Please cite and quote from published version:
***Cognitive Neuropsychiatry*, 26 (2021), 213–230.**

Failure of hypothesis evaluation as a factor in delusional belief

MAX COLTHEART AND MARTIN DAVIES

Abstract

Introduction: In accounts of the two-factor theory of delusional belief, the second factor in this theory has been referred to only in the most general terms, as a failure in the processes of hypothesis evaluation, with no attempt to characterise those processes in any detail. Coltheart and Davies (2021) attempted such a characterisation, proposing a detailed eight-step model of how unexpected observations lead to new beliefs based on the concept of abductive inference as introduced by Charles Sanders Peirce.

Methods: In this paper, we apply that model to the explanation of various forms of delusional belief.

Results*:* We provide evidence that in cases of delusion there is a specific failure of the seventh step in our model: the step at which predictions from (delusional) hypotheses are considered in the light of relevant evidence.

Conclusions: In the two-factor theory of delusional belief, the second factor consists of a failure to reject hypotheses in the face of disconfirmatory evidence.

1. Introduction

William James said: “The delusions of the insane are apt to affect certain typical forms, very difficult to explain. But in many cases they are certainly theories which the patients invent to account for their bodily sensations” (James, 1890, Volume 2, p. 114).

We think this is correct. So did one of the most influential recent thinkers about delusion, Brendan Maher: “Delusional beliefs, like normal beliefs, arise from an attempt to explain [unusual] experience” (1999, p. 550).

If James and Maher were correct, then for any delusional condition it must be possible to identify some unusual (unexpected, unpredicted) phenomenon to which a person with that delusion has been exposed and must also be the case that, if the delusional belief were true, the unexpected phenomenon would be expected. When this is so, the delusional belief does provide an explanation of the unexpected phenomenon, just as James and Maher claimed.

Davies, Coltheart, Langdon and Breen (2001) considered six distinct types of delusional condition, identifying for each a specific unexpected phenomenon which would be present for the delusional person and which would no longer be unexpected were the delusional belief true. Columns 1–3 of Table 1 summarise their considerations.

[Insert Table 1 about here]

For all the delusional conditions referred to in Table 1, there is so intimate a relationship between the unexpected phenomenon (Column 2) and the content of the delusional belief (Column 3) that it would be very difficult to claim that the unexpected phenomenon played no causal role in the genesis of the delusional belief. So Davies et al. (2001) concurred with James and Maher that there is such a causal role.

However, Davies et al. (2001) departed from James and Maher in proposing that abnormalities of the kind listed in Column 2 of Table 1, though causal, are not *sufficient* to result in a delusional condition. The reason is simple: for every one of these abnormalities, patients exist who possess this abnormality but who are not delusional. For each abnormality, such nondelusional patients are referred to in Column 4 of Table 1.

A comment is necessary concerning the work of Vuilleumier et al. (2003) cited in Table 1. In that study, autonomic responding was not directly measured. But the authors concluded that in this patient unfamiliar faces evoked “unconscious signals of familiarity” and hence “affective meaning” (p. 903). The patient’s exaggerated affective response to unfamiliar faces is vividly illustrated in this excerpt:

> During the first grand round on the ward (4 days after admission), J.R. greeted one of the authors (T.L.) by using the familiar form of personal pronoun (*tu*, unusual in French with an unknown physician), smiling at him as if he was somebody known to her, though she could not retrieve his name right away. Realizing her mistake, she excused herself and said: “Sorry, you must be the professor, it got me again, I cannot trust my own perceptions. When you entered this door I thought I knew you well, well enough so that

> you would embrace me and call me by my first name. Apparently we do not know each other, but I still have that feeling of having met you many times." (pp. 892–893)

The fact that, as Column 4 shows, not all patients with one of the abnormalities listed in Column 2 of Table 1 are delusional gave rise to the two-factor theory of delusional belief (Langdon and Coltheart, 2000; Davies et al., 2001; Coltheart, 2007; Coltheart, Langdon and McKay, 2011)[1], which asserts that when any of the unexpected phenomena listed in Column 2 of Table 1 are present (this is Factor 1 of the two-factor theory), a delusion will only result when a second factor is also present. This Factor 2 is considered to be present in all of the cases referred to in Column 1 of Table 1 (that is why they are delusional) but is absent in the cases referred to in Column 4 of Table 1 (that is why they are not delusional).

So the findings reported in Column 4 of Table 1 provide evidence against any one-factor account of delusional belief, such as the accounts offered by James and Maher.

Some have been persuaded by this reasoning. For example,

> It is important to clarify that a lack of SCR [skin conductance response] to familiar faces is not sufficient to produce the CD [Capgras Delusion], since, i.e. fronto-ventromedial lesions produce the same dissociation between autonomic response and overt recognition (Tranel et al., 1995) but do not cause the CD. (Bobes et al., 2016, p. 31)

and

> As has been cogently argued[2], two deficits are necessary to explain these delusions: a primary deficit (paralysis or memory loss) and a failure to suppress the implausible responses that result from this deficit. In the case of neurological patients, false beliefs seem to derive from the coincidence of damage in two locations, with the abnormal belief formation associated with damage to the prefrontal cortex. (Fletcher and Frith, 2009, pp. 50–51)

But not all have been persuaded. There is a "predictive-coding" one-factor account of delusion according to which there is a single deficit that generates delusion: an abnormality in the encoding of the precision of predictive-error signals.

> We will consider hallucinosis, abnormal eye movements, sensory attenuation deficits, catatonia, and delusions as various expressions of the same core pathology: namely, an aberrant encoding of precision.
>
> The basic idea is that faulty inference leads to false concepts (delusions) or percepts (hallucinations) and that this failure is due to a misallocation of precision to hierarchical representations in the brain (Adams, Stephan, Brown, Frith, & Friston, 2013, p. 1).

and

[1] https://maxcoltheart.wordpress.com/the-two-factor-theory-of-delusion/ lists all 27 papers from our group concerning this theory of delusion.

[2] The text cites Coltheart (2007) and Coltheart et al (2011) as the source of these arguments.

> While some psychotic patients get paranoid, others experience passivity, others still have multiple bizarre delusions. We posit a single factor, prediction error dysfunction for delusion formation and maintenance (Corlett, Taylor, Wang, Fletcher, & Krystal, 2010, p. 361).

