# Explaining Delusional Beliefs: A Hybrid Model

Kengo Miyazono1\* and Ryan Thomas McKay2

1 Graduate School of Integrated Arts and Sciences, Hiroshima University, 1-7-1 Kagamiyama, Higashi-Hiroshima City Hiroshima, Japan 739-8521; +81-82-424-6306; miyazono@hiroshima-u.ac.jp

2 Department of Psychology, Royal Holloway, University of London, Egham, Surrey TW20 0EX, United Kingdom; +44 (0) 1784 41 4993; Ryan.McKay@rhul.ac.uk

\* Corresponding author

# Explaining Delusional Beliefs: A Hybrid Model

## *Introduction*. In this paper we present and defend a hybrid theory of the development of delusions that incorporates the central ideas of two influential (yet sometimes bitterly opposing) theoretical approaches to delusions – the two-factor theory and the prediction error theory.

## *Method*. After introducing the central ideas of the two-factor theory and the prediction error theory, we describe the motivations for our conciliatory project, explain the theoretical details of the hybrid theory we propose, and answer potential objections to our proposal.

## *Results*. According to the hybrid theory we advance, the first factor of a delusion is physically grounded in an abnormal prediction error, and the second factor is physically grounded in the overestimation of the precision of the abnormal prediction error. Against anticipated objections, we argue that the hybrid theory is internally coherent, and that it constitutes a genuine hybrid between the two-factor theory and the prediction error theory.

## *Conclusion*. A rapprochement between the two-factor theory and the prediction error theory is both possible and desirable. In particular, our hybrid theory provides a parsimonious and unified account of delusions, whether monothematic or polythematic, across a wide variety of medical conditions.

## Key words: Delusions; Belief formation; Two-factor theory; Prediction error theory; Cognitive neuropsychiatry.

## Introduction

A delusion is a belief that is held despite obvious counterevidence and that is not explained by the person's social, cultural or religious background. Delusions feature in an array of psychiatric and neurological conditions including schizophrenia, dementia, and acquired brain injury. Delusions are typically (but not necessarily) false, and their content can vary greatly, although certain recurrent themes are recognised by theorists and clinicians. The delusion of persecution, for example, involves the idea of being harmed or harassed by an individual or group. The Capgras delusion involves the idea that a familiar individual has been replaced by an imposter, and so on.

A range of “local” theories have been put forward to explain particular delusions, such as the Capgras delusion (e.g., Ellis & Young, 1990), the delusion of persecution (e.g., Bentall, Kinderman, & Kaney, 1994; Freeman, Garety, Kuipers, Fowler, & Bebbington, 2002) and the delusion of control (e.g., Blakemore, Wolpert, & Frith, 2002). But what are the prospects for a “global” theory that explains many (if not all) delusions? (Note that a global theory might not rule out local theories. For example, a global theory might leave out some details that are filled in by local theories.) There are at least two influential candidates for a global theory in the recent literature: the two-factor theory, developed in the field of cognitive neuropsychiatry, and the prediction error theory, which comes from the area of computational psychiatry. In brief, according to the two-factor theory, delusions are explained by two distinct neurocognitive factors with different explanatory roles. According to the prediction error theory, delusions are explained by the disrupted processing of prediction errors (i.e., mismatches between expectations and actual inputs).

While the two-factor theory and the prediction error theory are often construed as rival accounts (Corlett, 2019; Corlett, Taylor, Wang, Fletcher, & Krystal, 2010; Fletcher & Frith, 2009; Hohwy, 2010; Williams, 2018), we have previously argued that the two theories might not be irreconcilable alternatives (McKay, 2012; Miyazono, 2019; Miyazono, Bortolotti, & Broome, 2014; see also Coltheart, 2010; Corlett, Honey, & Fletcher, 2016). In this paper we advance a hybrid theory, arguing that key contributions of the two theories can be combined in a powerful way. In developing our hybrid theory, we follow Coltheart (2010) who points out that “the first factor in most delusions is a violation of expectation [= prediction error]” and proposes to “make clearer what constitutes the second factor in the [prediction error] approach” (p. 25).1 The core idea of our hybrid theory is that the distinction between the first factor and the second factor in the two-factor framework corresponds to a crucial distinction in the prediction error framework, namely, the distinction between prediction errors and their estimated precision.

We begin by describing the central ideas of the two-factor theory and the prediction error theory. Then, in the main part of this paper, we describe the motivations for our conciliatory project, explain the theoretical details of the hybrid theory, and answer potential objections to the theory.

