**Production version**
**To appear in KWM Fulford, M Davies, RGT Gipps, G Graham, J Sadler, G Stanghellini and T Thornton (eds),**
***The Oxford Handbook of Philosophy and Psychiatry* (Oxford: Oxford University Press, 2013).**

# Delusion: Cognitive approaches
# Bayesian inference and compartmentalisation*

MARTIN DAVIES AND ANDY EGAN

Delusions in individuals with schizophrenia are personal-level phenomena and no account of delusion could be complete unless it included a rich phenomenological description of individuals' experience of their delusions (Sass and Pienkos, this volume). Cognitive approaches aim to contribute to our understanding of delusions by providing an explanatory framework that extends beyond the personal level to the subpersonal level of information-processing systems (Cratsley and Samuels, this volume). At other subpersonal levels, contributions are also offered by neurobiological, neurocomputational, and psychopharmacological approaches.[1]

There are questions to be asked about the relationships between these different subpersonal levels (for example, about the relationship between the level of cognitive psychology and the level of neurobiology; Gold and Stoljar 1999). There are also questions about the relationship between the personal level, where description extends to phenomenology and normativity and where there are distinctive practices of rationalising explanation, and subpersonal levels of mechanistic description and explanation. According to one extreme view of this relationship, all that is literally true at the personal-level can be recast in the terms favoured by the sciences of the mind. According to the opposite extreme, what is distinctive and important at the personal level is independent from the sciences of the mind. But intermediate options, between reduction and independence, are available (inter-level interaction without reduction; Davies 2000). It is a familiar point that there seems to be an explanatory gap between the objective sciences of the mind and the subjective character of conscious experience (Nagel 1974; Levine 1983). But it would be an overreaction to this point to maintain that the science of colour vision, for example, could contribute nothing at all to our understanding of the normal or impaired experience – the 'what it is like' – of seeing colours. In a similar way, we can agree that phenomenological description is essential while maintaining that cognitive psychology and cognitive neuroscience can contribute to our understanding of

* Our debt to the papers by Max Coltheart, Peter Menzies, and John Sutton (2010) and Ryan McKay (2012) will be evident on almost every page of this chapter. MD also acknowledges intellectual debts to Max Coltheart extending over thirty years. AE would like to thank Adam Elga, in particular, for informing and improving his thinking about compartmentalisation and Spinozan belief-formation. We are grateful to Richard Gipps, Matthew Parrott and Nicholas Shea for comments on, and conversations about, an earlier version.

[1] The distinction between personal and subpersonal levels of description was introduced by Dennett (1969). The personal level of description is the folk psychological level at which we describe people as experiencing, thinking subjects and agents. It includes descriptions of conscious mental states as such and descriptions of normative requirements of rationality, for example. Personal-level descriptions figure in explanations in which actions are rationalised in terms of mental states such as beliefs and desires. Subpersonal levels, in contrast, are suited to the mechanistic descriptions and explanations of the objective sciences of the mind, such as information-processing psychology and neurobiology. (For discussion, see Davies 2000; Shea, this volume.)

personal-level phenomena, including personal-level pathologies (such as addiction or hearing voices; Shea, this volume).

Cognitive approaches to understanding delusions have focused first, not on the elaborated polythematic delusional systems or worlds of some individuals with schizophrenia, but on monothematic delusions – islands of delusion in a sea of apparent normality – and particularly, on monothematic delusions of neuropsychological origin. A starting point for understanding monothematic delusions is provided by Maher's (1974, 1988, 1992) anomalous experience hypothesis: a delusion arises as a normal response to an anomalous experience. The methodology of cognitive approaches has been that of cognitive neuropsychiatry (David 1993; Halligan and David 2001); that is, the application of the methods of cognitive neuropsychology to psychiatric disorders. Thus, it has been assumed that the anomalous experience that figures in Maher's proposed aetiology of delusions is the product of a neuropsychological deficit (Coltheart 2007, p. 1047):

> The patient has a neuropsychological deficit of a kind that could plausibly be related to the content of the patient's particular delusion – that is, a deficit that could plausibly be viewed as having prompted the initial thought that turned into a delusional belief.

Maher himself shares the assumption that '[t]he origins of anomalous experience lie in a broad band of neuropsychological anomalies' (1999, p. 551) but, as we see in the quotation from Coltheart, cognitive approaches allow that Maher's anomalous conscious experience might not always be essential. In some cases, the route from neuropsychological deficit to delusional belief might be wholly hidden from consciousness so that the delusional belief is 'the first delusion-relevant event of which the patient is aware' (Coltheart, Menzies and Sutton 2010, p. 264).

Because the neuropsychological deficit is supposed to be related to the content of the delusion in some plausible way, the type of deficit (and anomalous experience, if any) will vary from delusion to delusion. The type of deficit may also vary between individuals with the same delusion if different deficits can prompt the same 'initial thought'. Thus, cognitive approaches need to document types of neuropsychological deficit (and perhaps also types of experience) that could plausibly give rise to each of a variety of monothematic delusions, such as: Capgras delusion – 'This [the subject's wife] is not my wife. My wife has been replaced by an impostor' (Capgras and Reboul-Lachaux 1923; Edelstyn and Oyebode 1999), Cotard delusion – 'I am dead' (Cotard 1882; Young and Leafhead 1996), Fregoli delusion – 'I am being followed around by people who are known to me but who are unrecognisable because they are in disguise' (Courbon and Fail 1927; de Pauw, Szulecka and Poltock 1987; Ellis, Whitley and Luauté 1994), mirrored-self misidentification – 'The person I see in the mirror is not really me' (Breen et al. 2000; Breen, Caine and Coltheart 2001), somatoparaphrenia – 'This [the subject's left arm] is not my arm' (Halligan, Marshall and Wade 1995; Bottini et al. 2002), and the delusion of alien control – 'Other people can control the movements of my body' (Frith and Done 1989; Frith 1992).

Candidate neuropsychological deficits have been proposed as factors in the aetiology of each of these delusions and others. But, in each case, there are examples of individuals who have the proposed deficit but not the delusion. The conclusion that is drawn from

this dissociation is that there must be some additional factor or factors implicated in the aetiology of delusions. In this chapter, we shall be concerned with the *two-factor* cognitive neuropsychological approach to understanding delusions (for early expositions, see e.g. Davies and Coltheart 2000; Langdon and Coltheart 2000; Davies et al. 2001; for recent reviews, see e.g. Aimola Davies and Davies 2009; Coltheart 2007, 2010; Coltheart, Langdon and McKay 2011; McKay 2012).

Before moving on, we shall illustrate the proposal of a neuropsychological deficit as a first factor, and the dissociation argument for a second factor, in the widely discussed case of Capgras delusion. We shall draw on important early work in cognitive neuropsychiatry (Ellis and Young 1990), which in turn built on a well-supported model of normal face processing (Bruce and Young 1986). In the Bruce and Young model, information about known faces is stored in face recognition units (FRUs), one for each known face. When a known face is seen, one FRU will be activated to a high level and biographical information stored in a corresponding personal identity node (PIN) – such as information about the person's occupation – will be accessed, as will the person's name. An important functional difference between FRUs and PINs is that only a seen face will activate an FRU, whereas a PIN can be accessed from the person's seen face or heard voice, or in other ways.