However, these modern proponents of a one-factor account of delusional belief have never confronted the arguments against such an account that are based on the cases listed in Column 4 of Table 1. Until that is accomplished, we concur with Braun & Suffren (2011, p. 2) that the two-factor theory is "the most influential neurocognitive account of delusion in the scientific literature".

This second factor has been described as an *impairment of belief evaluation* in papers on the two-factor theory. The general idea is that the beliefs invoked as explanations for the unexpected phenomena ought to be negatively evaluated and hence rejected, because there is substantial evidence against them (and also because in most cases they are implausible or even bizarre). Such rejection is what successfully happens with the cases in Column 4 of Table 1. That it does not happen in the cases in Column 1 of Table 1 is, according to the two-factor theory of delusional belief, because belief evaluation does not proceed as it should.

That being said, proponents (and critics) of the two-factor theory have pointed out that too little has been said by such proponents about what the processes are by which belief evaluation – or, more properly, hypothesis evaluation – is carried out (and how these processes are impaired in delusional people):

> [V]ery little has yet been said about Factor 2. All that has been said is that it is a defect of a hypothesis evaluation system, a defect claimed to be present in all deluded patients regardless of the form of the delusion and responsible for the deluded person accepting the hypothesis as a belief despite all the evidence available that is inconsistent with the belief. (Coltheart, 2010, p. 21)

More needs to be said about this to flesh out the two-factor theory. As Marshall and Halligan (1996, p. 8) said:

> One would … hope that theories of normal belief-formation will eventually cast light on both the content of delusions and on the processes by which the beliefs came to be held.

That is what we aim to achieve in this paper.

Before we consider hypothesis *evaluation*, there is a critical question to ask about hypothesis *generation*: By what procedure could people generate, given some unexpected phenomenon, any proposition which has the property that, if it were true, this phenomenon would no longer be unexpected? How do people arrive at James's explanatory "theories" or Maher's explanatory "hypotheses" when they are confronted with observations that they did not expect and so want to explain?

This critical issue was not considered by James or Maher, and has also not been discussed by proponents of the two-factor theory of delusional belief.

The process by which explanatory hypotheses are generated from surprising observations is *abduction*, a concept introduced and elaborated by the American pragmatist philosopher Charles Sanders Peirce:

> Long before I first classed abduction as an inference it was recognized by logicians that the operation of adopting an explanatory hypothesis,—which is just what abduction is,—was subject to certain conditions. Namely, the hypothesis cannot be admitted, even as a hypothesis, unless it be supposed that it would account for the facts or some of them. The form of inference therefore is this:
>
> > The surprising fact, *C*, is observed;
> > But if *A* were true, *C* would be a matter of course.
> > Hence, there is reason to suspect that *A* is true.
>
> (Peirce, 1903/1998, p. 231)

The close connection between Peirce's ideas about abduction and the genesis of delusional belief is immediately obvious if one instantiates Peirce's *C* as "This person looks like my spouse but does not evoke a response of the autonomic nervous system" and his *A* as "This person is a stranger, not my spouse", to yield:

> The surprising fact *C* "This person looks like my spouse but does not evoke a response of the autonomic nervous system" is observed;
> But if "This person is a stranger, not my spouse" were true, *C* would be a matter of course.
> Hence, there is reason to suspect that "This person is a stranger, not my spouse" is true.

2. The Peircean pathway and the explanation of delusional belief

Peirce had much to say about how abduction might work, but his views on the topic were scattered throughout his writings. Davies and Coltheart (2020) gathered together Peirce's views on abduction and assembled them into a coherent whole. They concluded that his otherwise very promising account suffered from two problems which they referred to as *the problem of hypothesis generation* and *the problem of multiple hypotheses*.

Coltheart and Davies (2021) considered these two problems, and suggested solutions to each one. This allowed them to arrive at a viable model, developed from Peirce's work, of the pathway by which people arrive at new beliefs in response to the observation of surprising (unexpected, unpredicted) facts. This Peircean pathway is depicted in Figure 1.[3]

[3] Maher (1974, 1999) argued that delusions arise as explanations of unusual perceptual phenomena and Maher and Ross (1984) set out a five-step model that shares some components with our eight-step 'Peircean pathway' model. Maher and Ross's first two steps, *initial observation* of an unexpected event and *experience of puzzlement* (which includes development of a preliminary hypothesis), correspond to the first two steps of our model; and their third step, *additional observations*, corresponds to the hypothesis evaluation stage (Steps 5 through 7) of our model. Maher and Ross's fourth step, *the explanatory insight*, corresponds to the hypothesis

[insert Figure 1 about here]

How might the model of normal belief formation depicted in Figure 1 make the characterisation of the second factor as "an impairment of hypothesis evaluation" less vague? Can we say any more about the second factor than that it is an impairment at one or more of Steps 5 through 7 (which constitute the Hypothesis Evaluation system)?

2.1 Identifying the source of the failure of hypothesis evaluation: Observations of patients' behaviour

Consider the following account of the Capgras patient YY, who believed her father had been replaced by an impostor:

> At the time of testing the delusion was persistent and confirmed by YY's attempts to remove what she believed was the impostor's mask. Furthermore, YY repeatedly tested her father (whom she misidentified) with questions about her childhood (e.g., "Where did we celebrate my 10-year-old birthday?") with the aim of exposing him as an impostor. (Brighetti et al., 2007, p. 192)

Here YY generated the prediction, "This person will not be able to answer questions about my childhood" from the Capgras delusion hypothesis "This person is not my father, but a stranger". Furthermore, since it is surprising that a stranger should look just like one's father, YY generated the ancillary hypothesis that the impostor appeared like her father because he was wearing a mask and from this hypothesis generated the prediction that it should be possible to locate the mask on the impostor's face and to remove it. Step 5 of the Figure 1 model is deductive inference of predictions from the delusional hypothesis; clearly, it was functioning in YY.

YY also attempted to *test* these predictions, making efforts to determine whether the impostor could answer questions about her childhood, and to determine whether it was possible to locate a mask on the impostor's face and remove it. Thus Step 6, experimental testing of the predictions derived from the delusional hypothesis, was also functioning in YY.

The results of these tests carried out by YY were inconsistent with her predictions. Her father when questioned was able to answer questions about her childhood correctly[4]; and she was unable to locate a mask on his face. So if Step 7 were functioning normally, YY should have recognised that the delusional hypothesis had been disconfirmed and rejected it. But she did not. Therefore, we may conclude that for YY Factor 2 was not just some general impairment of the Hypothesis Evaluation system (Steps 5 through 7). It was instead an impairment of one specific component of that system, namely, Step 7.