## Two theories

### *The two-factor theory*

The two-factor theory (Coltheart, 2007; Coltheart, Menzies, & Sutton, 2010; Coltheart, Langdon, & McKay, 2011; Davies, Coltheart, Langdon, & Breen, 2001; McKay, 2012) primarily aims to explain monothematic delusions (i.e., delusions concerning a single theme) that can arise in the context of neuropsychological deficits (but, for two-factor theoretic discussions of delusions in schizophrenia, see Coltheart, 2013; Coltheart, Langdon, and McKay, 2007).

The Capgras delusion is commonly used to illustrate the basic idea of the two-factor theory (indeed, Corlett [2019] has called Capgras delusion “the flagship of the two-factor explanatory fleet”). It is likely that a reduced autonomic response to familiar faces (Brighetti, Bonifacci, Borlimi & Ottaviani, 2007; Ellis, Young, Quayle, and De Pauw, 1997; Hirstein & Ramachandran, 1997) is relevant to the formation of the Capgras delusion. Ex hypothesi, this reduced autonomic activity constitutes an abnormal datum2 concerning a familiar face, and the Capgras delusion is formed in response to this datum, either as an explanation (Ellis & Young, 1990) or an endorsement (Bayne & Pacherie, 2004) of it.

Two-factor theorists accept that the abnormal datum concerning a familiar face is a causal factor in the formation of the Capgras delusion but claim that it is causally insufficient; an additional factor is needed. The main reason for this claim comes from the observation that the reduced autonomic response to familiar faces can be dissociated from the Capgras delusion. For example, just like Capgras patients, patients with ventromedial prefrontal cortex (vmPFC) damage exhibit reduced autonomic responses to familiar faces (Tranel, Damasio, & Damasio, 1995), but they typically do not become delusional about the identity of familiar people. The dissociation between the reduced autonomic response and the Capgras delusion can be explained by the two-factor theoretic hypothesis that the abnormal datum concerning a familiar face is causally insufficient for the formation of the delusion (for similar arguments concerning other monothematic delusions, see Coltheart, 2007 or Coltheart et al., 2011).

The fundamental claim of the two-factor theory is that there are two explanatory factors in the delusion formation process, “the first factor” and “the second factor”. These factors are supposed to play different explanatory roles; the first factor explains the content of a delusion**,** while the second factor explains either the adoption of a delusional hypothesis (McKay, 2012) or the persistence of an adopted delusional hypothesis (Coltheart et al., 2010). For example, the first factor of the Capgras delusion (i.e., the abnormal datum concerning a familiar face) is putatively responsible for the content of the Capgras delusion (i.e., the content that a familiar individual has been replaced by an imposter; see Corlett, 2019, for a critique of this view). The second factor (about which we will say more later on) is putatively responsible for the adoption or persistence of the Capgras hypothesis.

### *The prediction error theory*

The prediction error theory (Adams, Stephan, Brown, Frith, & Friston, 2013; Corlett et al., 2010; Corlett et al., 2016; Fletcher & Frith, 2009; Frith & Friston, 2012; Sterzer et al., 2018) is another influential theory of delusions, in particular, delusions in schizophrenia, which are often polythematic (but, for prediction error theoretic discussions of monothematic delusions such as the Capgras delusion, see Corlett et al., 2010).

The prediction error theory is based upon an increasingly popular account of the brain and its functions, sometimes called “predictive processing theory” or “predictive coding theory” (Clark, 2013; Friston, 2010; Hohwy, 2013), which has the following core commitments.

The brain does not simply reconstruct internal representations of the world from sensory inputs. Rather it predicts sensory inputs on the basis of the representations it already has, and compares the predicted and the actual sensory inputs. When the predicted input does not match the actual input, the brain tries to minimise the mismatch, for example, by revising the representations from which the predictions are derived. This prediction error minimisation is a kind of Bayesian inferential process in which internal representations in the brain are updated based on the prior, predicted inputs and the new, actual inputs. Prediction errors are processed at many different levels in the hierarchy within the brain, where different levels deal with information with different degrees of abstractness. Processes at different levels interact with each other in such a way that the prediction error at a given level serves as the bottom-up input for processing at the higher-level and the feedback from a level serves as the top-down expectation for processing at the lower-level.

Belief updating, in this framework, is a process of updating internal representations at a high level in the hierarchy in accordance with the principle of prediction error minimisation. The crucial insight of the prediction error theory is that delusions are the result of dysfunction in this process. In short, delusions are regarded as the product of abnormalities in processing prediction errors.

