatments see, e.g., Clark 2015 and Hohwy 2013.

more helpfully construed as an inability to distinguish the factual from the counterfactual hypotheses (or, perhaps more accurately, a failure to appropriately use the counterfactual hypotheses as they should be used). In other words, psychosis could sometimes be less about generating inaccurate hypotheses *de novo,* as existing PPF accounts have suggested, and more about wrongly identifying or using counterfactual hypotheses as such. Paradoxically, it is these counterfactual departures from reality, that give our experience of reality its *counterfactual depth*, which contribute to the sense of reality.

This updated view has a couple of important consequences. First, it paints a picture where the brain of an individual in a state of psychosis is more similar to the brain of the individual who is not in such a state, since human cognition is rife with unreal hypothesising. Second, it fits better with the phenomenology of psychosis, its subtlety and heterogeneity, and also in the way that hallucinations are not simply like normal perceptual experiences that happen to be inaccurate (and delusions are not simply like normal beliefs that happen to be false): they are phenomenologically more exotic and unfamiliar than that (Ratcliffe, 2017; Humpston & Broome, 2016).

We proceed as follows. We start by presenting existing predictive processing accounts of psychosis and discuss their virtues (section 2). Then we introduce the notion of *counterfactual depth* in theoretical work on cognition in general (section 3). We further explore the idea that psychosis can be understood in terms of failures to recognise or use counterfactual hypotheses as such and discuss four distinct types of breakdowns in the counterfactual depth and how they may produce the symptoms associated with psychosis (section 4). We argue that failures of reality monitoring result not only in the taking (by the brain) of inaccurate, counterfactual hypotheses to be accurate and factual, and hence to feature as the primary rather than auxiliary drivers of experience and belief (i.e. hallucinations and delusions), but also erode the structure of experience, because the auxiliary counterfactual hypotheses are no longer playing that role but also because such failure results in various subjective reality markers being misapplied. This fits nicely with the idea that what Ratcliffe calls "Real Hallucinations" are not simply like normal perceptual states that are inaccurate: they are new and unfamiliar kinds of states. It also fits with the uncanny, but pre-hallucinatory aspects of the psychosis prodrome: reality looks flat or strange (Ratcliffe, 2013), experience has a lost or altered counterfactual depth. We conclude by outlining some implications of our approach and future directions (section 5).

2. Existing predictive processing accounts of psychosis

The predictive processing framework (PPF) for thinking about cognition, perception, and action has recently gained a lot of attention in computational psychiatry, with PPF-based models being proposed in relation to anxiety (Chekroud, 2015), depression (Barrett, Quigley & Hamilton, 2016; Stephan et al., 2016), PTSD (Wilkinson, Dodgson & Meares, 2017), autism (Pellicano & Burr, 2012; Van de Cruys et al., 2014; Lawson et al., 2014), schizophrenia (Adams et al., 2013; Horga et al., 2014; Fletcher & Frith, 2009), and general accounts of emotion (Seth, 2013; Miller & Clark, 2018; Smith, Parr & Friston, 2019).

The popularity of the PPF in psychiatric research does not come as a surprise; the framework, even in its most basic form, provides at least three distinct sources of potential problems for the functioning of a cognitive system (namely, predictions, priors, and prediction errors with associated precision weightings), setting a clear direction for addressing the first of the two main questions of computational psychiatry:

Given a specific model of the mind, what could possibly go wrong (Huys, Maia & Frank, 2016; Ford et al., 2014)? On the other hand, problems with the above elements of the framework map well onto various psychopathologies, providing simple potential answers to the second question: How can a specific disorder be explained using the model at hand (Wilkinson, 2014)? Furthermore, the PPF is increasingly conceived by its proponents as a general paradigm about brain functioning. Hence, any explanation of psychopathology in PPF terms automatically fits within a much broader research program.

Such symptoms of psychosis as delusions and hallucinations are probably the most explored psychopathological phenomena within the PPF. An early account of psychosis in predictive processing can be found in Fletcher and Frith's paper (2009), which sketched a hierarchically arranged prediction error minimizing architecture. The main point of this model was to show how one basic mechanism can account for both delusion-like (belief-like) phenomena and hallucination-like (experience-like) phenomena. According to Fletcher and Frith, both hallucinations and delusions present erroneously selected winning hypotheses, due to excessive prediction error signalling, while the difference between the two types of phenomena is a question of degree determined by where in the hierarchy the mis-selection occurs. The "higher up" the hypotheses are, the more belief-like they are (e.g., delusions); the "lower down" they are, the more experience-like they are (e.g., visual hallucinations or voice-hearing) (see, e.g., Adams et al., 2013; Horga et al., 2014; Fletcher & Frith, 2009). In other words, hallucinations and delusions are treated as a problem of inference with a single cause — associating too little precision with sensory information (/too much with predictions), which results in the selection of a "wrong" hypothesis about the world. Depending on where in the system erroneous selection occurs, it may be experienced, for example, as voice-hearing, visual hallucinations, delusions, etc.

This idea that psychosis is the result of excessive prediction error in the system or, to similar results, excessive precision-weighting on prediction errors is common to many accounts of psychosis (see Sterzer et al., 2018 for a nice review of PPF and excessive prediction error in psychosis) and has some empirical support. For example, there is evidence that certain conscious effects that are explained in terms of prediction error minimization are experienced less or differently in people with diagnoses of schizophrenia (see, e.g., the hollow mask illusion (Dima et al., 2009)). Other evidence that supports this hypothesis comes from eye tracking data, namely, impaired tracking during visual occlusion (Hong, Avila & Thacker, 2005), impaired repetition learning (Avila et al., 2006), and “paradoxical improvement” (Adams et al., 2013), where clinical populations are better at responding to sudden changes of direction in visual tracking targets. All of this points to a general over-reliance on bottom-up prediction error, and less reliance on the top-down predictions. Furthermore, these approaches provide plausible stories about why delusions and hallucinations co-occur (similar bottom-down processing “bias”) and why delusions can arise both due to biological factors and life events, since both can impact on predictive processing mechanisms (Wilkinson, 2014).