---

generation stage (Steps 2 through 4) of our model, but only in cases where the generated hypothesis is assessed (at Step 3) as one which, if true, would provide a great deal of understanding of why the surprising fact was observed and, consequently, is considered (at Step 4) to be especially pursuit-worthy. Maher and Ross's final step, *the process of confirmation*, might be understood as skipping the hypothesis evaluation stage and proceeding directly from the fourth Peircean step to the eighth. (Peirce, himself, acknowledged this possibility. For discussion, see Davies & Coltheart, 2020, Section 2.9 and Coltheart & Davies, 2021, p. 3.)

[4] Rosita Borlimi, personal communication, 15 July 2020.

Is this true only of YY, or more generally true of delusional cases?

Frazer and Roberts (1994) studied a woman with Capgras delusion (their Case 3) who believed that her son had been replaced by an impostor. Here again it would seem that Steps 5 and 6 were functioning. The patient said the impostor's eyes were not the same colour as her son's eyes (which were, say, blue). From the impostor hypothesis – thus elaborated – the patient could derive the prediction that the impostor's eyes would be found *not to be blue* (Step 5). It would be a straightforward matter to test this prediction by observation (Step 6) and the result of such observation would be inconsistent with the patient's prediction. So if Step 7 were functioning normally, the patient should have recognised that it was the nondelusional hypothesis ("This is my son"), rather than the *non-blue-eyed impostor* hypothesis, that was supported by observation. She should have rejected the delusional hypothesis; yet that is not what happened.

Bisiach (1988) reported the following conversation with a patient who had the delusional condition somatoparaphrenia. The patient believed that his left arm (which was paralysed following a right-hemisphere stroke) was not his:

> The examiner, placing the patient's left hand in the patient's right visual field, asks: "Whose hand is this?"
> Patient: Your hand.
> The examiner then places the patient's hand between his own hands, and asks: "Whose hands are these?"
> Patient: Your hands.
> Examiner: How many of them?
> Patient: Three.
> Examiner: Ever seen a man with *three* hands?
> Patient: A hand is the extremity of an arm. Since you have three arms it follows that you must have three hands. (p. 469)

Here Step 5 was used correctly to make deductive inferences from the delusional (somatoparaphrenic) hypothesis "This arm is not mine", such as "This arm must be someone else's arm" and "If this arm is the examiner's arm, then since I can see two other arms attached to him, he must have three arms" and "If he has three arms, it follows that he must have three hands".

It is unclear whether the patient derived, from the hypothesis, predictions that could be tested by experiment or observation. But the patient presumably knew that people do not have three arms or three hands and would have agreed with this had he been questioned about it (Step 6).[5] Nevertheless, this knowledge did not lead him to reject the somatoparaphrenic hypothesis at Step 7. On the contrary, he confabulated an explanation of his statement that the Examiner had three hands.

[5] In Figure 1, Step 6 of our 'Peircean pathway' model follows Peirce, himself, who usually talked about experimental testing of predictions. But predictions can also be tested by observations (without there being any experimental manipulation involved) or just by using already-existing knowledge.

This failure of patients to use their everyday knowledge to falsify their delusional hypotheses is also evident in this exchange (Alexander, Stuss and Benson, 1979) between an examiner E and a delusional patient S who believed that he had two families of identical composition.

> E. Isn't that [two families] unusual?
> S. It was unbelievable!
> E. How do you account for it?
> S. I don't know. I try to understand it myself, and it was virtually impossible.
> E. What if I told you I don't believe it?
> S. That's perfectly understandable. In fact, when I tell the story, I feel that I'm concocting a story … . It's not quite right. Something is wrong.
> E. If someone told you the story, what would you think?
> S. I would find it extremely hard to believe. I should be defending myself.
> (p. 335)

As Alexander et al. (1979) wrote: "He [the patient] could not use his awareness of contradictory statements to modify his beliefs in accordance with the environmental cues" (p. 336).

What these examples have in common is that there is evidence available to the patient (coming either from a patient's everyday knowledge or from "experiments" or observations initiated by the patient) that disconfirms the delusional hypothesis; yet the patient does not abandon that hypothesis.

2.2 Identifying the source of the failure of hypothesis evaluation: Bias against disconfirming evidence

Our claim is that in the kinds of delusions with which we are concerned, one or more of Steps 5 through 7 of our model fail to be properly executed: this failure is our Factor 2. Our model would thus be falsified if it could be shown that in any such delusional cases these three steps were executed properly. In Section 2.1 we gave a number of examples of delusional cases in which it seems that both Step 5 and Step 6 were properly executed. It may be that there are other delusional cases where Step 5 and/or Step 6 are not properly executed – it remains for future work to investigate this (perhaps by finding cases where the patient makes no attempt at deducing predictions from the delusional hypothesis). But for cases of the kind discussed in Section 2.1, where execution of Steps 5 and 6 does occur, our model demands that there is some form of failure of Step 7. What form might that failure take? Here, we explore the possibility that the failure to reject the delusional hypothesis at Step 7 results from a more general cognitive impairment or bias: a failure to reject (or at least downgrade) disconfirmed hypotheses or bias against disconfirmatory evidence.

In the literature on delusions in individuals with schizophrenia, this bias has been investigated using the BADE paradigm, which Sanford, Veckenstedt, Moritz, Balzan and Woodward (2014) describe thus:

> [T]he BADE task involves a set of scenarios that are accompanied by four possible interpretations, individually rated after each of three successively presented statements; one absurd interpretation, which seems implausible from the first statement and remains

> so throughout the trial; two lure interpretations (one neutral and one emotional), which seem plausible initially but are disconfirmed after the second or third statement; and one true interpretation, which does not seem to be the most plausible from the start but is confirmed by the final statement. … The BADE has been characterized as an unwillingness to down-rate initially plausible interpretations ('lures') as they are revealed to be implausible. (p. 2730)

Woodward, Moritz, Cuttler and Whitman (2006) found a stronger bias against disconfirmatory evidence in delusional compared to nondelusional schizophrenia patients. This was also reported by Riccaboni, Fresi, Bosia, Buonocore, Leiba, Smeraldi and Cavallaro (2012); and in their study regression analysis showed that a general disconfirmatory index (across lure and absurd ratings) was a significant predictor of delusion scores on the PANSS scale for measuring symptom severity.