In some individuals with severely impaired face recognition (prosopagnosia), skin conductance responses continue to discriminate between familiar and unfamiliar faces – there is covert recognition (Tranel and Damasio 1985, 1988; see also Bauer 1984). So, although the primary face-recognition system is damaged in these individuals, there must be a preserved connection between an early stage of face processing (the FRUs) and the autonomic nervous system. Ellis and Young (1990) proposed that the neuropsychological deficit in Capgras delusion is the mirror image of the deficit in prosopagnosia with covert recognition. In Capgras delusion, the primary face-recognition system is intact but the connection between the FRUs and the autonomic nervous system is damaged. Normally, the seen face of a loved one, such as the spouse, causes activity in the autonomic nervous system and the experience of the loved one's face has a strong affective component. But now, with the connection between the face-recognition system and the autonomic nervous system disrupted, this component of the experience is missing.

Ellis and Young's proposal about the neuropsychological deficit and anomalous experience in Capgras delusion made a clear empirical prediction that the skin conductance responses of individuals with Capgras delusion would not discriminate between familiar and unfamiliar faces. This prediction was subsequently confirmed in four studies using photographs of familiar (famous or family) faces and unfamiliar faces (Ellis et al. 1997, 2000; Hirstein and Ramachandran 1997; Brighetti et al. 2007) – an 'exemplary vindication' of the new discipline of cognitive neuropsychiatry (Ellis 1998). Thus, it is plausible that a neuropsychological deficit, disconnection of the primary face-recognition system from the autonomic nervous system, is a factor in the aetiology of Capgras delusion. But there is a dissociation between this deficit and the delusion. There are individuals (patients with damage to ventromedial regions of frontal cortex; Tranel, Damasio and Damasio 1995) whose skin conductance responses do not discriminate between familiar and unfamiliar faces, but who do not have Capgras delusion (or any other delusion). There is also a report of an individual who (following temporal lobe

surgery for relief of epilepsy) had an anomalous 'Capgras-like' experience of her mother – 'she was different, something was different about her … you can look different by, you know, doing your hair or whatever, but it wasn't different in that way … it didn't feel like her' (Turner and Coltheart 2010, pp. 371–2) – but did not have the Capgras delusion.[2] Thus there must be a second factor in the aetiology of Capgras delusion – presumably, in cases of neuropsychological origin, a second deficit.

## 1. The two-factor framework for explaining delusions

Coltheart (2007 p. 1044) has proposed that, in order to explain any delusion, we need to answer two questions. First, where did the delusion come from? Second, why does the patient not reject the belief? The leading idea of the two-factor framework for explaining delusions (Coltheart 2007, 2010; Coltheart et al. 2011) is that the two factors will provide answers to these two questions.

> The first question is always: *where did the delusion come from?* – that is, what is responsible for the *content* of the delusional belief? The second question is always: *why does the patient not reject the belief?* … – that is, what is responsible for the *persistence* of the belief? (Coltheart 2007 p. 1044)
>
> Factor 1 is what is responsible for the belief having occurred to the person in the first place … : this factor determines the *content* of the delusional belief. Factor 2 is responsible for the failure to reject the hypothesis despite the presence of (often overwhelming) evidence against it … this factor determines the *persistence* of the delusional belief. (Coltheart 2010, p. 18)
>
> 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)

The two-factor framework has provided explanations (at least in outline) of a range of delusions including those that we have already mentioned and also anosognosia for motor impairments (Davies, Aimola Davies and Coltheart 2005; Aimola Davies et al. 2009, Aimola Davies and Davies 2009). It has been proposed that the second factor is the same in all cases of delusion (or at least in all cases of delusion of neuropsychological origin) and that it consists in an impairment of normal processes of belief evaluation, associated with pathology of right lateral prefrontal cortex (e.g. Coltheart et al. 2011, p. 285). The nature of the putative task of belief evaluation suggests that the second factor could be an impairment of executive function or working memory (or both), consistent with its proposed neural basis (Aimola Davies and Davies 2009). But the cognitive nature and neural basis of the second factor have not been specified as precisely as the nature and basis of putative first factors.

[2] This patient was studied by Nora Breen and Mike Salzberg.

1.1 Adoption and persistence: Two options for the two-factor framework

It is important to notice that the three quotations listed earlier leave open two possible interpretations of the second question (that is, the question to which the second factor is supposed to provide an answer). In the third quotation, the second question is about *adoption* of the delusional belief. But it is possible initially to adopt a belief and then, on reflection, to reject it. Sometimes we initially believe what we see and then realise that we are subject to an illusion or we initially believe what we are told and then realise that our informant is unreliable. In the case of a delusional belief we can ask why the belief, once adopted, is not subsequently rejected. Why is it 'firmly sustained' (American Psychiatric Association 2000, p. 821), why does it *persist*? This is more like the version of the second question that is posed in the first two quotations.

Explaining a delusion requires answers to both the adoption question and the persistence question. In principle, it might turn out that two factors (two pathologies or departures from normality) are needed to answer the adoption question and that a third factor is needed to answer the persistence question. That would, of course, be incompatible with the two-factor framework, but compatible with a less specific multi-factor framework. An exactly-two-factor account must say either:

> (A) that no pathology or departure from normality beyond the first factor is needed to answer the adoption question and the second factor answers the persistence question;

or else:

> (B) that two factors are needed to answer the adoption question and no additional pathology or departure from normality is needed to answer the persistence question.

(In principle, it might be that option (A) is correct for some delusions and option (B) for others.)

According to option (A), we should expect each dissociation of the 'first deficit without delusion' form (e.g. ventromedial frontal damage without Capgras delusion; Tranel et al. 1995) to be a case in which the delusional belief is initially adopted, but does not persist. According to option (B), we should expect each dissociation to be a case in which the first deficit is present, but the delusional belief is not even initially adopted. This presents a potential problem for option (A) because there is no evidence that patients with ventromedial frontal damage, for example, initially adopt the Capgras delusion but subsequently reject it (nor is there evidence that this is not the case; see Coltheart et al. 2010, p. 281 and McKay 2012, pp. 341–2, for discussion).

On the other hand, there are reports of individuals who, after recovering from a delusion, still feel the attraction of the belief that they now reject. For example, a patient (HS) who had recovered from anosognosia reported that the idea that he could move his paralysed limbs still seemed credible even though he was able to reject it (Chatterjee and Mennemeier 1996, p. 227):

> E: What was the consequence of the stroke?
>
> HS: The left hand here is dead and the left leg was pretty much.

> HS: (later): I still feel as if when I am in a room and I have to get up and go walking . . . I just feel like I should be able to.
>
> E: You have a belief that you could actually do that?
>
> HS: I do not have a belief, just the exact opposite. I just have the feeling that sometimes I feel like I can get up and do something and I have to tell myself 'no I can't'.

Turner and Coltheart describe a patient in the early stages of recovery from Capgras delusion (2010, p. 371):

> I've started going through it, and seeing what could possibly happen and what couldn't happen. That was wrong, that couldn't happen. Even though it has happened it couldn't. Mary couldn't suddenly disappear from the room, so there must be an explanation for it. … And then I worked it out and I've wondered if it's Mary all the time. It's nobody else.

In summary, the standard examples of deficit without delusion, which figure in the dissociation argument for a second factor, are potentially problematic for option (A) and fit option (B) better. But the examples of recovery from delusion fit option (A) well, on the assumption that the recovery resulted from remission of the second factor. Thus, not only the cognitive nature and neural basis of the second factor, but also – and even more importantly – its role in the aetiology of delusions, requires further specification.[3]

### 1.2 Bayesian approaches

One of the aims of cognitive neuropsychology is to understand disorders of cognition in terms of theories or models of normal cognition. Cognitive impairments are understood in terms of damage to one or more components of the normal cognitive system. When the methods of cognitive neuropsychology are applied to delusions – pathologies of belief – what is required is an information-processing model of the normal formation, evaluation, and revision of beliefs. Thus, one of the problems faced by cognitive neuropsychiatry – in comparison with the cognitive neuropsychology of face recognition, for example – is that we do not have an articulated, still less a computationally implemented, model of normal believing. Indeed, there may be reasons of principle why it is difficult to understand believing in terms of the computational theory of mind (Fodor 1983, 2000; see Cratsley and Samuels, this volume, on Fodorian pessimism).