 A remarkable study by Corlett and colleagues (2007; see also Corlett et al., 2006; Gradin et al., 2011; Murray et al., 2008) suggests abnormal activities of right prefrontal cortex (which had been identified as a reliable marker of prediction error processing in previous studies) among patients with a diagnosis of first-episode psychosis (c.f., Corlett & Fletcher 2015; Griffiths, Langdon, Le Pelley, & Coltheart 2014). As Corlett and colleagues point out, this finding is “consistent with the possibility that in psychotic patients, prediction errors are signalled inappropriately and those errors maladaptively update the prefrontal representation of the world with irrelevant information” (Corlett et al., 2007, p. 2396).

## A hybrid theory

***Why hybrid?***

In this section, we present a hybrid theory that incorporates central ideas of the two-factor theory and the prediction error theory. There are good reasons to seek a rapprochement between the two theories. For instance, the two-factor theory (as a general framework) tends to be rather agnostic about mechanistic details. By adopting some ideas from the prediction error theory camp, we might achieve a better understanding of the nature (and neurophysiological cause) of the second factor (Coltheart 2010). Conversely, by adopting some ideas from the two-factor theory camp, we might better understand how alleged abnormalities in processing prediction errors manifest themselves at the psychological level of description.

However, one might worry that the idea of a hybrid theory is hopeless because of the theoretical incompatibility between the two-factor theory and the prediction error theory. Certainly there are important differences and disagreements between the two-factor theory camp and the prediction error theory camp (chief among these is the fact that the prediction error theory is often presented as a *one*-factor theory; “We posit a single factor, prediction error dysfunction for delusion formation and maintenance”; Corlett et al., 2010, p. 361). But these differences and disagreements should not be exaggerated. For instance, as Corlett and colleagues (2016) admit, there is nothing incoherent in the two-factor version of the prediction error theory; “[f]actor one (altered experience) could be specified lower in the hierarchy, and factor two (altered belief evaluation) higher up” (p. 1148) (for more on the one-factor/two-factor distinction, see Miyazono et al., 2014). We return to the problem of incompatibility later when we discuss potential objections to our hybrid proposal.

 Before moving on, we would like to mention a recent study by Darby, Laganiere, Pascual-Leone, Prasad and Fox (2016), which is coherent with the idea of a hybrid theory. Their work suggests that the association between delusions and the processing of prediction errors might not be limited to delusions arising in the context of schizophrenia. Darby and colleagues conducted a series of meta-analyses to identify brain regions activated by violations (versus confirmations) of participant expectations. A separate resting-state connectivity analysis revealed that the right frontal regions implicated in these meta-analyses (e.g., right ventral frontal cortex) were functionally connected to the lesion locations in two cohorts of lesion-induced delusional patients: patients with delusions of misidentification (e.g., the Capgras delusion) and patients with other delusions (e.g., the delusion of persecution). The logic of Darby and colleagues’ procedure (termed “lesion network mapping”) is that neuropsychiatric symptoms stem not just from lesion locations, but also from brain regions functionally connected to lesion locations. Their results are thus consistent with the hypothesis that delusions stem from the disrupted processing of expectation violations (=prediction errors).

Interestingly, Darby and colleagues’ findings are also consistent with the two-factor theory, insofar as they support the existence of two dissociable factors in the process of delusion formation. While expectation violation regions were functionally connected to the lesion locations in both of Darby and colleagues’ cohorts of lesion-induced delusional patients, a second region (left retrosplenial cortex, identified by a separate meta-analysis of studies on face/place familiarity as the region most activated by personally familiar [versus unfamiliar] stimuli) was functionally connected to the lesion locations in patients with delusions of misidentification but *not* in patients with other delusions. Thus, Darby and colleagues’ findings implicate disruption to two distinct networks in the genesis of delusions of misidentification: a familiarity network (factor one?) and a prediction error network (factor two?; for discussion see McKay, 2019; McKay & Furl, 2017).

***Basic ideas***

 Let us first be clear about what it means to provide a “hybrid” theory.

The two-factor theory and the prediction error theory, strictly speaking, are not formulated at the same level of explanation. The former, at least primarily, belongs to the psychological level of explanation and the latter to the neurophysiological level of explanation. To provide a “hybrid” theory is, then, to link two levels of explanations. In this regard, our project follows Kapur’s (2003), which aims to provide “a heuristic framework for linking the psychological and biological in psychosis” (p.13).