Sanford et al. (2014) used a BADE task with healthy controls, patients with a diagnosis of obsessive-compulsive disorder (OCD), patients with a diagnosis of schizophrenia who were low-delusional, and patients with a diagnosis of schizophrenia who were high-delusional. The high-delusional group differed from the other groups on the 'evidence integration' component, corresponding to "relatively high plausibility ratings for disconfirmed interpretations and low ratings for confirmed (true) interpretations".[6] The authors concluded that

> These data support the finding that a reduced willingness to adjust beliefs when confronted with disconfirming evidence may be a cognitive underpinning of delusions specifically … and illustrates a cognitive process that may underlie maintenance of delusions in the face of counter-evidence. (p. 2729)

A meta-analysis by McLean, Mattiske and Balzan (2017) of eight studies using the BADE task with schizophrenia patients with or without delusions confirmed the association between delusionality and a stronger bias against disconfirmatory evidence, as did a literature review by Eisenacher and Zink (2017).

As far as we know, there has been no work with the BADE paradigm on nonschizophrenic delusional patients. Buchy, Woodward and Liotti (2007) found that the tendency not to adjust plausibility ratings when presented with evidence disconfirming lure interpretations was greater in nonclinical subjects with high scores on a schizotypy scale (SPQ; Raine, 1991) than in those with low schizotypy scores. In an extension of that study, Woodward, Buchy, Moritz and Liotti (2007) found that the 'integration of disconfirmatory evidence' factor in a component analysis was "correlated with delusional aspects of the SPQ" (p. 1025).[7]

In Section 2.1 we cited examples of the behaviour of patients with delusions that supported the idea that that the defect of hypothesis evaluation which is the second factor in the two-

[6] The 'evidence integration' component is thus a kind of composite score for bias against disconfirmatory evidence and bias against confirmatory evidence.

[7] The SPQ includes three subscales that directly assess, in nonclinical populations, tendencies towards having delusional beliefs: Ideas of Reference, Odd Beliefs or Magical Thinking, and Suspiciousness.

factor theory of delusional belief is specifically a defect at Step 7 of the 'Peircean pathway' model in Figure 1. Given that a failure to reject disconfirmed hypotheses, or bias against disconfirmatory evidence, would affect Step 7 of our model but not Steps 5 or 6, reports of an abnormally strong bias against disconfirmatory evidence in schizophrenia patients with delusions also support this idea.

Delusional ideas are often bizarre or outlandish: that is, they conflict with other things the deluded person knows or believes. Such background knowledge constitutes disconfirmatory evidence and should operate to reject the delusional hypothesis at Step 7. But, as we argue in this paper, Factor 2 in our theory is, specifically, a failure to use disconfirmatory evidence at Step 7, and that would include failure to use background knowledge that implies the bizarreness or outlandishness of a potential belief.

It follows that in cases of delusion where Steps 5 and 6 of our model are properly executed, a general cognitive impairment or bias should be demonstrable; namely, a failure to reject (or at least downgrade) disconfirmed hypotheses or bias against disconfirmatory evidence. Suppose there were cases of delusion where Steps 5 and 6 were properly executed and where the use of disconfirmatory evidence was no different from that of nondelusional people. That would not constitute a falsification of our model; but it would be damaging to the specific proposal being put forward here, namely, that the second factor is a bias against disconfirmatory evidence affecting Step 7 of the model.

3. Anosognosia (for hemiplegia)

Most patients suffering from hemiplegia correctly believe that (a) the paralysed limb is their own limb and also (b) they can no longer voluntarily move that paralysed limb.

However, some hemiplegic patients, though they correctly believe that they can no longer move the paralysed limb, express the delusional belief that this limb is not their own limb. This delusion is somatoparaphrenia, one of the delusions in Table 1.

And some hemiplegic patients, though they believe that the limb in question is their own, express the delusional belief that they can still voluntarily move the paralysed limb. This delusion is anosognosia (for hemiplegia).

What account might the two-factor theory of delusional belief offer for anosognosia? A critical point here is that, unlike all of the delusions listed in Table 1, anosognosia does not involve the (delusional) generation of a new belief. Instead, it involves the (delusional) retention of an old – that is, previously-held – belief.

Thus the full Peircean pathway depicted in Figure 1 is not relevant to anosognosia. That pathway describes how a new belief is generated from a surprising observation; but in anosognosia there is no surprising observation and no generation of a new belief. So Steps 1 through 4 of this pathway are not executed by the anosognosic patient, since there is no need for generation of some new candidate-for-belief: the patient already holds the relevant belief (because it has always been held).

Can the two-factor theory of delusional belief, nevertheless, still be applied to anosognosia? This theory, in its most general form, seeks to answer two questions about any delusional belief:

> The first question is, what brought the delusional idea to mind in the first place? The second question is, why is this idea accepted as true and adopted as a belief when the belief is typically bizarre and when so much evidence against its truth is available to the patient? (Coltheart et al., 2011, p. 271)

In the case of anosognosia, the answer to the first question is that the idea "I can move my left arm and leg" is entertained by the patient simply because it has been true for the patient throughout that person's entire life, until the occurrence of the brain damage that resulted in the hemiplegia.

The answer to the second question can be the same as it is for all the delusions listed in Table 1: a failure of hypothesis evaluation (i.e., Factor 2). Patients have data available to them contradicting the hypothesis "I can move my left arm and leg" – they cannot stand up from their chair, or perform bimanual tasks such as tying shoelaces or picking up a tray of glasses – but observations of these data do not result in rejection of the hypothesis. The data should be taken as disconfirmatory of the hypothesis, but are not.

For such disconfirmation to occur, patients would have to execute Steps 5 through 7 of the Peircean pathway.

First (Step 5), patients would have to *derive* (from the hypothesis that they can still move the left limbs) *a prediction* about standing up from a chair, tying shoelaces, or picking up a tray of glasses.

Second (Step 6), patients would need to conduct an *experiment* to test the prediction or (more likely) would need to regard an *observation* (falling on the floor, being unable to tie shoelaces, or dropping the tray of glasses on the floor) as testing the prediction.

Then (Step 7), patients would need to recognise the outcome of the experiment (or observation) as inconsistent with the derived prediction – so constituting a disconfirmation of that hypothesis, which therefore should be rejected.