More than twenty-five years ago, Hemsley and Garety suggested a strategy for making progress in the absence of a model of normal believing (1986, p. 52): 'A normative theory of how people *should* evaluate evidence relevant to their beliefs can

[3] In this chapter, we shall be defending a version of option (A), but we do not offer a resolution of the potential problem associated with option (A). In the specific case of the ventromedial frontal patients, it might be suggested that they do not initially adopt the Cagras delusion because they do not, in fact, have exactly the same neuropsychological deficit as Capgras patients (see Ellis and Lewis 2001). That suggestion provides a response to the potential problem for option (A) but at the price of removing the standard dissociation argument for a second factor.

provide a conceptual framework for a consideration of how they do *in fact* evaluate it.' Their specific proposal was to begin from a probabilistic analysis of hypothesis evaluation and then to investigate whether individuals with delusions deviate from the normative Bayesian model. In pursuing this strategy and interpreting its results, it is important to distinguish the normative from the normal; it is important not to forget that, as Hemsley and Garety put it, there is '"normal" deviation from the prescriptive model' (p. 55).

Recently, the Bayesian approach has been married with the neuropsychological deficit approach in continuing development of the two-factor framework for explaining monothematic delusions (Coltheart et al. 2010; McKay 2012). A second body of work has adopted a Bayesian approach – and, specifically, the theoretical framework of predictive coding and prediction error signals, in which neural processing aims to minimise prediction error or 'free energy' (Friston 2005, 2009, 2010; Friston and Stephan 2007) – to delusions in schizophrenia (Corlett, Frith and Fletcher 2009; Fletcher and Frith 2009).[4] In this chapter, we shall focus on the Bayesian two-factor approach to explaining monothematic delusions and on the idea that delusions arise through a process of Bayesian inference or updating.

## 1.3 Bayesian inference

On a Bayesian approach, probabilities are updated on the basis of evidence, E, so that the new or *posterior* probability of a hypothesis, H, is equal to the old or *prior* conditional probability of H given E. This updating procedure is known as *simple conditionalisation*. By Bayes' theorem, the conditional probability, $P(H|E)$, can be further unpacked to give:

**Simple conditionalisation**

$$P'(H) = P(H|E) = P(H).\ P(E|H)/P(E).$$

(Here, $P'$ is the new distribution of probabilities.) The notions of prior and posterior probabilities are relative. The prior probability of H is prior only to the evidence E; it already takes account of antecedently available evidence – today's priors are yesterday's posteriors. In simple conditionalisation, the evidence is treated as certain: $P'(E) = 1$. A more general updating procedure, Jeffrey conditionalisation (Jeffrey 1983), allows that the evidence may be less than certain, so that $P'(E) < 1$.[5]

The posterior probability of a hypothesis, H, updated on the basis of evidence E by simple conditionalisation, is proportional to the prior probability of H, $P(H)$, and to the probability of E given H, $P(E|H)$, also known as the *likelihood* of H on E. The likelihood provides a measure of how well H predicts E. In this chapter, we shall usually be more interested in the balance of probabilities between two competing hypotheses than in the

[4] See Frith 2007, chapters 4 and 5, for an accessible introduction to the predictive coding approach and Corlett et al. (2007, 2009) for prediction error and delusion. See also Shea, this volume, for discussion of prediction error signals.

[5] In Jeffrey conditionalisation, the probability of H is updated to:

$$P'(H) = P(H|E).\ P'(E) + P(H|\text{not-}E).\ P'(\text{not-}E).$$

precise probability of each hypothesis. If we are considering two hypotheses, $H_1$ and $H_2$, then the ratio of posterior probabilities (the posterior odds) is the product of two other ratios, the ratio of prior probabilities (the prior odds) and the likelihood ratio:

**Bayes Ratio Formula**

$$\frac{P'(H_1)}{P'(H_2)} = \frac{P(H_1|E)}{P(H_2|E)} = \frac{P(H_1)}{P(H_2)} . \frac{P(E|H_1)}{P(E|H_2)}$$

Thus, on a Bayesian approach, the balance of probabilities between two candidate hypotheses, updated on the basis of evidence E, depends on (*a*) how probable each hypothesis is in the light of available evidence other than E – given by the prior probability $P(H_i)$ – and (*b*) how well each hypothesis predicts the evidence – given by the likelihood $P(E|H_i)$.

In the next two sections, we shall review two versions of the Bayesian two-factor approach (Coltheart et al. 2010; McKay 2012) in some detail. Section 2 is about the initial adoption of a delusional belief and section 3 is about the persistence of the belief. Before moving on, however, we note that there are complex and difficult issues surrounding the relationship between, on the one hand, Bayesian inference or updating and, on the other hand, abductive inference or inference to the best explanation (Lipton 2004).

Coltheart et al. (2010) sketch two models of abductive inference, the logical empiricist model based on an understanding of explanation as logical implication and the Bayesian model based on a probabilistic account of explanation (p. 271): 'the hypothesis H explains observations O to the degree *x* just in case the probability of O given H is *x*'. They adopt a Bayesian model of abduction, but we are not committed to the view that Bayesian inference is a model of inference to the best explanation. One reason is that the likelihood, $P(E|H)$, is not in general a good measure of how well a hypothesis H explains evidence E. A hypothesis about barometer readings (e.g. the barometer is falling) does not explain weather patterns (e.g. a storm is coming), however high the likelihood (that is, the probability of the weather patterns given the barometer readings) may be. Rather, causation and explanation run in the opposite direction, from weather patterns to barometer readings (van Fraassen 1980, p. 104).

More generally, Lipton's account of inference to the best explanation takes account, not only of whether a candidate explanatory hypothesis is the most probable given the available evidence, but also of whether it exhibits explanatory virtues such as parsimony, scope, depth, unifying disparate phenomena, and making new predictions. The question then arises whether it could ever be rational to accept an explanation because of its virtues, if an alternative explanation was more probable (van Fraassen 1989). Lipton (2004) aims to neutralise this concern about the relationship between inference to the best explanation and Bayesianism by suggesting that explanatory virtues are a guide to probability, but we take no stand on that issue.

In our discussion of delusions and Bayesian inference, it will be the standard Bayesian apparatus of probability assignments, likelihoods, and updating that bears the theoretical load. The notion of explanatory virtue will play only a peripheral role, in that it may influence the psychological accessibility of hypotheses. It is true that some of the literature that we shall engage with is couched in terms of inference to the best

explanation. But this seems to be largely inessential and, because of the issues that we have just mentioned, potentially distracting. Most or all of the theoretical work in explaining the adoption and persistence of delusional beliefs in terms of Bayesian *abductive* inference could be done just as well by talking about Bayesian inference *simpliciter*, thereby sidestepping those complex and difficult issues.

## 2. Bayes in the two-factor framework: Adoption of the delusional belief

Coltheart and colleagues (2010) propose that the answer, in outline, to the question where a delusion came from is that it arose through a process of Bayesian inference. In principle, this might be a process of inference carried out consciously by the person with the delusion, but Coltheart and colleagues focus on the case of unconscious inferential processes. To illustrate their approach, Coltheart and colleagues provide a worked example of how Bayesian inference could lead from a neuropsychological deficit to the initial onset of Capgras delusion.