 There are difficult philosophical issues concerning the psychological level of explanation and its relation to the neurophysiological level of explanation (Cummins, 1983; Fodor, 1968; Piccinini & Craver, 2011). Addressing these issues is beyond the scope of this paper. We will adopt a tentative terminology according to which psychological states and processes (such as those invoked in the two-factor theory) are “physically grounded” in neurophysiological states and processes (such as those in the prediction error theory). We will remain neutral with respect to the nature of the putative “physical grounding” relation, but it would be safe to assume at least that physical grounding implies counterfactual dependence; i.e., the claim that “a psychological state P is physically grounded in a neurophysiological state N” implies “if N had not occurred, then P would not have occurred”.

 Now we are ready to introduce the hybrid theory3.

According to the theory (which is based on our previous work; e.g., McKay, 2012; Miyazono, 2019; Miyazono et al., 2014), the first/second factor distinction in the two-factor framework corresponds to a crucial distinction in the prediction error framework, namely, the distinction between prediction errors and their estimated precision. More precisely, the first factor (at the psychological level) is physically grounded in an abnormal prediction error (at the neurophysiological level), and the second factor (at the psychological level) is physically grounded in the overestimation of the precision of this abnormal prediction error (at the neurophysiological level).

 Let us focus on the Capgras delusion and see how the theory works in this example.

*First Factor & Prediction Error*: We follow the standard account in the two-factor theory camp that the first factor in the Capgras delusion is the abnormal datum about a familiar face. This abnormal datum is physically grounded in an abnormal prediction error; i.e., a mismatch between the expected and actual autonomic response to a familiar face (Coltheart, 2010; Corlett et al., 2010).

*Second Factor & Estimated Precision*:Among others, a promising hypothesis is that the second factor is a “bias towards observational (or explanatory) adequacy” (“OA bias” hereafter); i.e., the tendency to form beliefs that accommodate perceptions, even where this entails adjustments to the existing web of belief (Stone & Young, 1997). One of us (McKay, 2012) has provided a probabilistic expression of this proposal; the OA bias is mathematically characterised as the bias of discounting the prior probability ratio4. This hypothesis can be connected to a prediction error theoretic idea.

 Not all prediction errors are equally trustworthy; some are more reliable than others. It is important, therefore, that we are sensitive to the trustworthiness of given prediction errors and that we deal with trustworthy prediction errors and untrustworthy ones differently; i.e., we take the former seriously but not the latter. To use an everyday example, a newspaper headline alleging that Elvis Presley is still alive could be expected to engender a prediction error in most media consumers. However, the *precision* readers assign to this prediction error might vary considerably, depending on the source of the headline (e.g., *Weekly World News* vs. the *Washington Post*)*.* In the predictive processing (or predictive coding) framework, the brain estimates the “precision” of prediction errors, which is the indicator of their trustworthiness. Prediction errors are prioritised over prior beliefs when the prediction errors are estimated to be precise, while prior beliefs are prioritised over prediction errors when the prediction errors are estimated to be imprecise.

Our proposal is that the putative OA bias involves misestimating the precision of relevant prediction errors. More precisely, the OA bias (which is the second factor at the psychological level of our hybrid model) is physically grounded in the *over*estimation of the precision of abnormal prediction errors (in which the first factor is physically grounded). When the precision of an abnormal prediction error is overestimated (as in overestimating the reliability of the journalism in *Weekly World News*), the abnormal prediction error is prioritised over prior beliefs, and it drives bottom-up belief updating processes (Adams et al., 2013; Fletcher & Frith, 2009; Frith & Friston, 2012). In effect, this is the OA bias.

 In sum, here is how the hybrid theory explains the Capgras delusion. Consider the case of a Capgras patient who believes that her husband has been replaced by an impostor. The patient encounters abnormal data concerning the face of her husband (the first factor). These data are physically grounded in autonomic prediction errors; i.e., mismatches between the expected and the actual autonomic responses. The Capgras patient finds herself in a situation where prior beliefs support the husband hypothesis but the abnormal data support the imposter hypothesis. She ends up adopting the imposter hypothesis because of the OA bias (the second factor), which is physically grounded in overestimation of the precision of the autonomic prediction error. A patient with vmPFC damage, in contrast, finds herself in a similar situation where prior beliefs and the abnormal data support competing hypotheses. Because of the lack of the second factor, however, she does not adopt the imposter hypothesis.