A sufficiently strong bias against disconfirmatory evidence, operating at Step 7, could lead to anosognosic patients failing to reject the delusional hypothesis. As discussed earlier (Sections 2.1 and 2.2), there is evidence of just such a bias in delusional patients. Thus despite the differences between the delusion of anosognosia and the various other delusions in Table 1, the two-factor theory is readily applicable to the explanation of anosognosia.

As noted above, the BADE task has not been used to assess a bias against disconfirmatory evidence in patients with delusions, but without a psychiatric diagnosis. Patients with anosognosia following right hemisphere stroke have, however, been assessed using a riddle task (Vocat, Saj, & Vuilleumier, 2013), which bears a family resemblance to the BADE task.

In the riddle task, subjects are given five successively more informative clues and, after each clue, asked to guess the target word (e.g., HEART). The first clue is sufficiently uninformative to create doubt in healthy subjects (e.g., "My weight is approximately 300 grams"); the final clue is intended to leave no doubt about the correct answer ("Lovers often draw me"). The measure of bias against disconfirmatory evidence is the number of times (across ten riddles) the subject proposes the same incorrect guess following two consecutive clues in the same riddle.

In a study of right hemisphere stroke patients – with or without anosognosia – and healthy control subjects, the key finding was that anosognosia patients were more than twice as likely as the other two groups to repeat the same incorrect guess. Thus patients with anosognosia "required a repeated signal of errors, or a larger incongruence between a new clue and the previous guess, in order to prompt a re-appraisal of their preceding responses and to trigger a new solution" (Vocat et al., 2013, p. 1778). In other words, these anosognosic patients showed a bias against disconfirmatory evidence.

3.1 The cognitive nature and neural basis of the second factor

As noted earlier, proponents of the two-factor theory of delusional belief have not said enough about the cognitive nature of the second factor. They have, however, proposed that its neural basis is damage to right dorsolateral prefrontal cortex (rDLPFC) (Coltheart, 2007, 2010).

In the present paper, we have argued that the second factor is an impairment at Step 7 of the 'Peircean pathway' model, resulting in failure to reject a hypothesis despite the availability of disconfirmatory evidence. We suggest that this claim is broadly consistent with the proposed neural basis of the second factor.

First, in an fMRI study of associative learning (Fletcher, Anderson, Shanks et al., 2001), healthy subjects learned associations between cues and outcomes (fictitious drugs and syndromes) and occasional trials were surprising given the previous learning. As Coltheart (2010) observes, rDLPFC activation was high when unexpected cue-outcome evidence (disconfirmatory of the current hypothesis) was presented, by comparison with presentation of predictable (confirmatory) evidence.

Second, using their riddle task (see above), Vocat et al. (2013) found that anosognosia patients were less sensitive to disconfirmatory evidence than controls; that is, more likely to repeat an incorrect guess after a disconfirming clue. They suggested (citing Coltheart, 2010) that the patients' failure to update current beliefs despite the presence of incongruent information "might reflect damage to prefrontal cortical areas" (p. 1778). Note that these patients' damage to prefrontal cortical areas was only in the *right* hemisphere; indeed, their brain scans revealed no left hemisphere damage at all.

Third, Coltheart, Cox, Sowman et al. (2018) found that impairing rDLPFC activity by applying transcranial magnetic stimulation increased hypnotic susceptibility – that is, made it less likely that a hypnotic suggestion would be rejected. In a control condition (TMS to *left* DLPFC) there was no effect on hypnotic suggestibility.

Fourth, in a model-based lesion-mapping study of 328 patients with focal lesions, Gläscher, Adolphs and Tranel (2019) found that lower rates of hypothesis updating in response to disconfirmatory evidence during performance of the Wisconsin Card Sorting Task were associated with lesions "located primarily in the right PFC [prefrontal cortex] reaching from dorsolateral PFC to the frontal pole and mostly focused in the underlying white matter" (p. 5).

Finally, we do not mean to imply here that rDLPFC is the sole neural substrate of hypothesis evaluation. No doubt the cognitive processes leading from presentation of disconfirmatory evidence to rejection of the hypothesis depend upon brain circuits of which rDLPFC is just one component. In an fMRI study of detection and integration of disconfirmatory evidence (Lavigne, Metzak & Woodward, 2015), subjects were presented with two ambiguous images of animals. After presentation of the first image (e.g., 70/30 bird/dolphin), they gave a rating as to which animal was depicted; after presentation of the second image, they gave a new rating. The results revealed two functional networks that were active during performance of the task.

A *salience network*, including areas of right lateral prefrontal cortex (rLPC), reached a brief peak of activation when the second image was first presented and this peak was higher when the second image provided evidence that was disconfirmatory (e.g., 10/90 bird/dolphin), compared with confirmatory (e.g., 90/10 bird/dolphin), of the first rating. Lavigne et al. explicitly related the role of rLPC in this network to Fletcher et al.'s (2001) associative learning study and to Coltheart's (2010) proposal about the neural basis of hypothesis evaluation.

An *integration network*, including specifically rDLPFC, reached a peak of activation after the second image was presented – later than the peak activation of the salience network – and activation remained elevated until the end of the trial. Again, activation was higher in the disconfirm than in the confirm condition. The later peak and continuing activation in the integration network suggests a role in evaluation of the hypothesis (based on the first image) in the light of the evidence (provided by the second image).

Lavigne et al. (2015) interpreted the two networks as underlying two cognitive processes that are required for hypothesis evaluation and updating. First (the salience network), detection of conflict or mismatch between available evidence and an existing hypothesis is a crucial initial step; failure to detect that evidence is disconfirmatory will allow unwarranted credence in the hypothesis to be maintained. Second (the integration network), once the evidence is recognised as disconfirmatory, the hypothesis must be evaluated in the light of the evidence, to determine whether it can be modified or must be rejected outright.

Subsequent fMRI studies have shown that lower levels of activation of the integration network (now referred to as the *cognitive evaluation network*) in the disconfirm condition are associated with (i) poorer performance on the BADE task and (ii) delusions in schizophrenia patients (Lavigne, Menon & Woodward, 2020) and also with (iii) delusional ideation assessed by the SPQ delusion-related subscales (Lavigne, Menon, Moritz & Woodward, 2020).

