is a strikingly unproductive way to work with delusions (11). Rather, the therapy has always involved understanding the grounds for the person's belief – the unusual experiences and events underpinning it – while validating and empathizing with emotional distress; and exploring with the patient, collaboratively, alternative possibilities, cognitive, emotional and behavioural, in the light of the person's history and social environment.

More recently, informed by the more precise empirical evidence of mechanisms of specific symptom persistence, whether of reasoning biases, negative affect or negative self-concept, the CBT approach is to work with the process (if you like, with the mode or manner of the thinking rather than the content) and thereby to alleviate the delusion and its accompanying distress and impact on everyday life (2).

## References

1. Sass L, Byrom G. Phenomenological and neurocognitive perspectives on delusions: a critical overview. World Psychiatry 2015;14:164-73.
2. Freeman D, Garety P. Advances in understanding and treating persecutory delusions: a review. Soc Psychiatry Psychiatr Epidemiol 2014;49:1179-89.
3. Garety PA, Gittins M, Jolley S et al. Differences in cognitive and emotional processes between persecutory and grandiose delusions. Schizophr Bull 2013;39:629-39.
4. Iyassu R, Jolley S, Bebbington P et al. Psychological characteristics of religious delusions. Soc Psychiatry Psychiatr Epidemiol 2014;49:1051-61.
5. Garety PA, Freeman D. The past and future of delusions research: from the inexplicable to the treatable. Br J Psychiatry 2013;203:327-33.
6. Kahneman D. Thinking, fast and slow. New York: Farrar, Strauss, Giroux, 2011.
7. Garety P, Waller H, Emsley R et al. Cognitive mechanisms of change in delusions: an experimental investigation targeting reasoning to effect change in paranoia. Schizophr Bull 2015;41:400-10.
8. Fowler D, Hodgekins J, Garety P et al. Negative cognition, depressed mood and paranoia: a longitudinal pathway analysis using structural equation modelling. Schizophr Bull 2012;38:1063-73.
9. Freeman D, Dunn G, Murray RM et al. How cannabis causes paranoia: using the intravenous administration of Δ9-tetrahydrocannabinol (THC) to identify key cognitive mechanisms leading to paranoia. Schizophr Bull 2015;41:391-9.
10. Wickham S, Taylor P, Shevlin M et al. The impact of social deprivation on paranoia, hallucinations, mania and depression: the role of discrimination social support, stress and trust. PLoS One 2014; 9:e105140.
11. Milton F, Patwa VK, Hafer RJ. Confrontation vs belief modification in persistently deluded patients. Br J Med Psychol 1978;51:127-30.

DOI 10.1002/wps.20210

# Answering some phenomenal challenges to the prediction error model of delusions

**Philip R. Corlett**

Department of Psychiatry, Yale University, New Haven, CT, USA

Delusions are a challenge; a menagerie of odd beliefs with a diverse set of differential diagnoses and candidate pathologies (1). Their phenomenology has led many to deem them un-understandable (2). On the other hand, their susceptibility to treatment with D2 receptor antagonists has led many clinicians and scientists to consider them understood.

Delusions are neither fully understood nor un-understandable. 20-50% of patients have residual delusions even after adequate D2 blockade (3). A recent model challenges the un-understandability conclusion, suggesting a bridging hypothesis that unites neural and experiential aspects of delusions through computational theory (1). That hypothesis involves prediction errors (PEs) – the mismatches between expectation and experience that guide learning, attention, and belief formation and maintenance. If PEs are signaled inappropriately, delusions result (1).

Sass and Byrom (4) highlight some phenomenological challenges to this explanation. Here, I meet those challenges. I will argue that the aberrant PE model can indeed account for some of the more puzzling aspects of delusions, for example the central role of self-experience in delusions, the curious double bookkeeping in which patients may engage, the role of hyposalience (the bizarre as banal), the "anything goes" inferences made by many people with delusions, as well as bizarre delusions that appear to defy understanding.

Sass and Byrom also speculate that the brain default mode network (DMN) may mediate these latter phenomena. I will join them in this speculation, but I will argue that the DMN too is PE driven (5). As such, I will maintain that PE is still a single factor explanation of delusions, even the most bizarre ones.

Sass and Byrom question whether aberrant PE can explain the centrality of self-experience to delusions, as well as some of the contents of delusions related to inflated self-concept or metaphysics. There is a nascent field examining self-representation in the brain (indeed, this circuitry often overlaps with the DMN). I believe we can conceive of healthy self-experience and ipseity disturbance in the context of PE theory.

PE theory posits that the brain builds hierarchical models to predict the causes of its sensory data (6). Any mismatch between prediction and data can have two consequences: a) it is ignored or overridden by prior beliefs (as is the case with optical illusions), or b) it is transmitted up the hierarchy where it updates the top-down prior with new learning (so expectation is different in the future) (6).