 This account can be easily generalised to many other delusions. Let us focus on the first factor. (We assume that the second factor, the OA bias, is shared in many, if not all, delusions.) As Coltheart (2010) points out, often the first factor of a delusion has something to do with prediction errors. For instance, the first factor of the delusion of mirrored-self misidentification might be physically grounded in an abnormal prediction error; i.e., a mismatch between the actual visual input concerning one’s own face and the expected image of one’s face that is derived from long-term representations (Coltheart, 2007). The first factor of the delusion of control might also be physically grounded in an abnormal prediction error; i.e., a mismatch between actual sensory inputs and the expected sensory inputs that are computed from the motor command (Fletcher & Frith, 2009; Frith & Friston, 2012).

The idea can also be extended to delusions in schizophrenia. For example, in the delusion of reference, some events and objects are experienced as being abnormally attention-grabbing or salient (Kapur, 2003). Abnormal experience of this kind, which could be seen as the first factor of the delusion of reference (Coltheart, 2000), is physically grounded in abnormal signalling of prediction errors.

***Objections***

We turn now to some potential objections to the hybrid theory, in particular objections concerning the internal coherence of the hybrid theory (Objection 1 and Objection 2) and the hybridity of the theory (Objection 3).

*Objection 1*: The alleged hybrid theory is internally incoherent because the two-factor theory and the prediction error theory have incompatible commitments with regard to the distinction between experience and inference. The two-factor theory draws a sharp distinction between (abnormal) experience and (abnormal) inference, corresponding to the distinction between the first factor and the second factor. The prediction error theory, in contrast, rejects a sharp distinction between them; both are united by the principle of prediction error minimisation.

This objection, however, ignores levels of explanation. It is true that two-factor theorists tend to presuppose a relatively sharp distinction between experiential and inferential factors and, relatedly, between perceptual and doxastic (i.e., belief-related) processes. However, this is typically a commitment about the psychological level of explanation (e.g., McKay, 2012). And this commitment might be consistent with the idea that there is no sharp distinction between experiential and inferential factors at the neurophysiological level: e.g., “[S]uch separation [between perceptual and doxastic processes], though it functions well at a descriptive level (some mental phenomena can meaningfully be described as beliefs and some as perceptions) does not require a separation at a deeper level” (Corlett et al., 2016, p. 1148).

*Objection 2*: The alleged hybrid theory is internally incoherent because the two-factor theory and the prediction error theory have incompatible commitments with regard to the causal interaction between perceptual and doxastic states. In particular, according to the two-factor theory, abnormal perceptual states cause delusional doxastic states in a bottom-up manner (Bayne & Pacherie, 2004). On the other hand, the prediction error theory insists that lower-level states, such as perceptual states, and higher-level states, such as doxastic states, interact in both ways in the hierarchy (Hohwy, 2004).

We propose to finesse this objection by distinguishing two separate issues. One is the issue of whether, in general, perceptual states only influence doxastic states in a bottom-up manner, or whether the latter can also influence the former in a top-down manner. On this issue, one might think, the two theories are incompatible. The prediction error theory (or, more precisely, the predictive processing theory of the brain) is committed to the two-way interaction model where perceptual and doxastic states are mutually influential. The two-factor theory, by contrast, is “not committed to perception being cognitively penetrable” (Ross, McKay, Coltheart, & Langdon, 2016, p. 47). But an absence of commitment to the cognitive penetrability of perception does not entail a commitment to cognitive impenetrability. Two-factor theorists can be neutral on that issue.

The second issue concerns the dominant direction of influence in the particular cases that we are interested in; i.e., the cases of delusional beliefs. The question is whether it is the bottom-up influence or the top-down influence that is dominant in the process of delusion formation. On this score, the two-factor theory (at least, our favourite version of the theory; but see endnote 4 for a version of the theory that is potentially top-down) is a bottom-up theory. Similarly, many versions of the prediction error theory are clearly committed to bottom-up influence being dominant in the delusion formation processes (an exception is Hohwy, 2013): e.g., “[T]he problem that leads to the positive symptoms of schizophrenia starts with false prediction errors being propagated upwards through the hierarchy. These errors require higher levels of the hierarchy to adjust their models of the world. However, as the errors are false, these adjustments can never fully resolve the problem” (Fletcher & Frith, 2009, p. 55). As far as this issue is concerned, therefore, there is no fundamental gap between the two-factor theory and the prediction error theory.

*Objection 3*: The alleged hybrid theory is not really a “hybrid” theory at all; it is just a prediction error theory in which all crucial explanatory work is done by prediction error theoretic ideas. The two factor theoretic terms like “first factor” or “second factor” are red herrings; they do not play any significant explanatory roles. In fact, the alleged hybrid theory does not seem to add much to existing theories in the prediction error theory camp (e.g., Adams et al., 2013; Fletcher & Frith, 2009; Frith & Friston, 2012).