4. On the scope of the two-factor theory of delusional belief

A distinction has been proposed (see e.g., Davies et al., 2001; Coltheart, Langdon and McKay, 2007; Radden, 2010; Coltheart 2013) between monothematic and polythematic delusional conditions. In cases of monothematic delusion, the patient has a single delusional belief or at most a small set of delusional beliefs concerning a single theme. In polythematic delusion, the patient exhibits a variety of delusional beliefs concerning many different themes.

Proponents of the two-factor theory have generally taken the following position:

> [T]here is a well-worked-out general neuropsychological theory of monothematic delusion, the two-factor theory. Whether polythematic delusion might be explained in a similar way is an open question. (Coltheart, 2013, p. 103)

All of the delusional conditions listed in Table 1 are instances of monothematic delusion.

When considering the question of the intended scope of the two-factor theory, the distinction between psychotic and non-psychotic cases is not quite right, because many of the monothematic delusions listed in Table 1 have been reported in people with a diagnosis of psychosis. Nor is the distinction between neurological and non-neurological cases quite right, because monothematic delusions have been reported in people where is there is no evidence of brain damage, such as believers in alien abduction (for discussion see Coltheart et al., 2011, p. 291). This is why the appropriate distinction here is between monothematic and polythematic delusion. The two-factor theory is committed to being able to explain monothematic delusional beliefs. Whether the two-factor theory is also applicable to the explanation of polythematic delusions remains to be seen; but there are some reasons to think that it might be applicable.

In Maher and Ross's (1984) model of delusion, an experience of puzzlement or feeling of significance attaches to the initial observation of an unexpected event (see above, footnote 3). Maher and Ross also note that a feeling of significance "may be activated independently by central neuropathology" (1984, p. 404) with the result that quite ordinary experiences are interpreted as having special significance, and delusions arise as explanations of this significance. Kapur (2003) proposed dysregulated dopamine transmission as underpinning these aberrant experiences of significance or salience and an fMRI study by Menon, Schmitz, Anderson et al. (2011) provided evidence linking this account of aberrant salience with delusions of reference in patients with schizophrenia. Thus, an impairment of the mesolimbic dopamine system could serve as the first factor in a two-factor account of delusions of reference – and perhaps other polythematic delusional conditions – with a severe impairment of the hypothesis evaluation system as the second factor (Coltheart et al., 2011; Coltheart, 2013, Section 3.2).

We argued earlier that a one-factor approach cannot explain the kinds of delusion listed in Table 1 (because of the types of nondelusional case listed in Column 4 of that Table) – that is, it cannot explain monothematic delusional beliefs such as Capgras delusion. The problem with a one-factor approach is that it predicts that all people with the abnormality specified in

Column 2 of Table 1 should be delusional, and the data in Column 4 of Table 1 show that this is not the case.

5. Conclusions

One-factor theories of delusional belief are contradicted by the cases listed in Column 4 of Table 1; the existence of these nondelusional cases indicates that a two-factor theory is needed (at least for monothematic delusions). In this paper we have attempted to flesh out the two-factor theory of delusional belief by using an explicit model (Coltheart and Davies, 2021) of the processes by which unusual experiences give rise to new beliefs. We argue that the second factor in this theory takes the form of a failure at Step 7 of the model – a specific failure to respond appropriately to evidence that is disconfirmatory of predictions made from delusional hypotheses.

Acknowledgements

We thank three anonymous reviewers for their insightful and constructive critiques of an earlier version of this paper.

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.

Alexander, M. P., Stuss, D. T., & Benson, D. F. (1979). Capgras syndrome: A reduplicative phenomenon. *Neurology*, *29*, 334–339.

Assal, G. (1983). "No, I am not paralyzed; it is the hand of my husband." *Schweizer Archiv fur Neurologie, Neurochirurgie und Psychiatrie, 133*, 151–157.

Binkofski, F., Buccino, G., Dohle, C., Seitz, R. J., & Freund H.-J. (1999). Mirror agnosia and mirror ataxia constitute different parietal lobe disorders. *Annals of Neurology*, *46*, 51–61.

Bisiach, E. (1988). Language without thought. In L. Weiskrantz (Ed.), *Thought Without Language* (pp. 464–484). Oxford: Oxford University Press.

Bobes, M. A., Góngora, D., Valdes, A., Santos, Y., Acosta, Y., Garcia, Y. F., Lage, A., & Valdés-Sosa, M. (2016). Testing the connections within face processing circuitry in Capgras delusion with diffusion imaging tractography. *NeuroImage: Clinical*, *11*, 30–40.

Braun, C. M. J., & Suffren, S. (2011). A general neuropsychological model of delusion. *Cognitive Neuropsychiatry, 16*, 1–39.

Breen, N., Caine, D., Coltheart, M., Roberts, C., and Hendy, J. (2000). Towards an understanding of delusions of misidentification: Four case studies. *Mind & Language, 15*, 74–110.

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*, 189–197.

Buchy, L., Woodward, T. S., & Liotti, M. (2007). A cognitive bias against disconfirmatory evidence (BADE) is associated with schizotypy. *Schizophrenia Research*, *90*, 334–337.

Coltheart, M. (2007). Cognitive neuropsychiatry and delusional belief. *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., and Davies, M. (2021). How unexpected observations lead to new beliefs: A Peircean pathway. *Consciousness and Cognition*, *87*, 103037

Coltheart, M., Cox, R., Sowman, P., Morgan, H., Barnier, A., Langdon, R., Connaughton, E., Teichmann, L., Williams, N., & Polito, V. (2018). Belief, delusion, hypnosis, and the right dorsolateral prefrontal cortex: A transcranial magnetic stimulation study. *Cortex*, *101,* 234–248.

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.

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.

Davies, M. & Coltheart, M. (2020). A Peircean pathway from surprising facts to new beliefs. *Transactions of the Charles S. Peirce Society*, *56*, 400–426.

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

Edelstyn, N. M. J., & Oyebode, F. (1999). A review of the phenomenology and cognitive neuropsychological origins of the Capgras syndrome. *International Journal of Geriatric Psychiatry, 14*, 48–59.

Eisenacher, S., & Zink, M. (2017). Holding on to false beliefs: The bias against disconfirmatory evidence over the course of psychosis. *Journal of Behavior Therapy and Experimental Psychiatry*, *56*, 79–89.

Fletcher, P. C., Anderson, J. M., Shanks, D. R., Honey, R., Carpenter, T. A., Donovan, T., Papadakis, N., & Bullmore, E. T. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature Neuroscience*, *4*, 1043–1048.