### 2.1 From deficit to delusional belief: Capgras delusion

Suppose that, as the result of a stroke, a patient suffers disconnection of the primary face processing system from the autonomic nervous system (while the two disconnected systems themselves remain intact). Before the patient suffered the stroke, the appearance of his wife caused activation of the face recognition unit for the wife ($FRU_W$), which normally led to activation, not only of the corresponding personal identity node ($PIN_W$), but also of the autonomic nervous system. As a result of the learned association between the appearance of the patient's wife and activation of his autonomic nervous system, the appearance of his wife generated an unconscious prediction of activity in the autonomic nervous system, and this prediction was reliably fulfilled. Following the stroke, some things remain the same and some things are different. When the patient sees his wife, the face recognition unit $FRU_W$ and the personal identity node $PIN_W$ are still activated, and activity in the autonomic nervous system is still predicted. But, because of the disconnection, the prediction is not fulfilled. The abnormal absence of the predicted autonomic activity, resulting from the neuropsychological deficit (disconnection), stands in need of explanation.

The aim of the Bayesian approach is to show that a delusional hypothesis may be initially adopted as a belief as a result of Bayesian inference or updating on the basis of abnormal data, D (in this case, the absence of activity in the autonomic nervous system). Consequently, the next stage of Coltheart and colleagues' (2010) worked example involves two competing hypotheses. One is the true hypothesis, $H_W$, that the woman that the patient sees in front of him, who looks like the patient's wife and says that she is the patient's wife is, indeed, his wife. The other is the delusional hypothesis, $H_S$, that the woman is not the patient's wife but a stranger.[6]

[6] Coltheart et al. (2010) do not consider hypotheses that are incompatible with both $H_W$ and $H_S$, such as the hypothesis $H_A$, that the person that the patient sees in front of him is aunt Agatha, or the hypothesis $H_B$, that the person that the patient sees in front of him is Bob the bank teller. As a result of this simplification, $H_S$ is treated as the negation of $H_W$.

What needs to be shown is that the ratio of posterior probabilities, $P(H_S|D) / P(H_W|D)$ could favour the stranger hypothesis, $H_S$, over the wife hypothesis, $H_W$. The Bayes ratio formula tells us that this ratio is equal to the product of the ratio of prior probabilities and the likelihood ratio. So, how might the balance between those two ratios favour $H_S$ over $H_W$? The prior probabilities, $P(H_W)$ and $P(H_S)$, are prior only to the to-be-explained abnormal data D. They take account of antecedently available evidence including, in particular, the evidence that the woman that the patient sees in front of him looks just like his wife and says that she is his wife. Consequently, the probability that the woman is the patient's wife is much higher than the probability that she is a stranger and the ratio $P(H_S) / P(H_W)$ is correspondingly low. In contrast, Coltheart and colleagues say, the likelihood ratio, $P(D|H_S) / P(D|H_W)$ is high: 'It would be highly improbable for the subject to have the low autonomic response [D] if the person really was his wife, but very probable indeed if the person were a stranger' (2010, p. 277).

According to the worked example, then, the ratio of prior probabilities favours the wife hypothesis, $H_W$, but the likelihood ratio favours the stranger hypothesis, $H_S$. Whether the ratio of posterior probabilities favours $H_W$ or $H_S$ depends on the relative values of these ratios and, specifically, on whether the likelihood ratio is sufficient to outweigh the ratio of prior probabilities. Suppose, for example, that the prior probabilities favoured $H_W$ in the ratio 100:1 but the likelihoods favoured $H_S$ in the ratio 1000:1. Then the posterior probabilities would favour $H_S$ in the ratio 10:1. If these were the only two hypotheses to consider, their probabilities would be $P(H_S) = 0.91$ and $P(H_W) = 0.09$. Thus, Bayesian inference might lead from the abnormal data D to the assignment of a high probability to the hypothesis that the woman who looks just like the patient's wife and also claims to be the patient's wife is not his wife but a stranger, and so an impostor. Equally, if the prior probabilities favoured $H_W$ in the ratio 100:1 but the likelihoods favoured $H_S$ only in the ratio 10:1, then the posterior probabilities would be reversed: $P(H_S) = 0.09$ and $P(H_W) = 0.91$.

Coltheart et al. suggest that the likelihood ratio does outweigh the ratio of prior probabilities (2010, p. 278):

> The delusional hypothesis provides a much more convincing explanation of the highly unusual data than the nondelusional hypothesis; and this fact swamps the general implausibility of the delusional hypothesis. So if the subject with Capgras delusion unconsciously reasons in this way, he has up to this point committed no mistake of rationality on the Bayesian model.[7]

One difficulty in evaluating this suggestion about Bayesian inference is that it is somewhat unclear which probabilities are to figure in the worked example. Are they supposed to be, for example, realistic probabilities or the patient's subjective probabilities

[7] Coltheart and colleagues move from the fact that the delusional hypothesis, $H_S$, provides a much more convincing explanation of the data than the wife hypothesis, $H_W$, does to the claim that the likelihood ratio strongly favours $H_S$. It is worth noting, however, that explanatoriness is not always a good indicator of likelihood. From the fact that a hypothesis, H, utterly fails to explain evidence, E, it does not follow that the likelihood, $P(D|H)$, is close to zero.

(credences)? As we shall see later (section 5.3), there are problems with each of these options. But the main point here is that if Coltheart and colleagues' suggestion is correct then the disconnection of the patient's face processing system from the autonomic nervous system is the only pathology or departure from normality that need be implicated in the processes leading up to the initial onset of Capgras delusion. Consequently, the role for a second factor in the aetiology of the delusion must lie beyond that point, in an impairment of post-onset belief evaluation that explains the persistence of the delusion. This is option (A) (see section 1.1).

### 2.2 Biased Bayesian inference

McKay (2012) challenges Coltheart and colleagues' (2010) description of the unconscious inferential processes that lead up to the initial onset of the delusional belief. The general point that Coltheart and colleagues make is that the superior explanatory (better: predictive) potential of the stranger hypothesis can outweigh the prior odds in favour of the wife hypothesis. McKay objects that the prior probability of the stranger hypothesis in their worked example ($P(H_S) = 0.01$) is 'unrealistically high' (2012, p. 339). A more realistic estimate would give a ratio of prior probabilities that would much more strongly favour the wife hypothesis and would be much more difficult for the likelihood ratio to outweigh.[8]

The exact value of $P(H_S)$ is, of course, a point of detail. But the moral that McKay (p. 340) draws is that, with a realistic distribution of prior probabilities, unbiased Bayesian inference would not lead from abnormal data to the onset of delusional belief. Specifically, the likelihood ratio in Coltheart and colleagues' worked example (999:1 in favour of $H_S$) would be insufficient to outweigh a realistic ratio of prior probabilities favouring $H_W$. Unbiased updating of probabilities gives weight to both the prior probabilities and the likelihoods of competing hypotheses. Actual updating departs from the normative Bayesian model if it discounts either of these components in the Bayes ratio formula. McKay proposes that, if updating is to lead to the onset of Capgras delusion, there must be a departure from the normative Bayesian model. If posterior probabilities are to favour $H_S$ over $H_W$ then prior probabilities must be discounted to increase the relative weight given to likelihoods.[9]

---

[8] McKay points out, in effect, that if Coltheart et al.'s (2010) proposed prior probability, $P(H_S) = 0.01$, were correct then one should expect that one time in a hundred, an encounter with a person who looked just like one's spouse and claimed to be one's spouse would be an encounter with a stranger (impostor). He suggests that a more realistic prior probability for $H_S$ would be two orders of magnitude lower; that is, of the order of 0.0001. (McKay's proposal is $P(H_S) = 0.00027$.)