We claim that our theory deserves the name “hybrid” for four reasons.

First, our theory is coherent with what we take to be the core commitments of the two-factor theory. Perhaps it is obvious that the prediction error theory can be framed so as to remain faithful to those commitments. But Corlett (2019), a prediction error theorist, has recently published a scathing critique of the two-factor theory, so it is at least not obvious to everyone that the two theories are consilient.

Second, our theory is coherent with the particular version of the two-factor theory according to which the OA bias is the second factor (McKay, 2012; Stone & Young, 1997).

Third, our theory, because of its hybrid nature, inherits theoretical merits from both theories. For example, the hybrid theory, just like the two-factor theory, provides a plausible account of the dissociation between the first factors in delusion formation and the delusional beliefs they occasion. This dissociation is rarely, if ever, acknowledged by prediction error theorists. In the case of vmPFC damage, for example, the autonomic prediction errors are present but the Capgras delusion is not formed. This is because the mere presence of autonomic prediction errors is not sufficient; it is only when the precision of the autonomic prediction errors is overestimated that the Capgras delusion is formed. More generally, the hybrid theory – like the two-factor theory – provides an account of the diversity and recurrence of delusional themes; explaining why, for instance, some delusional patients have Capgras delusion and others mirrored-self misidentification (different delusions implicate different first factors).

Fourth and finally, our theory, because of its hybrid nature, has a wide scope of application (although it might not be applicable to all delusions; see Concluding remarks). The two-factor theory provides a plausible account of the monothematic delusions (e.g., the Capgras delusion) that can arise due to neuropsychological deficits. In contrast, the prediction error theory provides a plausible account of delusions (e.g., the delusion of reference) in schizophrenia. Our hybrid theory provides a unified explanation of both types of delusions.

## Concluding remarks

It is sometimes said that the two-factor theory and the prediction error theory are competing accounts of the delusion formation process. In our view, however, a rapprochement between the two-factor theory and the prediction error theory is both possible and desirable. We have outlined a hybrid account that incorporates central ideas of both theories. Against anticipated objections, we have argued that the hybrid theory is internally coherent, and that it constitutes a genuine hybrid between the two-factor theory and the prediction error theory.

 Of course, the hybrid theory as it stands does not answer all questions about the process of delusion formation. For example, it is not clear how the hybrid theory accommodates a role for motivational factors in delusion formation (McKay, Langdon, & Coltheart, 2005). Relatedly, although the hybrid theory has a wide scope of application, it might not explain all delusions. A particularly difficult example would be anosognosia, which we may need a separate account of (for more on the hybrid theory and delusion in anosognosia, see Miyazono, 2019).

 Finally, we have not touched on certain other theories of delusion, such as the phenomenological theory (e.g., Sass & Byrom, 2015). An interesting question for future research would be whether our hybrid account is compatible with such theories.

**Acknowledgements**

We received helpful feedback at the seminars and conferences in which early versions of this paper were presented, including the PERFECT Reading Group at the University of Birmingham (March 2017), the Philosophy, Psychology, and Informatics Group at the University of Edinburgh (March 2017), The 31st International Congress of Psychology at PACIFICO Yokohama (July 2016), and the Delusions Lunchtime Seminar at the University of Birmingham (June 2015). We also thank two anonymous referees for their insightful comments and suggestions. Some parts in this paper come from revising Section 4.4 of *Delusions and Beliefs: A Philosophical Inquiry*. Routledge: Abingdon and New York. © 2019, Kengo Miyazono. Kengo Miyazono acknowledges the support of JSPS KAKENHI (grant number 16H06998 & 18H00605). Ryan McKay acknowledges the support of the Cogito Foundation (grant number R10917).