The first-person self is perceived as a result of the same hierarchical modeling process. Ultimately it encodes the evidence for (or belief in) the existence of the self in the world – when all is intact, we model ourselves as agents in our world that can act on our environment and, through acting, change the sensory feedback we receive (7). Under this account, ipseity arises when the agent identifies with its model of the world (8). Aberrant PE would lead to the ipseity disturbances that Sass and Byrom outline through a disruption of this self-modeling process. For example, a surprising lack of self-agency experience, due to a deficit in predicting one's intentions, could lead to passivity delusions (1). According to this account, self-experiences, beliefs and delusions arise as the best explanation for the available data incident upon the organism. This explanation overrides other potential explanations in a winner-take-all manner (1).

Sass and Byrom go on to highlight a phenomenon that challenges this winner-take-all notion: double bookkeeping. Here, patients with delusions lack manifest conviction in their beliefs, e.g., claiming that their food is being poisoned, but eating it nevertheless. It seems that people do not always act on their delusions and that they may simultaneously endorse and deny them (9).

A phenomenon from animal conditioning, extinction learning, might be relevant to double bookkeeping (1). Extinction involves new learning, for example to no longer expect reward or electric shock in a previously reinforced situation. There is a transition from expecting a salient event, to no longer expecting it. Patients recovering from their delusions describe a similar duality of belief and disbelief regarding their delusions (10). Under an extinction account of recovery from delusions, new learning (of a non-delusional belief) competes with and overrides the original reinforced situation (the delusion) (1). Extinction learning (of a new belief) is driven by appropriate PE: when the expected event fails to transpire, a negative PE triggers updating of future expectancies (1). Likewise, the relationship between endorsing and rejecting the delusion is modulated by PE; if a surprising salient event occurs (perhaps one that is reminiscent of the delusion), the old belief may be renewed (1). More broadly, in the face of constant aberrant PE (either in terms of magnitude, timing or precision), new belief formation is necessary. If PE signals remain variable, inconsistent, and difficult to accommodate, it is possible that a new causal model is required – that is a new set of causal associations, a new mechanism that might pertain. Thus PE guides the exploration of the space of possible explanatory beliefs (11) until this PE "over beliefs" is minimized by the adoption of a new higher order causal belief (1). Under constant aberrant PE, one can imagine switching back and forth between delusional and non-delusional interpretations (or double bookkeeping) (1).

Relatedly, Sass and Byrom posit that, for patients with delusions, the bizarre can become banal. Indeed, the PE model appreciates and can account for this. In the face of persistent aberrant PE, patients may learn a hyper-prior – *a prior over priors* – that anything is possible, even the surprising experiences and associations on which their delusions are based (11). This hyper-prior, that the world is always surprising, renders subsequent PE experiences expected – unsurprising, banal. This is a potential explanation for delusion maintenance, double bookkeeping (with respect to manifest conviction) and negative symptoms – if goal-directed actions have proven repeatedly ineffectual, why engage in actions at all (12), and if all beliefs lack explanatory adequacy, why bother acting on them or updating them?

Sass and Byrom suggest that PE may account better for non-bizarre delusions (particularly what they call the paranoid type). I suggest instead that the PE model best explains aberrant salience and delusions of reference. However, our empirical data – linking delusion severity to aberrant PE using functional neuroimaging – were gathered from patients with a range of delusion contents (13). They also claim that bizarre delusions (e.g., "I am the right foot of Christ") are more problematic for PE theory. Here, I point to the overlap between causal belief formation, associative learning and propositional cognition; causal representation may involve linguistic expressions like metaphors (14). Bizarre delusions then represent inappropriate use of metaphor in an attempt to establish some intersubjective meaning, albeit futile.

During the formative delusional mood, the world becomes ineffable. Prodromal patients use relative terms (similes) to describe their experiences: "It is *as if* people are actors, walking down the street wearing masks" (15). As these experiences persist, the relative terms subside (people *are* wearing masks, they *are* in disguise); the simile becomes a metaphor as the delusion develops and the metaphor becomes a top-down prior around which perception and cognition are organized.

Delusional priors form as the best way to account for a noisy and uncertain PE. But if they don't accommodate this PE, the PE will eventually be disregarded and won't update the prior (16). However, the prior will be engaged with, reactivated and therefore strengthened (1). Similar so-called *backfire effects* have been observed with political beliefs (17). They relate to the process of memory reconsolidation, through which memories are reactivated, updated and consolidated once more (1). Aberrant PE may drive inappropriate reactivation, rumination upon and strengthening of delusional priors (1). This ruminative engagement with delusional priors may also be a mechanism through which simile becomes metaphor in bizarre delusions.

Such rumination engages autobiographical memory (perhaps a contributor to ipseity) and the DMN circuitry (18). Sass and Byrom point out that delusion severity has been related to the inappropriate DMN engagement, perhaps as a result of its unconstrained operation in the absence of control from dorsolateral prefrontal cortex (DLPFC) (19). They argue that DMN responses have been related to self-processing and so the DMN represents a neural locus for ipseity and its disturbance in individuals with delusions. I urge caution in ascribing any function, particularly one as multifaceted as self-processing, to one set of regions.