[9] In a case of comparison between two hypotheses, if the ratio of prior probabilities were 1:1 then the ratio of posterior probabilities would be equal to the likelihood ratio. So an updating bias in favour of likelihoods can be implemented by treating the ratio of prior probabilities as being closer to 1:1 than it really is. This could be achieved by treating the absolute magnitude of the logarithm of the prior odds as being closer to zero than it really is (by subtraction, so that the logarithm would approach zero linearly, or by division, so that the logarithm would approach zero asymptotically). Conversely, a reduction in the absolute magnitude of the logarithm of the likelihood ratio would implement an updating bias in favour of prior probabilities. See McKay (2012, pp. 350–2) for a different, and more general, way of implementing these biases.

Once again, it is difficult to evaluate this proposal because it is unclear how we are to understand the probabilities that are supposed to figure in the worked examples. But we set aside issues of interpretation for the time being (see section 5.3) in order to highlight a clear and important point of contrast between McKay (2012) and Coltheart et al. (2010). According to McKay, the disconnection of the patient's face processing system from the autonomic nervous system is *not* the only pathology or departure from normality that is implicated in the processes leading up to the onset of Capgras delusion (McKay 2012, p. 345): 'the second factor in delusion formation comprises a bias towards explanatory adequacy'. Here, we should recall that a bias or other deviation from the normative Bayesian model might be normal (Hemsley and Garety 1986). But if McKay's proposed updating bias does amount to a departure from normality then, in order for his account to remain within the exactly-two-factor framework, no additional pathology should be needed to answer the persistence question. This is option (B) (see section 1.1).

## 3. Bayes in the two-factor framework: Persistence of the delusional belief

The Capgras patient believes that the woman who looks like his wife and says that she is his wife is really a stranger; his wife has been replaced by an impostor.[10] We have been considering two Bayesian answers to the question how this belief came to be adopted. Now we turn to the question why, given that this belief has been initially adopted, it is not subsequently rejected when evidence and implausibility count against it. One possible form of answer to this question is that there is a separate impairment of post-adoption belief evaluation – as in option (A). The other is that no additional pathology or departure from normality is needed to explain the delusional belief's persistence once it has been adopted – as in option (B).

### 3.1 Two accounts of persistence

Coltheart and colleagues (2010) envisage that, after adopting the delusional belief, the patient will be presented with new evidence, N, that is better explained (predicted) by the wife hypothesis than by the stranger hypothesis. Friends and relatives assure the patient that the woman he believes to be an impostor is really his wife, the woman knows many things about the patient's life, and she arrives home from his wife's place of employment, driving his wife's car, wearing his wife's clothes, and carrying his wife's briefcase with his wife's initials in gold lettering.

So there should be a new round of Bayesian updating of probabilities. The new prior probabilities are $P'(H_S) = P(H_S|D)$, and $P'(H_W) = P(H_W|D)$ and, because the delusional belief has been adopted, the ratio of new prior probabilities favours $H_S$ (perhaps in the ratio 10:1). In contrast, the likelihood ratio, $P'(N|H_S) / P'(N|H_W)$, favours $H_W$. For

[10] The proposition, $H_S$, that the woman that the patient sees in front of him is a stranger is strictly speaking distinct from the proposition that the woman is an impostor – that is, a stranger who looks like the patient's wife and says that she is the patient's wife. But since probabilities have already been updated to take into account the evidence that the woman that the patient sees in front of him looks like his wife and says that she is his wife, the stranger hypothesis and the impostor hypothesis have the same probability and we can elide the distinction between them.

example, friends and relatives would be more likely to give such assurances if the woman were the patient's wife than if she were a stranger. If the likelihood ratio is sufficient to outweigh the ratio of prior probabilities then, according to the normative model, the ratio of posterior probabilities, $P'(H_S|N) / P'(H_W|N)$, should favour $H_W$ over $H_S$ and the delusional belief should be rejected on the basis of the new evidence N. However, the delusion persists. According to Coltheart and colleagues (2010), this persistence is to be explained by the second factor, which is some kind of impairment – to be specified – in the evaluation of already adopted beliefs in the light of new evidence.

It is crucial to this account of persistence as being explained by a second factor that, according to the normative Bayesian model, the delusional belief, $H_S$, *should* be rejected on the basis of the new evidence N. This, in turn, depends on the claim that the likelihood ratio, $P'(N|H_S) / P'(N|H_W)$, favours $H_W$ strongly enough to outweigh the ratio of prior probabilities in favour of $H_S$. There is a potential problem for the account here. An estimate of the likelihood ratio, and particularly an estimate of $P'(N|H_S)$, depends on what exactly the new evidence N is taken to be.

On a Bayesian approach, evidence – such as the new evidence, N – is a proposition. Suppose, for example, that a trusted friend of the patient asserts that the woman who looks just like the patient's wife *is* the patient's wife. Is the evidence, N, the proposition *that the woman is the patient's wife* (essentially, the proposition $H_W$) or the proposition *that the trusted friend asserted that the woman is the patient's wife*? On the first (more committed) interpretation, the likelihood, $P'(N|H_S)$, is low; in fact it is zero (or very close to zero). But on the second (more cautious) interpretation, $P'(N|H_S)$ might be quite high and almost as high as $P'(N|H_W)$: even a trusted friend may be taken in by a very good impostor.[11] The potential problem for Coltheart and colleagues' explanation of persistence is that, on the more cautious interpretation of what the evidence is, the likelihood ratio might not favour $H_W$ strongly enough to outweigh the ratio of prior probabilities in favour of $H_S$. (We shall consider the more committed interpretation shortly; see section 3.3.) If the ratio of posterior probabilities still favours $H_S$ rather than $H_W$ then, of course, there is no need for a second factor (a pathology or departure from normality) to explain the persistence of the delusional belief.

On McKay's (2012) account, in contrast to that of Coltheart et al. (2010), there are already two factors involved in the answer to the question how the delusional belief came to be initially adopted (the neuropsychological deficit and the updating bias in favour of likelihoods). But we still need an answer to the question why the belief is not

[11] For some other pieces of new evidence, there will be similar, though less dramatic, differences in likelihood estimates depending on what exactly the evidence is taken to be. Suppose, for example, that the woman is carrying what is in fact the wife's briefcase. Is the evidence, N, the proposition *that the woman is carrying the wife's briefcase* or the proposition *that the woman is carrying a briefcase that looks just like the wife's*? However, the distinction between more committed and more cautious interpretations of what the evidence is might not extend smoothly to all pieces of new evidence (e.g. the fact that the woman is able to give correct accounts of many past events involving the patient and his wife). Different kinds of new evidence require more discussion than they can receive here.

subsequently rejected when there is so much evidence against it? Is a *third* factor needed in order to explain the persistence of the delusion?

Our discussion (two paragraphs back) of Coltheart and colleagues' (2010) account suggests that there may be no need for an additional factor to explain persistence and this is, indeed, McKay's (2012) view. He agrees that, after the initial adoption of the delusional belief, there will be new evidence that counts in favour of the wife hypothesis and against the stranger hypothesis. But he points out that much of this new evidence (e.g. trusted friends say that the woman is the patient's wife) is similar in kind to evidence that was taken into account in the previous prior probabilities of the competing hypotheses (e.g. the woman herself says that she is the patient's wife). The patient's evidential situation remains very similar to the situation that led to the initial adoption of the delusional belief and so the same processes of updating (processes biased in favour of likelihoods – McKay's second factor) will lead to similar posterior probabilities. Thus, McKay says (2012, p. 343): 'if the neuropsychological data that stimulate the stranger hypothesis are so salient that they can compensate for the low prior probability of that hypothesis, then supplementary testimony from friends and clinicians will be powerless to overwhelm those data'.