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Notes**

1. As Coltheart and colleagues have often characterised the first factor in their scheme as a prediction error (e.g., Coltheart et al., 2011), contrasting the two-factor account with the “prediction error” theory may be misleading. Nevertheless, we use this terminology to be consistent with much commentary about the two theories.
2. Following Coltheart and colleagues (2010), we use the phrase “abnormal datum” rather than “abnormal experience” because the former is more neutral as to whether the data are consciously accessible or not. We assume that the abnormal data can be, but need not be, consciously accessible. Conscious or not, we take “abnormal datum” to be a psychological notion that belongs to the psychological (rather than neurophysiological) level of explanation. This will be crucial later in the discussion of our hybrid theory, which connects the two-factor theoretic explanation, which belongs to the psychological level, with the prediction error theoretic explanation, which belongs to the neurophysiological level.
3. The hybrid theory we advance is not the only way of combining the two-factor theory and the prediction error theory. Other hybrid theories are possible (e.g., see endnote 4; see also Gadsby and Hohwy (forthcoming) for a related discussion).
4. According to another proposal, the second factor is a bias of discounting or neglecting evidence against adopted delusional hypotheses (Coltheart et al., 2010). This proposal can also be combined with a (top-down) prediction error story to form another hybrid theory. For instance, the bias of discounting or neglecting counterevidence might be physically grounded in the underestimation of the precision of prediction errors; i.e., the failed predictions from delusional hypotheses.

## 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).

Bayne, T., & Pacherie, E. (2004). Bottom-up or top-down: Campbell's rationalist account of monothematic delusions. *Philosophy, Psychiatry, & Psychology*, *11*(1), 1-11.

Bentall, R. P., Kinderman, P., & Kaney, S. (1994). The self, attributional processes and abnormal beliefs: Towards a model of persecutory delusions. *Behaviour Research and Therapy*, *32*, 331-341.

Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences*, *6*(6), 237-242.

Brighetti, G., Bonifacci, P., Borlimi, R. & Ottaviani, C. (2007). ‘Far from the heart far from the eye’: Evidence from the Capgras delusion. *Cognitive Neuropsychiatry*, *12*(3), 189-97.

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

Coltheart, M. (2007). Cognitive neuropsychiatry and delusional belief: The 33rd Sir Frederick Bartlett lecture. *The Quarterly Journal of Experimental Psychology*, *60*, 1041-1062.

Coltheart, M. (2010). The neuropsychology of delusions. *Annals of the New York Academy of Sciences*, *1191*, 16-26.

Coltheart, M. (2013). On the distinction between monothematic and polythematic delusions. *Mind & Language*, *28*, 103-112.

Coltheart, M., Langdon, R., & McKay, R. (2007). Schizophrenia and monothematic delusions. *Schizophrenia Bulletin*, *33*, 642-647.

Coltheart, M., Langdon, R., & McKay, R. (2011). Delusional belief. *Annual Review of Psychology*, *62*, 271-298.

Coltheart, M., Menzies, P., & Sutton, J. (2010). Abductive inference and delusional belief. *Cognitive Neuropsychiatry*, *15*, 261-287.

Corlett, P. R. (2019). Factor One, familiarity and frontal cortex: A challenge to the two-factor theory of delusions. *Cognitive Neuropsychiatry, 24*(3), 165-177.

Corlett, P. R., & Fletcher, P. C. (2015). Delusions and prediction error: Clarifying the roles of behavioural and brain responses. *Cognitive Neuropsychiatry*, *20*, 95-105.

Corlett, P. R., Honey, G. D., Aitken, M. R., Dickinson, A., Shanks, D. R., Absalom, A. R., … Robbins, T. W. (2006). Frontal responses during learning predict vulnerability to the psychotogenic effects of ketamine: Linking cognition, brain activity, and psychosis. *Archives of General Psychiatry*, *63*(6), 611-621.

Corlett, P. R., Honey, G. D., & Fletcher, P. C. (2016). Prediction error, ketamine and psychosis: An updated model. *Journal of Psychopharmacology*, *30*, 1145–1155.

Corlett, P. R., Murray, G. K., Honey, G. D., Aitken, M. R., Shanks, D. R., Robbins, T., … Fletcher P. C. (2007). Disrupted prediction-error signal in psychosis: Evidence for an associative account of delusions. *Brain*, *130*, 2387-2400.

Corlett, P. R., Taylor, J. R., Wang, X. -J., Fletcher, P. C., & Krystal, J. H. (2010). Toward a neurobiology of delusions. *Progress in Neurobiology*, *92*, 345-369.

Cummins, R. (1983). *The nature of psychological explanatio*n. Cambridge, MA: MIT Press.

Darby, R., Laganiere, S., Pascual-Leone, A., Prasad, S. & Fox, M. (2016). Finding the imposter: Lesions causing delusional misidentifications arecharacterized by a unique pattern of brain connectivity. *Brain,* *140*(2), 497-507.

Davies, M., Coltheart, M., Langdon, R., & Breen, N. (2001). Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, & Psychology*, *8*, 133-158.

Ellis, H. D., & Young, A. W. (1990). Accounting for delusional misidentifications. *The British Journal of Psychiatry*, *157*(2), 239-248.