Having said this, they have some outstanding issues (see Sterzer et al., 2018 for a full review). First, delusions and hallucinations co-occur to varying degrees in different cases of psychosis, and not as often as the traditional approach might suggest. Second, the persistence of delusions in psychosis may be tricky to accommodate by precision estimation breakdowns, especially the kind that result in a bottom-up processing bias. It is a defining feature of delusions that they persist despite contradicting evidence. This suggests an excessive influence of delusional beliefs on the perception of new information, which would entail an increased precision of delusion-related priors (in direct contradiction with the excessive error signalling proposal). Additionally, given that the hypotheses are updated on many conditionally independent levels simultaneously, arriving at (and sustaining) a drastically wrong hypothesis about the world requires a fairly large breakdown in precision estimation machinery. It is a common assumption in PPF that different parts of the generative model are deeply inter-connected and integrated. This, of course, does not imply unconstrained holism: the network reflects the inferred causal structure of the world, which imposes certain constraints.[2] Still, a specific story must be told about how these constraints would prevent an individual suffering a radical systematic precision estimation breakdown from losing the ability to function in the world at all, as selecting a hypothesis leading to persistent delusions and hallucinations could possibly affect the rest of the system.

Third, this approach to psychosis is not always consistent with the PPF-based explanations of other kinds of symptoms that may co-occur with it. For example, many features associated with autistic perception, are often attributed to an imbalance of precision ascribed to sensory evidence relative to prior beliefs towards bottom-up processing (that is, having too much precision on sensory evidence) (see, e.g, Pellicano & Burr, 2012; Van de Cruys et al., 2014). This paradigm in autism has some empirical support, such as increased visual cortical activation and decreased prefrontal activation in participants with autism (Lee et al., 2007; Manjaly et al., 2007). This is consistent with the increased bottom-up visual processing, which corresponds to precision weighting skewed towards sensory signal. Yet, although autistic behaviour sometimes

[2] We thank one of the anonymous referees for clarifying this point.

can co-occur with psychosis, this is rather unusual (Larson et al., 2017).

Finally, when it comes to hallucinations, an alternative view, in which hallucinations occur due to enhanced rather than weakened top-down predictive signaling, has also been proposed (Friston, 2005; Corlett et al., 2019). This approach suggests that perception would rely less on the sensory input and more on the prior beliefs, a claim even more problematic for explaining the occasional co-existence of psychotic and autistic symptoms in one individual. To complicate matters, some evidence indirectly supports this reverse approach to hallucinations. For example, it was shown that people who hear voices are more susceptible to conditioning-induced hallucinations (Powers et al., 2017) and that hallucinations in schizophrenia patients correlate with a top-down perceptual bias in auditory tasks (Cassidy et al., 2018).

What is relevant is that all the accounts discussed above take psychosis to involve the generation of hypotheses that are inaccurate portrayals of the world, hypotheses that do not feature in the brains of those who aren't in states of psychosis. This core commonality across several different accounts of psychosis is precisely what we would like to question to update the PPF treatment of psychosis. Whereas standard accounts take psychosis to involve the "atypical" brain generating inaccurate hypotheses that the neurotypical brain does not generate, we explore the idea that the neurotypical brain actually constantly generates inaccurate, de-coupled, counterfactual hypotheses, and that this is an integral part of the rich tapestry of healthy cognition. This is consistent with the evidence that psychosis-like experiences are much more common than we typically think (McGrath et al., 2015). So, perhaps, at least sometimes, what matters is how the mind *treats* and *uses* such "inaccurate" hypotheses and how the treatment differs in the pathological cases. In the next section, we argue that generation of inaccurate hypotheses is indeed a crucial feature of our cognition as it heavily relies on counterfactuals everywhere from the lowest levels of perception to intentional, conscious reasoning.

3. Predictive processing, counterfactual depth, and offline cognition

The focus of our paper is on the difference between cognition in psychosis and neurotypical cognition. In this section we discuss the latter. So, putting psychosis to one side, we would like to draw attention to the developments in predictive processing that go beyond the earliest versions of PPF and, specifically, to one recent feature, namely, the emphasis on the rich “counterfactual depth” of the generative models (Seth, 2014; Wilkinson, 2020, 2021).

In the literature related to PPF, the word “counterfactual” is mostly used not in the strict linguistic sense but in relation to the capacity to form hypotheses about the non-factual, about past and present possibilities, as well as about other possible (or even impossible) worlds. In such cases, the notion of counterfactual hypotheses often also refers simply to the hypotheses that present alternative possibilities that are mutually exclusive. This means that at least some of them do not correspond to the actual state of affairs (they are *counter to the facts*), although which ones are such may not be known from an agent’s perspective (see, e.g, Clark, Friston, and Wilkinson, 2019). In other cases (see, e.g., Corcoran, Pezzulo, and Hohwy 2020), the hypotheses may pertain to the states the agent could possibly find herself in *if* she were to act in a certain way. Philosophers and cognitive scientists have emphasized the importance of such (broadly construed) counterfactuals in human cognition long before the PPF. Counterfactual reasoning is thought to be central to planning, decision-making, and intentional, goaloriented behaviour more broadly (see Byrne, 2016). When it comes to predictive architectures specifically, counterfactual reasoning may be what underlies the human ability to learn priors for decision making in the absence of direct feedback (Zylberberg et al., 2018). Generation of counterfactual alternatives may also be implicated in imagination. Here, a common idea is that imagination has emerged as a way of predicting consequences of anticipated possible actions (Burr & Jones, 2016; Friston et al., 2012; Seth, 2014). "Since many actions will be mutually exclusive, many such representations will inevitably be about merely fictional circumstances, representing possible sensory consequences of actions that never occur" (Jones & Wilkinson 2020). [3] Corcoran and colleagues (2020) further propose a counterfactual active inference model that allows an agent to evaluate a variety of different contexts before settling on a specific action by reflecting on previous actions ('retrospective' inference) or imagining possible future scenarios ('prospective' inference). Although their discussion is shaped through the lens of the free energy principle (see., e.g., Friston 2009, 2010) not touched upon in this paper, similar observations hold for the standard predictive processing formulation.

Less intuitively, but consistent with the perception-cognition continuity in PPF, rich counterfactuality may be not only a property of our conscious reasoning, decision-making, and imagination, but lower-level processes, such as those that generate perception and perceptual phenomenology. For example, Seth (2014) (building on Noe's (2006) notion of sensorimotor contingencies) argues that when the subject is engaged in "factual", actual world-directed experience, that experience is the way it is,

[3] As Jones and Wilkinson 2020 note, however, deliberate imaginative acts are a very specific form of personal-level counterfactual cognition, and the imaginative capacity is not fully exhausted by the ability to generate counterfactual alternatives.

has “perceptual presence” (Noe, 2006), in part due to the activation of a range of counterfactual predictions within the generative model. Indeed, Seth claims, a lack of such counterfactual underpinning of the generative model leads to some atypical perceptual phenomena such as synaesthesia, which notably lack this presence, and hence do not feel real. The same could be also said about after-images. If you have looked directly into a bright light bulb, when you look elsewhere, you may see an after-image, but the fact that it occurs in the same patch of your visual field wherever you look is one of several things that tells you that it is not a real thing in the world, but an anomalous product of your visual system.