### 3.2 Diagnostic evidence, similar evidence, and 'unfalsifiability'

Consider just a subset of the evidence against the Capgras patient's delusional belief: the evidence that the woman that the patient sees in front of him looks just like his wife and says that she is his wife. This evidence is predicted by the patient's impostor hypothesis no less than by the true hypothesis, $H_W$. The evidence is not *diagnostic* between the two hypotheses – it cannot shift the prior balance of probabilities towards $H_W$, because the likelihood ratio is 1:1. In this limited way, the Capgras patient's hypothesis is similar to sceptical hypotheses (such as Descartes's evil demon hypothesis) in being 'unfalsifiable'. The evil demon hypothesis, or the 'brain in a vat' hypothesis, is not falsified by the evidence of my senses, 'Here is one hand, and here is another' (Moore 1939), because a perceptual experience as of hands is predicted by the sceptical hypothesis no less than by the external world hypothesis. (It is important to note that, on a Bayesian approach, this unfalsifiability depends on the evidence being the cautious proposition *that I am having a perceptual experience as of hands* rather than the committed proposition *that I have hands*.)

It is only a small set of evidence (only the evidence that the woman looks just like the patient's wife and says that she is his wife) that is utterly unable to shift the balance of probabilities towards $H_W$. But, as we have just seen, the value of a larger set of evidence is diminished because accepted evidence makes other similar evidence more probable. It is improbable that a trusted friend should assert, concerning a stranger, that she is the patient's wife. But it is not so improbable that a trusted friend should assert, concerning a stranger who looks just like the patient's wife and says that she is his wife (an impostor, and a good impostor at that), that she is the patient's wife. So, although the evidence of the trusted friend's testimony is somewhat better predicted by the wife hypothesis, $H_W$, it does not favour $H_W$ very strongly – arguably not strongly enough to outweigh the ratio of prior probabilities in favour of $H_S$. It is for this reason that McKay's (2012) account and, equally, our reservations about Coltheart et al.'s (2010) account, seem plausible. It is

difficult to identify a role for a pathology or departure from normality that would distinctively explain the persistence of a delusional belief after its initial adoption had been accounted for – that is, a role for a second factor according to option (A).

Coltheart and colleagues' view is, of course, that they have identified such a role. So let us consider their example (2010, p. 279) again.[12]

### 3.3 Rejecting evidence

The patient has adopted the delusional belief and the ratio of new prior probabilities, $P'(H_S) / P'(H_W)$, favours $H_S$. But now, new evidence N is presented, and the likelihood ratio, $P'(N|H_S) / P'(N|H_W)$, favours $H_W$. If the likelihood ratio is sufficient to outweigh the ratio of prior probabilities then the delusional belief should be rejected – but it persists. We pointed to a potential problem for Coltheart and colleagues' explanation of persistence in terms of a second factor. On the more cautious interpretation of which proposition the evidence is, the likelihood ratio might not favour $H_W$ strongly enough to outweigh the ratio of prior probabilities. In fact, we suggested that the likelihood $P'(N|H_S)$ might be only slightly lower than $P'(N|H_W)$.

Coltheart and colleagues (2010) go on to say that, before the new evidence is actually presented, the patient should attach a low credence to the evidence proposition. In the normative model, the prior probability of N is given by:

$$P'(N) = P'(N|H_S) \,.\, P'(H_S) + P'(N|H_W) \,.\, P'(H_W).$$[13]

If, as Coltheart et al. say, $P'(N|H_S)$ and $P'(H_W)$ are both low, then $P'(N)$ is also low: the patient should not expect N to be true. We suggested, however, that $P'(N|H_S)$ is not low and that the prior probability of N might be quite high. (This is what diminishes N's evidential value.)

The potential problem that we saw for Coltheart and colleagues' explanation of persistence disappears if we read them as adopting the more committed interpretation of the evidence (e.g. taking N to be the proposition *that the woman is the patient's wife* rather than the proposition *that the trusted friend asserted that the woman is the patient's wife*). On that interpretation (as the 'more committed' terminology suggests), the prior probability, $P'(N)$, and the likelihood, $P'(N|H_S)$, are lower – as Coltheart et al.'s account requires. Consequently, it is more plausible that the likelihood ratio outweighs the ratio of

[12] Coltheart and colleagues say (2010, p. 282): 'subjects are impaired at revising preexisting beliefs on the basis of new evidence relevant to any particular belief'. McKay interprets this as a Bayesian proposal that the second factor is an updating bias in which likelihoods are discounted to increase the relative weight given to prior probabilities. Against the proposal, he objects that it is hard to see how the delusional belief would ever have been adopted if a bias in favour of prior probabilities were in place. With such a bias, the patient would have updated his credences much more conservatively on the basis of the unexpected absence of activity in the autonomic nervous system (McKay 2012, section 3.3). This is an interesting objection, but we set it aside here and consider a strand in Coltheart et al.'s (2010) account of the second factor that is rather different from the idea of a general bias in favour of prior probabilities.

[13] This is an instance of the total probability theorem. We follow Coltheart et al. (2010, p. 280) in assuming (for simplicity) that hypotheses inconsistent with both $H_S$ and $H_W$, such as the aunt Agatha hypothesis and the Bob the bank teller hypothesis, have probability zero (equivalently, that $H_W$ is the negation of $H_S$).

prior probabilities so that the ratio of posterior probabilities favours $H_W$ over $H_S$. It is also more plausible that, normatively, the delusional belief should be rejected on the basis of the new evidence, N. Why, then, does the delusional belief persist? Coltheart and colleagues answer this question as follows (2010, p. 279–80):

> What the deluded Capgras subject seems to be doing here is ignoring or disregarding any new evidence that cannot be explained by the stranger hypothesis. It is as though he is so convinced of the truth of the stranger hypothesis by its explanatory power that his conviction makes him either disregard or reject all evidence that is inconsistent with that hypothesis, or at least cannot be explained by the hypothesis. … it seems as if the new information does not even enter the deluded subject's belief system as data that need to be explained.

It is striking that there is nothing evidently Bayesian about this answer to the persistence question.[14]

One important idea behind Coltheart and colleagues' account of how the Capgras patient goes wrong is that, 'it is reasonable in suitable circumstances to accept the evidence of one's senses and the testimony of others' (2010, p. 281; see also p. 275). It is clear that, on their view, it would be reasonable for the Capgras patient to accept as evidence, and update his credences on the basis of, not just the cautious proposition *that the trusted friend asserted that the woman is the patient's wife* but also the committed proposition *that the woman is the patient's wife*. But Coltheart and colleagues also accept that 'it is sometimes reasonable to reject information that cannot be explained by the hypothesis that one is committed to' (p. 280). For example, if a trusted friend and distinguished scientist tells you that, in a tunnel under Switzerland and Italy, particles travel faster than the speed of light, it might be reasonable not to accept the proposition that particles travel faster than the speed of light as evidence – not to update all your credences on that proposition. So when is it reasonable to accept and when is it reasonable to reject, the evidence of perception and testimony?

As we noted earlier, Coltheart and colleagues draw attention to the fact that, in the case of the Capgras patient, the prior probability of the new evidence in favour of the wife hypothesis is low (on the more committed interpretation of which proposition the evidence is). But low prior probability is not supposed to be sufficient to make rejection reasonable (2010, p. 281; emphasis added):

> Capgras subjects are so much *in the grip of the stranger hypothesis* that they refuse to accept the evidence of their senses and the testimony of others. They fail to incorporate the new data into their belief systems *when it is reasonable to do so*.