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.

Fourneret, P., Paillard, J., Lamarre, Y., Cole, J., & Jeannerod, M. (2002). Lack of conscious recognition of one's own actions in a haptically deafferented patient. *NeuroReport*, *13*, 541–547.

Frazer, S. J., & Roberts, J. M. (1994). Three cases of Capgras' syndrome. *British Journal of Psychiatry*, *164*, 557–559.

Gläscher, J., Adolphs, R., & Tranel, D. (2019). Model-based lesion mapping of cognitive control using the Wisconsin Card Sorting Test. *Nature Communications*, *10:20*, 1–12.

Heims, H. C., Critchley, H. D., Dolan, R., Mathias, C. J., & Cipolotti, L. (2004). Social and motivational functioning is not critically dependent on feedback of autonomic responses: Neuropsychological evidence from patients with pure autonomic failure. *Neuropsychologia*, *42*, 1979–1988.

James, W. (1890). *The Principles of Psychology* (in two volumes). New York, NY: Henry Holt and Company.

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.

Langdon, R., & Coltheart, M. (2000). The cognitive neuropsychology of delusions. *Mind & Language*, *15*, 183–216.

Langdon, R., Connaughton, E., & Coltheart, M. (2014). The Fregoli delusion: A disorder of person identification and tracking. *Topics in Cognitive Science, 6*, 615–631.

Lavigne, K. M., Menon, M., Moritz, S., & Woodward, T. S. (2020). Functional brain networks underlying evidence integration and delusional ideation. *Schizophrenia Research*, *216*, 302–309.

Lavigne, K. M., Menon, M., & Woodward, T. S. (2020). Functional brain networks underlying evidence integration and delusions in schizophrenia. *Schizophrenia Bulletin*, *46*, 175–183.

Lavigne, K. M., Metzak, P. D., & Woodward, T. S. (2015). Functional brain networks underlying detection and integration of disconfirmatory evidence. *NeuroImage*, *112*, 138–151.

Maher, B. A. (1974). Delusional thinking and perceptual disorder. *Journal of Individual Psychology*, *30*, 98–113.

Maher, B. A. (1999). Anomalous experience in everyday life: Its significance for psychopathology. *The Monist*, *82*, 547–570.

Maher, B. A. & Ross, J. S. (1984). Delusions. In H. E. Adams & P. B. Sutker (Eds.), *Comprehensive Handbook of Psychopathology* (pp. 383–409). New York: Springer.

Marshall, J. C., & Halligan, P. W. (1996). Towards a cognitive neuropsychiatry. In P. W. Halligan, & J. C. Marshall (Eds.), *Method in madness: Case studies in cognitive neuropsychiatry* (pp. 3–12). Hove, UK: Psychology Press.

McLean, B. F., Mattiske, J. K., & Balzan, R. P. (2017). Association of the jumping to conclusions and evidence integration biases with delusions in psychosis: A detailed meta-analysis. *Schizophrenia Bulletin*, 43, 344–354.

Menon, M., Schmitz, T. W., Anderson, A. K., Graff, A., Korostil, M., Mamo, D., Gerretsen, P., Addington, J., Remington, G., & Kapur, S. (2011). Exploring the neural correlates of delusions of reference. *Biological Psychiatry*, *70*, 1127–1133.

Peirce, C. S. (1903/1998). Harvard Lectures on Pragmatism (1903), Lecture 7: Pragmatism as the logic of abduction. In Peirce Edition Project (Eds.), *The Essential Peirce, Volume 2 (1893–1913)* (pp. 226–241). Bloomington: Indiana University Press.

Radden, J. (2011). *On Delusion*. London: Routledge.

Raine, A. (1991). The SPQ: A scale for the assessment of schizotypal personality based on DSM-III-R criteria. *Schizophrenia Bulletin, 17,* 555–564.

Ramachandran, V. S., & Blakeslee, S. (1998). *Phantoms in the brain: Probing the mysteries of the human mind*. New York: William Morrow.

Riccaboni, R., Fresi, F., Bosia, M., Buonocore, M., Leiba, N., Smeraldi, E., & Cavallaro, R. (2012). Patterns of evidence integration in schizophrenia and delusion. *Psychiatry Research*, *200*, 108–114.

Sanford, N., Veckenstedt, R., Moritz, S., Balzan, R. P., & Woodward, T. S. (2014). Impaired integration of disambiguating evidence in delusional schizophrenia patients. *Psychological Medicine*, *44*, 2729–2738.

Stirling, J. D., Hellewell, J. S. E., & Quraishi, N. (1998). Self-monitoring dysfunction and the schizophrenic symptoms of alien control. *Psychological Medicine, 28*, 675–683.

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

Vocat, R., Saj, A., & Vuilleumier, P. (2013). The riddle of anosognosia: Does unawareness of hemiplegia involve a failure to update beliefs? *Cortex*, *49*, 1771–1781.

Vuilleumier, P., Mohr, C., Valenza, N., Wetzel, C., & Landis, T. (2003). Hyperfamiliarity for unknown faces after left lateral temporo-occipital venous infarction: A double dissociation with prosopagnosia. *Brain*, *126*, 889–907.

Woodward, T. S., Buchy, L., Moritz, S., & Liotti, M. (2007). A bias against disconfirmatory evidence is associated with delusion proneness in a nonclinical sample. *Schizophrenia Bulletin*, *33*, 1023–1028.

Woodward, T. S., Moritz, S., Cuttler, C., & Whitman, J. C. (2006). The contribution of a cognitive bias against disconfirmatory evidence (BADE) to delusions in schizophrenia. *Journal of Clinical and Experimental Neuropsychology*, *28*, 605–617.

Young, A. W., Robertson, I. H., Hellawell, D. J., de Pauw, K. W., & Pentland, G. (1992). Cotard delusion after brain injury. *Psychological Medicine, 22*, 799–804.