Wilkinson (2020) builds on Seth’s account and draws on observations from virtual reality research (e.g., Meehan et al., 2003), arguing that different kinds of counterfactual predictions account for different aspects of perceptual experiences. In particular, “object-active” predictions are crucial for perceiving objects as having volumetric content, whereas “agent-active” predictions are involved in the subjective experience of presence associated with perceiving the world, of taking what one perceives to be real and present to us, and of us as present in the world.

In other words, it seems very plausible to suggest that “healthy” cognition involves far more, and constant, counterfactual hypothesising across the generative model, which stands in contrast with the previous more minimal versions of predictive processing, which emphasize processing efficiency and consistency of winning hypotheses across the generative model. Instead, on the updated view, our experience of the world is underpinned by *counterfactual depth*. The fact that we experience a real world at all (e.g. a real apple, in front of me), and that we experience parts of it in the way we do (e.g. as a real apple rather than a fake one) is grounded in a suite of counterfactual hypotheses that need never be tested or actualised (e.g. concerning what would happen if we were to bite into it, and the expectation that we could, even if we never in fact end up doing so).

When it comes to explaining psychosis, this realization tempts us to shift the focus from generation of inaccurate hypotheses by an "atypical" brain, toward the idea that the neurotypical brain is constantly generating inaccurate, de-coupled, counterfactual hypotheses, and that what is going on in psychosis (at least sometimes) is an inability to distinguish the factual from the counterfactual hypotheses. The question then is how and why such misidentification or mismanagement occurs. We argue that the counterfactual richness of the predictive mind requires additional mechanisms of monitoring, both when it comes to entertaining the alternative scenarios about how the world could be but also how the world could have been (but is not). Such mechanisms of monitoring may present an added source of vulnerability in the predictive brain that is not fully captured by precision weighting imbalances.

Clark, Friston, and Wilkinson (2019) argue that the commitment of PPF to the deep architecture of the generative models and latent variables ("hidden causes") allows for counterfactual reasoning to naturally arise in predictive agents. Specifically, they propose that one's posterior beliefs at intermediate levels of the hierarchical generative model may be estimated as highly certain in a way that leaves room for them to be paired with multiple potentially applicable higher-level hypotheses. For example, the hypothesis "it is snowing" may be compatible with both "the snow is H2O" and "the snow is synthetic". Being possible alternatives, these two hypotheses do not have the same degree of certainty as "it is snowing", and this is precisely what allows one to have doubts about their experience and entertain alternative causal scenarios.

However, such openness to causal alternatives on its own does not provide the mechanism of further counterfactual exploration, which may require simulation –

redistribution of precision weighting in a way that allows the system to treat the counterfactual parts of the model as if these hypotheses were indeed selected as the winning ones in order to generate rich counterfactual spaces. The case for simulation becomes even stronger for the task of generating counterfactuals **in the strict linguistic sense**, that is the kind of hypotheses explicitly incompatible with the current state of the system but rather related to how the world could have been but is not, and also for the kind of retroactive and prospective inference discussed by Corcoran, Pezzulo, and Hohwy (2020). Such hypothesizing presents an example of what Hoerl and McCormack (2019) call *temporal reasoning*. As they note, connectionist architectures generally have problems with the tasks that require temporal reasoning because they do not explicitly represent change, but rather simply update representations as new information comes along (Hoerl & McCormack 2019). In the absence of explicit representation of the past states of the system, direct simulation (Knight & Grabowecky, 1995; Corcoran, Pezzulo, and Hohwy, 2020) becomes a very likely candidate for the mechanism of counterfactual exploration. Here, by simulation we do not mean offline model updating that merely disregards sensory input (the feature often taken to be among those differentiating between cognition and perception). Rather, we suggest that, in the processes of both perception and cognition, the system is exploring, "trying on" different generative sub-models by altering the relevant weights. Such "role-play" requires keeping track of the parts of the model that pertain to the actual, as opposed to some possible, world.[4]

[4] There is, of course, an important difference between *agent*'s free, conscious exploration of alternative hypotheses (conscious voluntary cognition) and constant counterfactual-model generation by the *entire cognitive system,* for example such as those implicated in perception. In this paper we deliberately do not discuss agency and intentionality. The

In the literature, the process of differentiating between what is real from what is not is often referred to as "reality monitoring". There is no consensus as to which variables contribute to reality monitoring or what features reality monitoring specifically tracks. As we discuss in more detail in section 5.1, reality monitoring is typically understood as monitoring the source of one's experience (is it dependent on external stimuli or is it entirely self-generated?) – a task especially important for accounts such as predictive processing in which perception is a constructive process that has a significant top down component . Such monitoring is often taken to be metacognitive, that is, a kind of second-order process integrating multiple types of information and manifesting as a sense (or feeling) of reality (see, e.g., Dokic & Martin, 2017). However, as argued by Deroy and Rappe (under review), not all processes related to reality monitoring are necessarily aimed towards establishing the source of one's experience. For example, the experiences such as those in derealization or virtual reality are recognized by the agent as perceptual (coming from outside of the agent) but are characterized by distinct subjective reality signatures (namely, are experienced as not being "quite" real). When it comes to counterfactual reasoning, the focus shifts from monitoring the source of experiences (hypotheses) to monitoring the "actuality" status of the relevant generative (sub)model; that is, rather than caring about the distinction between perception and cognition/imagination (from world vs. from me), we care about the distinction between actual vs. counterfactual world models (which include not only here-and-now perceptual experiences, but general knowledge about the world). This distinction between "actuality" and perceptual reality monitoring, at least theoretically, leads to two different monitoring goals, although they may be accomplished by largely

assumption is that some form of tracking is required for both conscious and intentional, and unconscious and involuntary simulation-based counterfactual exploration.

overlapping mechanisms. For example, metacognitive reality monitoring may by default fulfil the role of tracking the actual world model, while various subjective markers of reality may ground the agent in the perceptual experience during the exploration of counterfactual scenarios. Here, we do not aim to provide the specific mechanisms of reality (or actuality) monitoring for counterfactual exploration (this is a whole research field in its own right), yet, following Deroy and Rappe (under review) we take the subjective experience of reality to be a composite that includes both categorical mechanisms of reality monitoring (is this real or not?) and a variety of qualitative, gradual, subjective signatures of reality that accompany different “non-imaginary” experiences. We argue that disturbances in the actual-model monitoring (as well as the subjective signature of reality) in some cases may play a role in explaining the sources and symptoms of psychosis.