This problem is further compounded by the challenges of inferring the precise function of DMN (since, by definition, it is engaged when subjects are disengaged, any inferences about its function cannot be corroborated by behavioral data). However, let's assume that DMN is involved in autobiographical processing (i.e., it contributes to ipseity) and that its activity is usually anti-correlated with DLPFC (19). My work has shown that, during causal belief formation, the DLPFC signals an explanatory gap – or PE (1). Others have suggested that the DMN may generate a narrative (possibly autobiographical) to explain such PEs (5).

It is possible then that ipseity may be perturbed through defective engagement of DLPFC, DMN or a faulty interaction between them. Future work should identify the relationship between PE and explanation, DLPFC and DMN function. Their usual anti-correlation is disrupted in psychotic states (19), but the specifics of their interaction in the genesis of experiences, beliefs and delusions is deserving of further scrutiny. Nevertheless, it is possible to explain the relationship between DMN and delusions in the PE framework.

Finally, Sass and Byrom express concern that belief formation in the PE model involves a conscious deliberative process that is cold and logical. This is not the case. All but the highest levels of the hierarchical generative model on which minimal self and delusion formation are based are unavailable to conscious awareness (8). We are not conscious of processing at lower levels of the hierarchy (below narrative self and first person perspective) and so beliefs and delusions form outside of conscious awareness.

Admittedly, our investigations so far have lacked an affective component (note, however, our work on the role of distress in PE signaling and delusion-like ideation (20)). A recent review of the dys-interaction between cognition and emotion in studies of schizophrenia highlighted the role of affect in generating aberrant salience. Across studies, neural and behavioral responses to affectively salient events were attenuated and neutral events garnered excessive affectivity – often these responses correlated with delusion severity (21).

In summary, Sass and Byrom highlight the importance of phenomenological data in generating an explanation of delusions. The PE theory likewise focuses on relating the experiences that characterize delusions to their underlying brain mechanisms (1). The devil though is in the details, and I trust that, by elaborating some PE theory details, I have answered some of the challenges leveled by Sass and Byrom.

It seems that ipseity and PE models are broadly consilient, but concerned with different levels of explanation (22). It is important that we capitalize on this consilience rather than focusing on prioritizing one level of explanation over the others.

## References

1. Corlett PR, Taylor JR, Wang XJ et al. Toward a neurobiology of delusions. Prog Neurobiol 2010;92:345-69.
2. Jaspers K. General psychopathology. Manchester: Manchester University Press, 1963.
3. Lindenmayer JP. Treatment refractory schizophrenia. Psychiatr Q 2000;71:373-84.
4. Sass L, Byrom G. Phenomenological and neurocognitive perspectives on delusions: a critical overview. World Psychiatry 2015;14:164-73.
5. Carhart-Harris RL, Friston KJ. The default-mode, ego-functions and free-energy: a neurobiological account of Freudian ideas. Brain 2010;133(Pt. 4):1265-83.
6. Friston K. A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 2005; 360:815-36.
7. Limanowski J, Blankenburg F. Minimal self-models and the free energy principle. Front Hum Neurosci 2013;7:547.
8. Blanke O, Metzinger T. Full-body illusions and minimal phenomenal selfhood. Trends Cogn Sci 2009;13:7-13.
9. Bleuler E. Die Prognose der Dementia praecox (Schizophreniegruppe). Allgemeine Zeitschrift für Psychiatrie 1908;65: 436-64.
10. Stanton B, David A. First-person accounts of delusions. Psychiatr Bull 2000;24:333-6.
11. FitzGerald TH, Dolan RJ, Friston KJ. Model averaging, optimal inference, and habit formation. Front Hum Neurosci 2014;8:457.
12. Fletcher PC, Frith CD. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat Rev Neurosci 2009;10:48-58.
13. Corlett PR, Murray GK, Honey GD et al. Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions. Brain 2007;130(Pt. 9):2387-400.
14. Mitchell CJ, De Houwer J, Lovibond PF. The propositional nature of human associative learning. Behav Brain Sci 2009; 32:183-98.
15. Gross G, Huber G. Sensory disorders in schizophrenia. Arch Psychiatr Nervenkr 1972;216:119-30.
16. Preuschoff K, Bossaerts P. Adding prediction risk to the theory of reward learning. Ann NY Acad Sci 2007;1104:135-46.
17. Bullock JG. Partisan bias and the Bayesian ideal in the study of public opinion. J Politics 2009;71:1109-24.
18. Spreng RN, Mar RA, Kim AS. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. J Cogn Neurosci 2009;21:489-510.
19. Whitfield-Gabrieli S, Ford JM. Default mode network activity and connectivity in psychopathology. Annu Rev Clin Psychol 2012;8:49-76.
20. Corlett PR, Fletcher PC. The neurobiology of schizotypy: fronto-striatal prediction error signal correlates with delusion-like beliefs in healthy people. Neuropsychologia 2012;50:3612-20.
21. Anticevic A, Corlett PR. Cognition-emotion dysinteraction in schizophrenia. Front Psychol 2012;3:392.
22. Marr D, Poggio T. From understanding computation to understanding neural circuitry. Neurosci Res Prog Bull 1977;204: 301-28.

DOI 10.1002/wps.20211