[14] There are two other notable features of this quotation. First, the answer that it gives to the persistence question seems to be completely different from the answer quoted in note 12 (Coltheart et al. 2010, p. 282; also quoted by McKay 2012, p. 343). Second, it does not seem to describe any impairment of information processing.

No explicit account is offered of what conditions, in addition to low prior probability, need to be met if rejection of evidence is to be reasonable. Furthermore, there seems to be a suggestion – depending on what being in the grip of a hypothesis amounts to – that the real problem is the patient's initial adoption of (assignment of a high credence to) the impostor hypothesis. But, if persistence flows from adoption without additional pathology or departure from normality, then there is no need for a second factor.

Even if we step back from any suggestion that there is no need for a second factor, it remains the case that Coltheart and colleagues have provided no distinctively Bayesian answer to the persistence question.[15]

**Interim summary**

We have considered two recent Bayesian developments of the two-factor framework for explaining delusions. There is a fairly clear disagreement between the two accounts over whether, in addition to the first neuropsychological deficit, a second pathology or departure from normality figures in the explanation of the initial adoption of the delusional belief. Coltheart et al. (2010) say that unbiased Bayesian inference could lead to the onset of Capgras delusion, whereas McKay (2012) says that a bias in favour of likelihoods at the expense of prior probabilities would be required. But we have not been able to identify a role for a further factor (a second factor in Coltheart et al.'s account, a third factor in McKay's) that would explain the persistence of the belief, once it had been adopted. When we adopt a Bayesian approach to the adoption question, persistence of the newly adopted belief appears to be the normal case – and even the normatively correct case.

The lesson that we draw from this is not that there is something wrong with the two-factor framework for explaining delusions, nor that there is something wrong, specifically, with the proposal that the second factor answers the persistence question – option (A). Our view is that the Bayesian approach does not reveal how post-adoption belief evaluation goes wrong in individuals with delusions because it provides no account at all of evaluation of beliefs, once they have been adopted and before new evidence becomes available, even in healthy individuals. In the next section, we argue that the

[15] There is, in fact, a fairly natural Bayesian answer to the specific question about the conditions under which acceptance or rejection of evidence (that is, evidence on the committed interpretation) is reasonable. Acceptance of (assignment of a high credence to) a committed evidence proposition, $E^+$ (e.g. *that the woman is the patient's wife*) is reasonable if the probability of $E^+$ given $E^-$, $P'(E^+|E^-)$, is high, where $E^-$ is the corresponding cautious evidence proposition (e.g. *that a trusted friend asserted that the woman is the patient's wife*). Similarly, rejection of (assignment of a low credence to) $E^+$ is reasonable if $P'(E^+|E^-)$, is low. This will be so if there is a proposition, E*, inconsistent with $E^+$, such that $P'(E^*|E^-) > P'(E^+|E^-)$ – that is, if a proposition that is inconsistent with $E^+$ has a higher probability than $E^+$ given $E^-$.

Evidently, however, that Bayesian proposal would take us back to the earlier potential problem for Coltheart and colleagues' (2010) account. Rejection of the committed evidence proposition, $E^+$ (e.g. *that the woman is the patient's wife*), *would* be reasonable because there is a proposition, $H_S$, inconsistent with $E^+$ (= $H_W$), such that $P'(H_S|E^-) > P'(E^+|E^-)$, where $E^-$ is the corresponding cautious evidence proposition (e.g. *that a trusted friend asserted that the woman is the patient's wife*). The reason for this is that the prior probability of $H_S$ ($P'(H_S)$) is substantially higher than the prior probability of $E^+$ (= $P'(H_W)$), while the likelihood $P'(E^-|H_S)$ is only slightly lower than $P'(E^-|E^+)$ (= $P'(E^-|H_W)$).

problem lies, not with the Bayesian approach as such, but with a standard idealising assumption.

## 4. Belief evaluation and the fragmented mind

If we conceive of believing as a binary (on/off) matter, so that each proposition is either believed or not believed, then the normative ideal for an individual's system of beliefs seems to require a single consistent set of beliefs (propositions for which believing is *on*) that guide action in all contexts. If we adopt a Bayesian approach and conceive of believing as a graded matter, so that each proposition is assigned a credence (subjective probability), then the ideal for a system of beliefs seems to require a single coherent distribution of credences that guide action in all contexts.

Notoriously, we fall short of these supposed ideals. Actual belief systems are fragmented or compartmentalised. Individual fragments are consistent and coherent but fragments are not consistent or coherent with each other and different fragments guide action in different contexts. We hold inconsistent beliefs and act in some contexts on the basis of the belief that P and in other contexts on the basis of the belief that not-P. Frequently we fail to put things together or to 'join up the dots'. It can happen that some actions are guided by a belief that P and other actions are guided by a belief that if P then Q, but no actions are guided by a belief that Q because the belief that P and the belief that if P then Q are in separate fragments. For example, some people are described as having their philosophical beliefs and their religious beliefs in separate compartments. David Lewis describes himself as having once had fragmented beliefs about the geography of Princeton. According to one fragment, Nassau Street ran north-south and was parallel to the railway track; according to another fragment, the railway track ran east-west and was parallel to Nassau Street (Lewis 1982, p. 436).

It is natural to assume that fragmentation is a consequence of our limited information-processing resources and that consistency and coherence across a single web of belief, though practically unattainable, is indeed the ideal. Fragmentation allows inconsistency and it allows failure to believe the consequences of our beliefs. But fragmentation also brings a benefit: it allows us to undertake reflective, critical evaluation of our own beliefs after their initial adoption.

### 4.1 Limits on belief evaluation imposed by consistency and coherence

It is a datum that sometimes we initially adopt a belief (assign a high credence to a proposition), subsequently reflect on the matter and, on the basis of other things that we already know and believe, without any new evidence, reject the belief (assign a lower credence to the proposition). As we shall now explain, this kind of critical reflection involves a departure from the supposed norm of a single consistent and coherent web of belief (Egan 2008).

The limits that are imposed on belief evaluation by the supposed normative ideals are particularly clear in the case of binary (on/off) belief. No consistent system of belief includes both *P* and *not-P*, or all or *P*, *Q*, and *at least one of P and Q is false*. No consistent system includes all of *P*, *my belief that P was formed by method M*, and *every belief formed by method M is false*. And no consistent system includes beliefs, *P*, *Q*, *R*, … formed on the basis of perception, *those are all the beliefs that I have formed on the*

*basis of perception*, and *at least one of the beliefs that I have formed on the basis of perception is false*. Obviously, if you are a consistent binary believer then you will never be in a position to reject an already adopted belief on the grounds that it is inconsistent with other things that you believe. Thus, the ideal of a single consistent system of binary belief excludes rejection of a belief on the basis of evaluation (without new evidence) after the belief has been adopted. The ideal requires that evaluation of a hypothesis should always precede adoption of the hypothesis as a belief.

In the case of graded belief, similar limits are imposed by the supposed ideal of a single coherent assignment of credences. No coherent assignment of credences includes both *P* and *not-P* amongst the propositions assigned a credence greater than 0.5. No coherent assignment includes *P* and *Q* amongst the propositions assigned a credence greater than 0.9 and assigns a credence greater than 0.2 to *at least one of P and Q is false*. And no coherent assignment includes *P*, *Q*, *R*, … amongst the propositions assigned a credence greater than 0.9 and assigns a credence greater than 0.2 to *at least half of the those propositions, P, Q, R, ... , are false*. If you are a coherent graded believer and you assign a credence greater than 0.9 to some propositions then, provided that you are well informed about your own credences, you cannot assign a credence greater than 0.2 to *half of the propositions to which I assign a credence greater than 0.9 are false*.[16]

The general situation is that, if you are a coherent graded believer then you will never be in a position to reject an already adopted belief (reduce the high credence that you have assigned to a proposition) on the grounds that it is not coherent with other things that you believe (the assignment of high credences to other propositions). If you have assigned a high credence to *P*, and *Q* is evidence against *P* (the conditional probability of *P* given *Q* is low) then, because your assignment of credences is coherent, you will not have assigned a high credence to Q. On a Bayesian approach, there is no place for post-adoption (that is, post-updating) belief evaluation, without new evidence.