4. Breakdowns in counterfactually rich models

Assuming the counterfactual richness of the generative models in PPF and its relevance to psychosis, the next step is to specify in more detail the different ways in which such rich counterfactuality may break down and lead to the symptoms such as delusions, illusions and hallucinations, derealisation, and the “uncanniness” of experience in psychosis.[5] As a starting point, we identify four possible ways in which

[5] Importantly, there are a lot of cases where an agent may “misperceive” without anything being broken in the system. For example, simple auditory illusions like misrecognizing the sound of fresh snow under one’s foot as a bird chirp can be easily explained by selecting a wrong prediction on the basis of one’s priors. Such erroneous selection does not signify any systematic problems; the bird noise may be simply more expected (although not corresponding to what is really going on). Other cases of misperception in healthy

such a breakdown could occur. The list is not necessarily exhaustive, and some options may work better than others and be better fits for different cases.

(1) Actuality monitoring breakdown: misidentifying the counterfactual parts of the model as pertaining to the actual world;
(2) Disconnectedness from (lack of access to) certain parts of the richly counterfactual model;
(3) Poor counterfactual underpinning: inability to generate enough alternative hypotheses for sufficient counterfactual depth;
(4) Perceptual reality grounding problems: abnormality in the subjective markers of reality.

Below we elaborate on these four options. Importantly, however, our proposal is not meant to substitute the more traditional precision estimation-based approaches to psychosis in PPF, rather it is complementary, providing *additional* possible sources of disorder. There is an increasing understanding in psychiatry that superficially similar symptoms may have distinct pathophysiological mechanisms and that diagnosis based simply on a cluster of symptoms may mistakenly group "heterogeneous syndromes with different pathophysiological mechanisms into one disorder" (Wong et al., 2010), which

population populations, however, at least seemingly go beyond simple misperception. They are not rooted in perceptual signals and indicate that there has been a failure of (or rather "incorrect") integration of information somewhere at a higher level of inference that includes cognitive beliefs. A prime example of such a situation is the Third Man Syndrome. However, the phenomenon typically occurs in the situations of high stress where all kinds of one-time abnormalities seem plausible. Here, we ignore such one-time cases and focus only on the mechanisms of hallucinations, delusions and derealisation in psychosis.

could ultimately result in a low efficiency of the administered treatment in individual cases and mismanagement of cognitive and financial resources when it comes to developing new treatments more broadly (Ford et al., 2014). Exhaustively mapping out the space of the possible pathophysiological mechanisms then becomes crucial for accurate diagnosis and treatment. We see our proposal as contributing towards this goal from the perspective of a specific theoretical framework, the PPF.

Option 1: Actuality monitoring breakdown.

The first option corresponds to the situation in which rich counterfactual models are generated but wrong hypotheses are selected "as real" because the parts of the model that pertain to the counterfactual alternatives to the current state of affairs are misidentified as pertaining to the actual world. Depending on how the actual-world monitoring mechanisms are conceived, this case may closely resemble some of the more traditional, precision-estimation (PE) accounts of psychosis.

According to the PE accounts, a self-generated stimulus may be mistaken for a stimulus caused by an external source due to a malfunction in precision estimation (Griffin & Fletcher, 2017). The consequences of self-generated stimuli are expected to be easy for the system to predict (the predictions have high precision), and so when they aren't well predicted, they generate levels of prediction error akin to external stimuli, and hence are experienced as such. This would lead to, for example, inner speech being experienced as having an external source, leading to auditory verbal hallucinations (voice-hearing) (see, e.g., Jones & Fernyhough, 2007). This is how self-monitoring accounts of psychosis (Frith 1992) are accommodated as a special case within the PPF (Wilkinson, 2014). There is, in effect, too much prediction error generated by self-produced stimuli, and hence the self-produced stimuli are erroneously deemed to not be self-produced.

Similar effects may also be achieved, however, not by the general dysregulation resulting in wrongly accommodating error signals across the system, but by misidentifying a part of the generative model related to a counterfactual scenario as pertaining to the actual world. This may manifest, for example, as delusions or hallucinatory experiences (again, depending on where in the hierarchy such mis-selection occurred). Here, however, the problem arises not as a global tendency towards top-down processing or a random precision estimation breakdown, but specifically in relation to the treatment of counterfactuals. This provides advantages over the PE dysregulation accounts, at least when it comes to accommodating certain cases. For example, hallucinations and delusions are not necessarily expected to co-occur because misidentification is now limited to very specific counterfactual sub-models. For the same reason, simultaneous occurrence of hallucinations and autistic behavior/decreased susceptibility to visual illusions in one individual are no longer in theoretical conflict: the hallucinatory/delusional part of one's experience is explained by the actuality monitoring breakdown, rather than a general top-down bias in the system. Hence, it is no longer incompatible with the bottom-up processing bias assumed to take place in individuals diagnosed with autism spectrum disorder (ASD) (Pellicano & Burr, 2012).[6] This is compatible with the reported co-occurrence of ASD and psychosis in

[6] A bottom-up processing bias, in fact, may serve as a compensatory mechanism in the situation where reality monitoring is somehow unreliable. Giving more "voice" to prediction errors is a helpful (although, perhaps, more effortful) strategy to keep grounded in the current, real word model when the monitoring mechanisms are less reliable.

individuals, and with the observation that the manifestation of psychosis in individuals with ASD is somewhat atypical (Larson et al., 2017).

Finally, if we treat actuality monitoring as (at least in part) a metacognitive process, this case also aligns with the metacognitive accounts of psychosis, supported by evidence that the patients diagnosed with schizophrenia often show certain metacognitive deficiencies (Cella, 2015, Lysaker et al., 2011, 2014).

Option 2: Loss of access.