Figure Caption:

Figure 1: An 8-step model of the adoption of a new belief in response to the observation of a surprising fact (Davies and Coltheart, 2020; Coltheart and Davies, 2021)

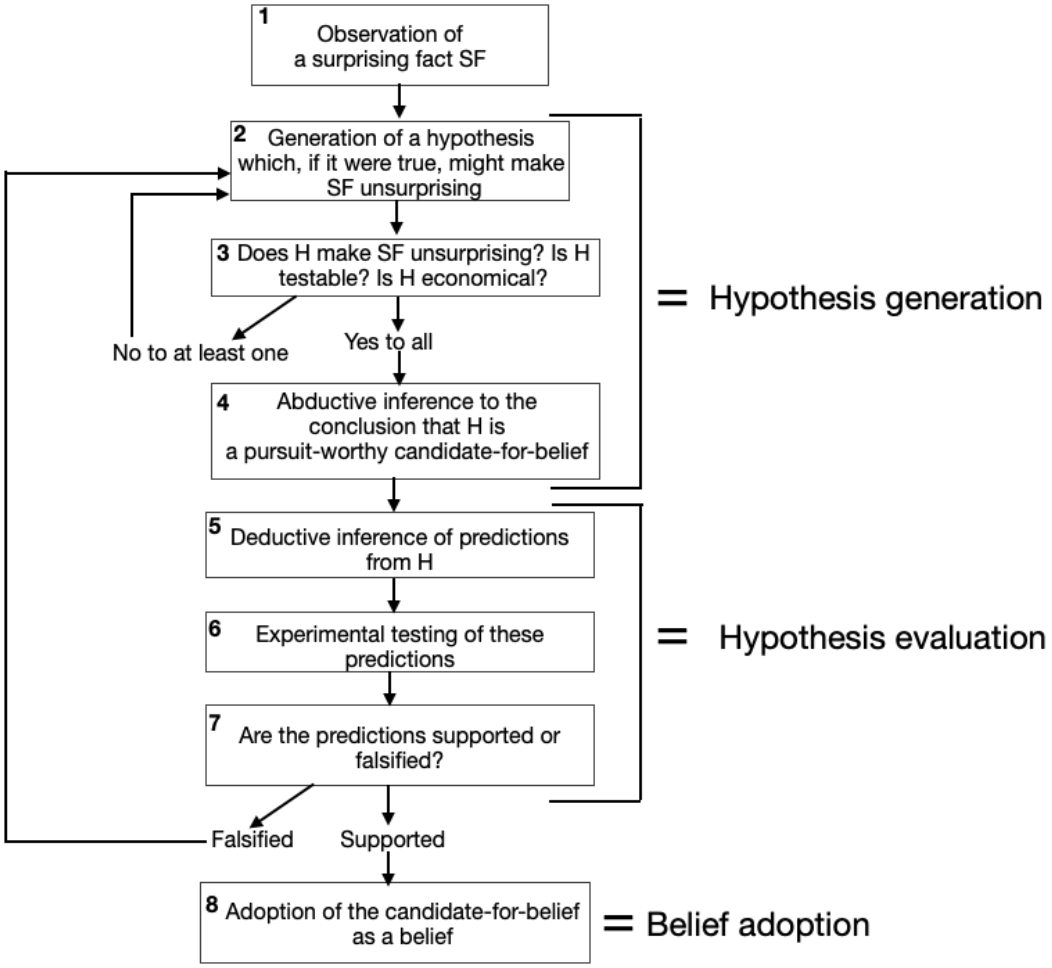

Figure 1: An 8-step model of the adoption of a new belief in response to the observation of a surprising fact (Davies and Coltheart, 2020; Coltheart and Davies, 2021)

Delusional condition	***Unexpected phenomenon***	***Explanatory hypothesis (delusional belief)***	***Nondelusional cases in which the unexpected phenomenon is present***
Capgras delusion (e.g., Edelstyn and Oyebode, 1999)	Failure of autonomic response to familiar faces (e.g., face of spouse).	This person I am looking at is a stranger, not my spouse.	Tranel, Damasio and Damasio (1995): patients with ventromedial frontal lesions lack autonomic responding to familiar faces but are not delusional.
Fregoli delusion (e.g., Langdon, Connaughton and Coltheart, 2014)	Presence of autonomic response even to unfamiliar faces.[8]	People with whom I am familiar are present in my environment, disguised.	Vuilleumier et al (2003): strong autonomic responses to unfamiliar faces (we presume) but no delusion.
Cotard delusion (e.g., Young, Robertson, Hellawell, de Pauw and Pentland, 1992)	Failure of autonomic response to any form of stimulus.[9]	I am dead.	"Pure autonomic failure" (Heims, Critchley, Dolan, Mathias and Cipolotti, 2004) is not always accompanied by delusion.
Mirrored-self misidentification (e.g., Breen, Caine, Coltheart, Roberts and Hendy, 2000: case FE)	Failure to recognise the face one sees when looking into a mirror.	The person I see when I look in the mirror is a stranger, not me.	Many patients with prosopagnosia are not delusional.
Mirrored-self misidentification (e.g., Breen, Caine, Coltheart, Roberts and Hendy, 2000: case TH)	Mirror agnosia present, so mirrors treated as windows. The seen person looks and behaves like a mirror image of the patient but appears to be in the space behind the glass.	The person I see when I look in the mirror is a stranger, not me.	Binkofski, Buccino, Dohle, Seitz and Freund (1999): mirror agnosia without delusion.
Passivity delusion ("alien control") (e.g., Stirling, Hellewell and Quraishi, 1998)	Failure of cancellation of feedback from motor response by efference copy.	Other people can cause my limbs to move without my volition.	In "haptic deafferentation", patient gets no sensory feedback from actions performed (Fourneret, Paillard, Lamarre, Cole and Jeannerod (2002). But no delusion present.

[8] Davies et al. (2001) adopted a suggestion by Ramachandran & Blakeslee (1998, p. 171). We note that Langdon et al. (2014) argue, against this suggestion, that "the Fregoli delusional content is generated when hyperexcitation from the cognitive system to the PINs [person identity nodes] causes a known person to be identified as present, even when no matching face is also present" (2014, p. 628).

[9] Hypothesised by Ramachandran & Blakeslee (1998, p. 167).

Somatoparaphrenia (e.g., Assal, 1983)	The unusual experience that results from paralysis and the loss of kinaesthetic and proprioceptive feedback from the arm.	This limb (the paralysed limb) is not mine, it is someone else's.	Many patients with a paralysed limb and without kinaesthetic and proprioceptive feedback are not delusional.

Table 1: Six types of delusional condition, the specific unexpected phenomenon associated with the delusion, the delusional belief which would explain this phenomenon, and cases where the specific unexpected phenomenon associated with each delusion is present in people who are nevertheless not delusional.