Ellis, H. D., Young, A. W., Quayle, A. H., & De Pauw, K. W. (1997). Reduced autonomic responses to faces in Capgras delusion. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *264*, 1085-1092.

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: A Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, *10*, 48-58.

Fodor, J. A. (1968). *Psychological explanation*. New York: Random House.

Freeman, D., Garety, P. A., Kuipers, E., Fowler, D., & Bebbington, P. E. (2002). A cognitive model of persecutory delusions. *British Journal of Clinical Psychology*, *41*, 331-347.

Friston, K. (2010). The free-energy principle: A unified brain theory?. *Nature Reviews Neuroscience*, *11*, 127-138.

Frith, C. D., & Friston, K. J. (2012). False perceptions and false beliefs: Understanding schizophrenia. in A. M. Battro, S. Dehaene, M. S. Sorondo, & W. J. Singer (Eds.), *The Proceedings of the Working Group on Neurosciences and the Human Person: New Perspectives on Human Activities* (pp. 134-148). Vatican City: The Pontifical Academy of Sciences.

Gadsby, S., & Hohwy, J. (forthcoming). Why use predictive processing to explain psychopathology? The case of anorexia nervosa. in S. Gouveia, R. Mendonça, & M. Curado (Eds.), *The Philosophy and Science of Predictive Processing*. Bloomsbury.

Gradin, V. B., Kumar, P., Waiter, G., Ahearn, T., Stickle, C., Milders, M., … Steele, J. D. (2011). Expected value and prediction error abnormalities in depression and schizophrenia. *Brain*, *134*(6), 1751-1764.

Griffiths, O., Langdon, R., Le Pelley, M. E., & Coltheart, M. (2014). Delusions and prediction error: Re-examining the behavioural evidence for disrupted error signalling in delusion formation. *Cognitive Neuropsychiatry*, *19*, 439-467.

Hirstein, W. S. & Ramachandran, V. S. (1997). Capgras syndrome: A novel probe for understanding the neural representation of the identity and familiarity of persons. *Proceedings of the Royal Society B: Biological Sciences*, *264*, 437–44.

Hohwy, J. (2004). Top-down and bottom-up in delusion formation. *Philosophy, Psychiatry, & Psychology*, *11*(1), 65-70.

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. *American Journal of Psychiatry*, *160*, 13-23.

McKay, R. (2012). Delusional inference. *Mind & Language*, *27*, 330-355.

McKay, R. (2019). Measles, magic and misidentifications: A defence of the two-factor theory of delusions. *Cognitive Neuropsychiatry*, *24*(3), 183-190.

McKay, R. & Furl, N. (2017). Doppelgängers and dissociations: Lesion network mapping illuminates misidentification delusions. *Brain, 140*(2), 262-265.

McKay, R., Langdon, R., & Coltheart, M. (2005). “Sleights of mind”: Delusions, defences, and self-deception. *Cognitive Neuropsychiatry*, *10*, 305-326.

Miyazono, K. (2019). *Delusions and Beliefs: A Philosophical Inquiry*. Abingdon and New York: Routledge.

Miyazono, K., Bortolotti, L., & Broome, M. R. (2014). Prediction-error and two-factor theories of delusion formation. In N. Galbraith (Ed.) *Aberrant Beliefs and Reasoning*. Psychology Press (pp. 34-54). New York: Routledge.

Murray, G. K., Corlett, P. R., Clark, L., Pessiglione, M., Blackwell, A. D., Honey, G., … Fletcher, P. C. (2008). Substantia nigra/ventral tegmental reward prediction error disruption in psychosis. *Molecular Psychiatry*, *13*(3), 267.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283-311.

Ross, R. M., McKay, R., Coltheart, M. & Langdon, R. (2016). Perception, cognition, and delusion. *Behavioral and Brain Sciences, 39*, e258.

Sass, L., & Byrom, G. (2015). Phenomenological and neurocognitive perspectives on delusions: A critical overview. *World Psychiatry*, *14*(2), 164-173.

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., … Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry*, *84*(9), 634-643.

Stone, T., & Young, A. W. (1997). Delusions and brain injury: The philosophy and psychology of belief. *Mind & Language*, *12*, 327-364.

Tranel, D., Damasio, H., & Damasio, A. R. (1995). Double dissociation between overt and covert face recognition. *Journal of Cognitive Neuroscience*, *7*, 425-432.

Williams, D. (2018). Hierarchical Bayesian models of delusion. *Consciousness and Cognition, 61*, 129-47.