The second option relates to a problem with accessing the right parts of the model. If the ability to navigate the entirety of the generative model is somewhat impaired, the agent may be stuck in certain possible world interpretations (subparts of the generative model) that would result in persistent hallucinatory and/or delusional experiences. The idea here is that, although a rich counterfactual model is generated, some of the alternative hypotheses are blocked from being selected and relied upon in further processing. This may include the alternatives corresponding to the real state of affairs, forcing the system to operate on a set of inevitable "false" options. This could explain the inability of an affected individual to properly process evidence against the model and re-assess. Indeed, delusion seem to be very resistant to evidence, even if such evidence is judged completely trustworthy (Wilkinson, 2015). Further, like with the previous option, because such a disconnect could theoretically occur at any specific part of the model, it would be rather natural, again, to assume that delusions and hallucinations would not always come together. There are no general top-down biases involved, only the impaired ability to break out of certain (counterfactual) "frames". Furthermore, an impairment in the ability to navigate the entirety of the generative model is directly associated with the problems integrating information from multiple parts of the model.

This could provide an alternative explanation to the observed decreased proportion of integrative “solutions” to the McGurk effect in populations with psychosis compared to controls (White et al., 2014) (as opposed to the standard explanation within the PPF that there is too much prediction error).

Beyond the waking (yet altered) states, such as those in psychosis, the substantial loss of access to subparts of the generative model is characteristic of dreaming. Impaired connectivity and limited access to episodic memory are indeed established features of REM sleep and may explain the subjectively “real” feeling of the dream environments (even though they can be deeply bizarre in content). Reality monitoring simply does not have the correct targets as viable options for applying itself. If the access is partially restored, however, different cues and monitoring processes may pick up on this, leading, for example, to (partial) lucidity, even if the dream is experienced as highly immersive. It also accounts for other interesting features of the strange phenomenology of dream content: often the counterfactual depth that tells us who an individual is, may clash with the surface imagery. In other words, you are convinced in your dream (indeed you never question) that someone is a certain individual, even though they look nothing like them. Consider this dream report: “I had a talk with your colleague, but she looked differently, much younger, like someone I went to school with, perhaps a 13-year-old girl” (Schwartz & Maquet, 2002, p. 26). Or this one: “I recognize A’s sister ... I am surprised by her beard, she looks much more like a man than a woman, with a big nose” (Schwartz & Maquet, 2002, p. 29). As Wilkinson (2015) notes, this bears significant similarity to delusional misidentification, which often occurs in association with first-episode psychotic disorders (Jocic, 1992; Salvatore et al., 2014; Gupta et al., 2021). Of course, another cause of delusional misidentification, and, one might suppose, of loss of access to relevant counterfactuals underpinning the generative model, is localised brain damage. Here the delusional individual may admit that the person they perceive looks just like a loved one but is not experienced this way because of a lack of the relevant counterfactuals (e.g., the individual is not experienced as huggable or as expected to behave in certain familiar ways).

Option 3: Poor counterfactual underpinning.

Third, there is also a possibility that not enough counterfactual alternatives are generated in the presence of the correct winning hypothesis. On the perceptual level, this could lead to insufficient counterfactual depth (see Seth, 2014). This would manifest itself experientially in things seeming flat, unreal, lacking in depth. This in turn might lead to delusions, for example, delusional misidentification of a loved one (e.g., they might be an android). On higher cognitive levels this could lead to impaired counterfactual reasoning and generation of hypothetical scenarios. In fact, counterfactual thinking is often impaired in patients diagnosed with schizophrenia precisely in the decreased ability to generate counterfactual scenarios (Hooker, Roese & Park 2000; Albacete et al., 2017). For example, Albacete and colleagues (2017) found that, although patients with schizophrenia do not differ from controls in their ability to identify an event most relevant for reversing a given scenario (their causal thinking is intact), they generate significantly fewer spontaneous alternative and counterfactual scenarios, especially in cases of spatial and temporal “nearly happened” events.

Interestingly, this is something that cannot be easily attributed to the traditional precision estimation breakdown/excessive prediction error accounts.[7]

Option 4: Subjective markers error.

Fourth and finally, a problem could arise with the subjective markers of reality. This could happen both due to the problem with reality monitoring, or independently, for individual subjective markers. The alteration in the subjective signature of reality, in either case could lead to the experience of derealization. For example, the latter case could correspond to derealization in healthy individuals (which are rather common, see, e.g., Aderibigbe et al., 2001) induced, for example, by sensory deprivation (Reed & Sedman,1964), extreme stress (Bernat et al., 1998) or drug/alcohol abuse (Melges et al., 1974). The former case, on the other hand, could be the cause of derealization commonly observed as an early symptom in patients with psychosis (Giersch & Mishara, 2017).

If we treat some cases of psychosis as resulting from counterfactual navigation impairment, it makes sense that one of the first symptoms of impaired actuality monitoring may be alteration of the reality signature of perception. Further, such alteration may on its own over time lead to precision redistribution in the generative model (without necessitating any consistent biases or precision estimation malfunction), resulting, for example, in different interpretation of the incoming sensory

[7] As one of the reviewers pointed out, our treatment presents a strong deficit account in which the relevant counterfactual hypotheses are **absent**. Yet, another possibility would be that these hypotheses are in fact still generated but **lack in precision**, which in principle, could give rise to similar consequences.

information/related higher-level causal inferences and, consecutively, perceptual hallucinations and delusions.

This may offer one explanation why hallucinations sometimes occur with or without accompanying sense-of-reality changes and can be judged by the suffering individual as either real or unreal. Depending on the individual's own differences in processing, including precision weightings assigned to certain parts of the model and specific types of evidence, as well as general proclivity towards more top-down/bottom-up processing, a malfunction in the sense of reality/actuality monitoring system may have stronger or weaker effect on the evaluation of the categorical reality status and the content of one's perception and vice versa. Although mutual reliance of the cognitive judgement, reality (meta-)monitoring, and the subjective signatures of reality on each other may occasionally lead to hypotheses selection errors, the functional redundancy in these processes is generally a helpful, rather than a hindering feature. Partial functional overlap among various types of evaluating the ontological nature of various submodels may make an individual less responsive to the processing errors in each subsystem making complex cognitive processing significantly more robust.

5. Consequences

5.1 Counterfactual depth and reality monitoring

We end up with a view that looks rather like the influential "reality monitoring" accounts introduced in the 1980's (Bentall & Slade, 1985). However, there are some important and illustrative differences that come with the counterfactually rich PPF interpretation. First, "reality monitoring" is task based: it gestures towards a task that an individual can do badly or well. We are delving beneath the success or failure of correctly ascertaining reality, to the mechanisms thanks to which this is possible.

Second, and most interestingly perhaps, reality monitoring was thought to be a subtype of "source monitoring". Source monitoring is a notion borrowed from memory research, where the source could be defined, for example, as "the spatial, temporal, and contextual characteristics of an event as well as the sensory modalities through which it was perceived" (Vinogradov et al. 1997, p.1530). In a classic review of source monitoring (in general, not in the context of psychosis), Johnson et al. (1993) claim that the term "source monitoring" subsumes at least three distinct abilities.

(1) *Internal source monitoring* - distinguishing one's real actions (verbal and bodily) from merely imagined ones.
(2) *External source monitoring* - distinguishing between outer sources (e.g. one third party from another).
(3) *Reality monitoring* - distinguishing between self-generated and outer events.

Bentall and Slade (1985) hypothesised that source monitoring could help to explain psychosis, and especially the third category of reality monitoring. In other words, the hypothesis at the centre of reality monitoring accounts is that people with schizophrenia/psychosis are bad at distinguishing between self-generated and outer events. Not only that, but they have a bias in a particular, externalising, direction: they have a *general propensity* to mistake self-produced events for external events. For example, a self-produced piece of imagery, either in the form of inner speech or episodic memory, is misattributed to an outside source. The basic idea is that if you misattribute something self-generated to the world (the non-self), then you will take fantasy (that which you made up), to be reality (that which is constrained by fact, by actuality). This basic logic is also what is behind a similar (but importantly different) approach to psychosis, namely, comparator-based self-monitoring (Frith, 1992).

This is very different to our understanding of reality monitoring. First of all, much of that which is self-generated is perfectly real. Since, according to the PPF, the world's contribution is so sparse and noisy, we have to construct our reality, albeit in a constrained manner. Furthermore, much of our cognition – both online and (more obviously) offline – is about inferring what is the case. In an important respect these inferential processes are self-produced, but they aren't by that same token inaccurate fantasies. Conversely, external elements are very much capable of leading us astray, either because we draw inferences in directions we ought not to, or because we are genuinely misled through no fault of our own cognition. Stated most generally, then, we differ both from (task-referencing) reality monitoring and (mechanism-referencing) self-monitoring by insisting that the equations between "from me" and "not real", and "not from me" and "real", do not hold. Ultimately, then, on our view, distinguishing reality from non-reality is not about recognising source: firstly, because it is more heterogeneous than that, but also because the relevant processes function at a lower level than experience: they help to generate the experience as the experience that it is, rather than characterising the response to the experience. In other words, these processes will be baked into the experience, rather than a judgement we make based on the experience. This seems very much in keeping with the phenomenological complexity of psychosis, and its "location" within phenomenology. It is not like psychosis involves strange judgements based on relatively normal experiences: it involves alterations to experience (Parnas & Henriksen, 2016; Giersch & Mishara, 2017; Berkovitch et al., 2021).

5.2 Summary and implications

Taking the PPF as a starting point for thinking about psychosis has been a fruitful approach. Here we suggest that the innovation of counterfactual depth in the PPF should be similarly extended to our thinking about psychosis. We do not intend this as a critique of the more straightforward view, but as a potential addition to them that may help to account for a wider array of the many things that fall under the category "psychosis".

Having said this, it does cast psychosis (or at least some forms of it) in a different light. It is no longer primarily cast as adopting a radically inaccurate hypothesis but rather as a subtle anomaly in something we all have, namely, mechanisms for distinguishing the actual, from the non-actual. This makes "accurate" cognition seem more fragile and delicate, but also closes the gap between psychosis and typical cognition. In other words, there is less difference between the brains of individuals in states of psychosis and the brains of those who are not in such states.

Our counterfactually enriched PPF account also allows for very heterogeneous forms of psychosis, and a richer understanding of its experience, beyond the presence of straightforward delusions and hallucinations, and into quasi-perceptual/uncanny/unreal etc. (see Ratcliffe, 2017). Sometimes theorists talk about psychosis as if, against a backdrop of otherwise normal experience, a voice is heard, or strange beliefs emerge, but the clinical and experiential reality of psychosis is often one of varied, pervasive, subtle, and unfamiliar changes to the basic fabric of experience.

5.3 Future directions

Our contribution is as speculative as it is modest. As we have clarified, we are not criticising existing views, but rather pointing in other un-explored directions. These speculations need to be tested and fleshed out through careful observation and experimentation. This requires a holistic, joined-up approach that examines everything from the neural and neurobiological underpinnings of counterfactual depth within the PPF, whether this be through imaging techniques, drug models (how might some drugs, for example, flatten counterfactual depth in ways that mimic certain form of psychotic experience?), etc. up to careful phenomenological investigation.

Another future direction is very straightforward. The counterfactually embellished PPF can be applied beyond psychosis, towards other conditions that have been given a more traditional PPF treatment, such as post-traumatic stress disorder (Wilkinson, Dodgson & Meares, 2017). It is worth noting that accounts within the related "Free Energy" framework have already cast depression in terms of changes to the experience of possibilities (Kiverstein, Miller & Rietveld, 2020), and to us this added subtlety seems very much in the right direction. Most generally, the appreciation that experience is not simply about perceiving sensory qualities of the here-and-now (hearing sounds, seeing colours and shapes), but experiencing a subtle patchwork of possibilities.

References

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

Aderibigbe, Y. A., Bloch, R. M., & Walker, W. R. (2001). Prevalence of depersonalization and derealization experiences in a rural population. *Social Psychiatry and Psychiatric Epidemiology*, *36*(2), 63-69.

Albacete, A., Contreras, F., Bosque, C., Gilabert, E., Albiach, Á., & Menchón, J. M. (2017). Symptomatic Remission and Counterfactual Reasoning in Schizophrenia. *Frontiers in psychology*, *7*, 2048. https://doi.org/10.3389/fpsyg.2016.02048

Avila, M. T., Hong, L. E., Moates, A., Turano, K. A., & Thaker, G. K. (2006). Role of anticipation in schizophrenia-related pursuit initiation deficits. *Journal of neurophysiology*, *95*(2), 593-601.

Bentall, R. P., & Slade, P. D. (1985). Reality testing and auditory hallucinations: a signal detection analysis. *The British journal of clinical psychology*, *24 (Pt 3)*, 159–169. https://doi.org/10.1111/j.2044-8260.1985.tb01331.x

Berkovitch, L., Charles, L., Del Cul, A., Hamdani, N., Delavest, M., Sarrazin, S., ... & Houenou, J. (2021). Disruption of conscious access in psychosis is associated with altered structural brain connectivity. *Journal of Neuroscience*, *41*(3), 513-523.

Bernat, J. A., Ronfeldt, H. M., Calhoun, K. S., & Arias, I. (1998). Prevalence of traumatic events and peritraumatic predictors of posttraumatic stress symptoms in a nonclinical sample of college students. *Journal of traumatic stress*, *11*(4), 645–664. https://doi.org/10.1023/A:1024485130934

Burr, C., and Jones, M. (2016). The Body as Laboratory: Prediction-Error Minimization, Embodiment, and Representation. *Philosophical Psychology*, 29(4), 586–600.

Byrne, R. M. (2016). Counterfactual thought. *Annual review of psychology*, *67*, 135-157.

Cassidy, C. M., Balsam, P. D., Weinstein, J. J., Rosengard, R. J., Slifstein, M., Daw, N. D., ... & Horga, G. (2018). A perceptual inference mechanism for hallucinations linked to striatal dopamine. *Current Biology*, *28*(4), 503-514.

Cella, M., Reeder, C., & Wykes, T. (2015). Lessons learnt? The importance of metacognition and its implications for Cognitive Remediation in schizophrenia. *Frontiers in Psychology*, *6*, 1259.

Chekroud, A. M. (2015). Unifying treatments for depression: an application of the Free Energy Principle. *Frontiers in Psychology*, *6*, 153.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing qualia: Consciousness as inference, not raw datum. *Journal of Consciousness Studies*, *26*(9-10), 19-33.

Corcoran, A. W., Pezzulo, G., & Hohwy, J. (2020). From allostatic agents to counterfactual cognisers: active inference, biological regulation, and the origins of cognition. *Biology & Philosophy*, *35*(3), 1-45.

Corlett, P. R., Frith, C. D., & Fletcher, P. C. (2009). From drugs to deprivation: a Bayesian framework for understanding models of psychosis. *Psychopharmacology*, *206*(4), 515–530.

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers III, A. R. (2019). Hallucinations and strong priors. *Trends in cognitive sciences*, *23*(2), 114-127.

Deroy, O., & Rappe S. (2021). *The clear and not so clear signatures of perceptual reality in the Bayesian brain.* [Manuscript submitted for publication].

Dima, D., Roiser, J. P., Dietrich, D. E., Bonnemann, C., Lanfermann, H., Emrich, H. M., & Dillo, W. (2009). Understanding why patients with schizophrenia do not perceive the hollow-mask illusion using dynamic causal modelling. *Neuroimage*, *46*(4), 1180-1186.

Dokic, J., & Martin, J. R. (2017). Felt reality and the opacity of perception. *Topoi*, *36*(2), 299-309.

Fletcher, P. C. & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience,* 10(1), 48.

Ford, J. M., Morris, S. E., Hoffman, R. E., Sommer, I., Waters, F., McCarthy-Jones, S., ... & Cuthbert, B. N. (2014). Studying hallucinations within the NIMH RDoC framework. *Schizophrenia bulletin*, *40*(Suppl_4), S295-S304.

Friston, K. J. (2005). Hallucinations and perceptual inference. *Behavioral and Brain Sciences*, *28*(6), 764-766.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, *13*(7), 293-301

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

Friston, K., Adams, R., Perrinet, L., and Breakspear, M. (2012). Perceptions as Hypotheses: Saccades as Experiments. *Frontiers in Psychology*, 3, 151.

Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia*. Psychology press.

Giersch, A., & Mishara, A. L. (2017). Is schizophrenia a disorder of consciousness? Experimental and phenomenological support for anomalous unconscious processing. *Frontiers in psychology*, *8*, 1659.

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

Gupta, M., Gupta, N., Zubiar, F., & Ramar, D. (2021). Delusional misidentification syndromes: untangling clinical quandary with the newer evidence-based approaches. *Cureus*, *13*(12).

Hoerl, C., & McCormack, T. (2019). Thinking in and about time: A dual systems perspective on temporal cognition. *Behavioral and Brain Sciences*, *42*.

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

Hong, L. E., Avila, M. T., & Thaker, G. K. (2005) Response to unexpected target changes during sustained visual tracking in schizophrenic patients. *Experimental Brain Research, 165*, 125–131.

Hooker, C., Roese, N. J., & Park, S. (2000). Impoverished counterfactual thinking is associated with schizophrenia. *Psychiatry*, *63*(4), 326-335.

Horga, G., Schatz, K. C., Abi-Dargham, A. & Peterson, B. S. (2014). Deficits in Predictive Coding Underlie Hallucinations in Schizophrenia. *The Journal of Neuroscience, 34*(24), 8072– 8082.

Humpston, C. S. & Broome, M. R. (2016). The Spectra of Soundless Voices and Audible Thoughts: Towards an Integrative Model of Auditory Verbal Hallucinations and Thought Insertion. *Review of Philosophy and Psychology* 7 (3):611-629.

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, *19*(3), 404-413.

Jocic, M. D. (1992). Delusional misidentification syndromes. *Jefferson Journal of Psychiatry*, *10*(1), 4.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological bulletin*, *114*(1), 3.

Jones, S.R. & Fernyhough, C. (2007) Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations, *Consciousness and Cognition*, **16** (2), pp. 391–399.

Jones, M., & Wilkinson, S. (2020). From Prediction to Imagination. In A. Abraham (Ed.), *The Cambridge Handbook of the Imagination* (Cambridge Handbooks in Psychology, pp. 94-110). Cambridge: Cambridge University Press.

Kiverstein, J., Miller, M., & Rietveld, E. (2020). How mood tunes prediction: a neurophenomenological account of mood and its disturbance in major depression. *Neuroscience of Consciousness*, *2020*(1), niaa003.

Knight, R. T., & Grabowecky, M. (1995). Escape from linear time: prefrontal cortex and conscious experience.

Larson, F. V., Wagner, A. P., Jones, P. B., Tantam, D., Lai, M. C., Baron-Cohen, S., & Holland, A. J. (2017). Psychosis in autism: comparison of the features of both conditions in a dually affected cohort. *The British Journal of Psychiatry 210*(4), 269–275.

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

Lee, P. S., Foss-Feig, J., Henderson, J. G., Kenworthy, L. E., Gilotty, L., Gaillard, W. D., & Vaidya, C. J. (2007). Atypical neural substrates of Embedded Figures Task performance in children with Autism Spectrum Disorder. *Neuroimage*, *38*(1), 184-193.

Lysaker, P. H., Erickson, M., Ringer, J., Buck, K. D., Semerari, A., Carcione, A., & Dimaggio, G. (2011). Metacognition in schizophrenia: The relationship of mastery to coping, insight, self-esteem, social anxiety, and various facets of neurocognition. *British Journal of Clinical Psychology*, *50*(4), 412-424.

Lysaker, P. H., Leonhardt, B. L., Pijnenborg, M., van Donkersgoed, R., de Jong, S., & Dimaggio, G. (2014). Metacognition in schizophrenia spectrum disorders: methods of assessment and associations with neurocognition, symptoms, cognitive style and function. *Isr J Psychiatry Relat Sci*, *51*(1), 54-62.

McGrath, J. J., Saha, S., Al-Hamzawi, A., Alonso, J., Bromet, E. J., Bruffaerts, R., ... & Kessler, R. C. (2015). Psychotic experiences in the general population: a cross-national analysis based on 31 261 respondents from 18 countries. *JAMA psychiatry*, *72*(7), 697-705.

Manjaly, Z. M., Bruning, N., Neufang, S., Stephan, K. E., Brieber, S., Marshall, J. C., ... & Fink, G. R. (2007). Neurophysiological correlates of relatively enhanced local visual search in autistic adolescents. *Neuroimage*, *35*(1), 283-291.

Meehan, M., Razzaque, S., Whitton, M. C., & Brooks, F. P., Jr. (2003). Effect of latency on presence in stressful virtual environments. *Proceedings of the IEEE Virtual Reality, 2003*, 141–148.

Melges, F. T., Tinklenberg, J. R., Deardorff, C. M., Davies, N. H., Anderson, R. E., & Owen, C. A. (1974). Temporal disorganization and delusional-like ideation: Processes induced by hashish and alcohol. *Archives of general psychiatry*, *30*(6), 855-861.

Miller, M., & Clark, A. (2018). Happily entangled: prediction, emotion, and the embodied mind. *Synthese*, *195*(6), 2559–2575.

Parnas, J., & Henriksen, M. G. (2016). Mysticism and schizophrenia: A phenomenological exploration of the structure of consciousness in the schizophrenia spectrum disorders. *Consciousness and Cognition*, *43*, 75-88.

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends in cognitive sciences*, *16*(10), 504-510.

Powers, A. R., Mathys, C., & Corlett, P. R. (2017). Pavlovian conditioning–induced hallucinations result from overweighting of perceptual priors. *Science*, *357*(6351), 596-600.

Ratcliffe, M. (2013). Phenomenology, Naturalism and the Sense of Reality. *Royal Institute of Philosophy Supplement* 72:67-88.

Ratcliffe, M. (2017). *Real Hallucinations: psychiatric illness, intentionality, and the interpersonal world*. Cambridge, MA, USA: MIT Press.

Reed, G. F., & Sedman, G. (1964). Personality and depersonalization under sensory deprivation conditions. *Perceptual and Motor Skills*, *18*(2), 659-660.

Salvatore, P., Bhuvaneswar, C., Tohen, M., Khalsa, H. M. K., Maggini, C., & Baldessarini, R. J. (2014). Capgras' syndrome in first-episode psychotic disorders. *Psychopathology*, *47*(4), 261-269.

Sass, L. A., & Parnas, J. (2007). Explaining schizophrenia: the relevance of phenomenology. *Reconceiving schizophrenia*, 63-95.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, *17*(11), 565–573.

Seth A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive neuroscience*, *5*(2), 97–118.

Smith, R., Parr, T., & Friston, K. J. (2019). Simulating emotions: An active inference model of emotional state inference and emotion concept learning. *Frontiers in Psychology*, *10*, 2844.

Stephan, K. E., Manjaly, Z. M., Mathys, C. D., Weber, L. A., Paliwal, S., Gard, T., ... & Petzschner, F. H. (2016). Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in human neuroscience*, *10*, 550.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., & Corlett, P. R. (2018). The Predictive Coding Account of Psychosis. *Biological psychiatry*, *84*(9), 634–643.

Van de Cruys, S., Evers, K., Van der Hallen, R., Van Eylen, L., Boets, B., de-Wit, L., & Wagemans, J. (2014). Precise minds in uncertain worlds: Predictive coding in autism. *Psychological Review*, *121*(4), 649.

Vinogradov, S., Willis-Shore, J., Poole, J. H., Marten, E., Ober, B. A., & Shenaut, G. K. (1997). Clinical and neurocognitive aspects of source monitoring errors in schizophrenia. *American Journal of Psychiatry*, *154*(11), 1530-1537.

White, T. P., Wigton, R. L., Joyce, D. W., Bobin, T., Ferragamo, C., Wasim, N., ... & Shergill, S. S. (2014). Eluding the illusion? Schizophrenia, dopamine and the McGurk effect. *Frontiers in Human Neuroscience*, *8*, 565.

Wiese, W., & Metzinger, T. (2017). Vanilla PP for philosophers: A primer on predictive processing.

Wilkinson S. (2014). Accounting for the phenomenology and varieties of auditory verbal hallucination within a predictive processing framework. *Consciousness and cognition*, *30*, 142–155.

Wilkinson, S. (2015). Delusions, dreams, and the nature of identification. *Philosophical Psychology*, *28*(2), 203-226.

Wilkinson, S. (2020). Distinguishing volumetric content from perceptual presence within a predictive processing framework. *Phenomenology and the Cognitive Sciences*, *19*(4), 791-800.

Wilkinson, S. (2021) What Can Predictive Processing Tell us about the Contents of Perceptual Experience? in *Purpose and Procedure in the Philosophy of Perception* (Logue and Richardson eds.) Oxford University Press.

Wilkinson, S., Dodgson, G., & Meares, K. (2017). Predictive processing and the varieties of psychological trauma. *Frontiers in Psychology*, *8*, 1840.

Wong, E. H., Yocca, F., Smith, M. A., & Lee, C. M. (2010). Challenges and opportunities for drug discovery in psychiatric disorders: the drug hunters' perspective. *International Journal of Neuropsychopharmacology*, *13*(9), 1269-1284.

Zylberberg, A., Wolpert, D. M., & Shadlen, M. N. (2018). Counterfactual reasoning underlies the learning of priors in decision making. *Neuron*, *99*(5), 1083-1097.