### 4.2 Compartmentalisation and Spinozan belief formation

A fragmented belief system and, specifically, a belief system in which newly adopted beliefs were compartmentalised, would escape the limits imposed by the ideal of a single consistent and coherent system. Assignments of credences that were not coherent with each other could be separated and, specifically, a new assignment of a high credence to a proposition could be separated from a prior assignment of credences with which it was not coherent. By retaining the prior credences, such compartmentalisation would allow the possibility of post-adoption belief evaluation, even without new evidence. It might be suggested that, although our belief systems are in fact fragmented in various ways, earlier evaluation of candidates for belief is always preferable to the post-adoption evaluation that compartmentalisation allows. But, as we explain in the next few paragraphs, default or prepotent doxastic responses to incoming information, particularly from perception, may leave little, if any, opportunity for early evaluation of candidates for belief.

[16] For the general theorem, see Egan and Elga (2005).

In a series of papers, Gilbert and colleagues (Gilbert et al. 1990, 1993; Gilbert 1991) have contrasted Cartesian and Spinozan views of belief and have presented experimental results in support of the Spinozan view.[17] Each view of belief can be summarised in terms of two stages, a representation stage and an assessment stage. On the Cartesian view, the representation stage involves *comprehension*. A hypothesis is first grasped and then, in a separate assessment stage, the hypothesis is either *accepted* as true and adopted as a belief, or else *rejected* as false. On the Spinozan view, in contrast, the representation stage involves both comprehension and *acceptance* (Gilbert 1991, p. 107): 'People believe in the ideas they comprehend, as quickly and automatically as they believe in the objects they see.' Then, in the separate assessment stage, the already adopted belief is either *certified* or else *unaccepted*.

Consider, for example, a case of belief based on perception, such as my belief that there is a pencil on the desk in front of me. The Cartesian view is that the representational content of my perceptual experience *as of* there being a pencil on the desk is first grasped as a hypothesis, which is then assessed. Such an assessment might consider the explanatory potential of the hypothesis – whether the hypothesis that there is a pencil on the desk explains my having an experience as of there being a pencil on the desk – and how probable the hypothesis is in the light of other available evidence. Whatever the exact nature of the assessment, its outcome is that the hypothesis is either accepted – I adopt the belief that there is a pencil on the desk – or else rejected – I do not believe my eyes. The Spinozan view, in contrast, is that comprehension and acceptance are inseparable. There is no gap between grasping how the experience represents the world to be and believing that the world is that way. On the basis of the perceptual experience, I automatically adopt the belief that there is a pencil on the desk. Assessment is still separate from comprehension and so it follows this initial acceptance. The outcome of the assessment is that the already adopted belief is either certified and retained or else unaccepted in 'a deliberate revision of belief' (1991, p. 108).

Gilbert assumes that assessment, whether Cartesian or Spinozan, is demanding of cognitive resources and that if resources are depleted or in use elsewhere then the pre-assessment state of the system will be revealed. He finds the Spinozan view to be supported by a range of empirical findings (see Gilbert 1991, for a review) including results from his own experiments (Gilbert et al. 1990, 1993). As Gilbert explains the logic of these experiments, 'resource-depleted Cartesians should be uncertain, uncommitted, but not persuaded', whereas the results indicate that 'resource depletion facilitate[s] believing' (1991, p. 111). Belief based on perception seems to be the most promising case for the Spinozan view but Gilbert's results also suggest that the view may be correct for belief based on heard testimony and read text, even when individuals are forewarned that some of what they hear or read is not true.

Spinozan belief formation, by definition, excludes assessment or evaluation of hypotheses before they are adopted as beliefs. If belief systems were required to be

[17] We shall not discuss the relationship between the two views that Gilbert describes and the historical philosophers for whom they are named.

consistent and coherent, then Spinozan belief formation would also exclude post-adoption belief evaluation without new evidence – acceptance would preclude assessment. The reason is that automatic adoption of an antecedently implausible belief would, given consistency and coherence, eliminate the very considerations in the light of which the belief was implausible. So there would be no basis for reflective, critical evaluation of the newly adopted belief. Spinozan belief formation would be consistent with post-adoption belief evaluation, however, if newly adopted beliefs were to be compartmentalised, at least until evaluation (Gilbert's assessment) had been undertaken. We assume that these compartmentalised beliefs would guide action in some, but not all, contexts.

### 4.3 Prepotent doxastic responses, compartmentalisation, and belief evaluation

We have said that it is a datum that beliefs are sometimes evaluated after they have been initially adopted and we have shown that a fragmented belief system in which newly adopted beliefs are compartmentalised allows post-adoption belief evaluation, even without new evidence. As we have just seen, this is important for the Spinozan view of belief formation, which places evaluation (assessment) after adoption (acceptance). The claim that belief is the default or prepotent doxastic response to perceptual experience (Davies et al. 2001, p. 153) allows that the prepotent response might sometimes be inhibited; so it is not quite as strong as the claim that belief formation based on perception is Spinozan.[18] Nevertheless, on this near-Spinozan view the default case will be that there is no assessment before acceptance; reflective evaluation will have to follow initial adoption of the belief. Evidently, if beliefs that are newly adopted on the basis of perceptual experience are compartmentalised then the near-Spinozan view is consistent with post-adoption belief evaluation. To the extent that belief is the prepotent doxastic response to heard testimony or read text, a similar argument will apply.

On a Bayesian approach, the near-Spinozan view is that, in the default case, a high credence is assigned to a proposition that specifies how a perceptual experience presents or represents the world as being (e.g. the proposition *that there is a pencil on the desk in front of me*). The ideal of a single coherent distribution of credences would require that all other credences should be automatically updated (presumably by simple conditionalisation or Jeffrey conditionalisation). Compartmentalisation separates the assignment of a high credence to the perception-based proposition from the prior assignment of credences and the two assignments coexist *pro tem*. Thus the prior credences remain available to figure in subsequent evaluation of the high credence that was initially assigned to the perception-based proposition.

With some understanding of how – thanks to compartmentalisation – there could be post-adoption belief evaluation in healthy individuals, we can return to the explanation of delusions. The two-factor framework for explaining delusions is also a three-stage framework (Aimola Davies et al. 2009; Aimola Davies and Davies 2009). At the first

[18] Hasson, Simmons and Todorov (2005) argue that the relationship between comprehension and acceptance is more complex than the Spinozan view allows.

stage there is a neuropsychological deficit, the first factor, which is supposed to be related to the content of the delusional belief in some plausible way. There is then a second stage that leads from the neuropsychological deficit to the initial adoption of the delusional belief. Finally, there is the third, post-adoption, stage of persistence rather than rejection of the delusional belief.

## 5. How did the patient come to adopt the belief?

Coltheart's first question is, 'Where did the delusion come from?' (2007, p. 1044). One kind of answer to this question simply cites the first factor; the answer specifies the neuropsychological deficit from which the delusion's content is supposed to have arisen in some plausible way. In the case of Capgras delusion, the deficit is disconnection of the primary face processing system from the autonomic nervous system. But a neuropsychological deficit is not a belief and an answer to the adoption question must specify, not only the neuropsychological deficit, but also the route from the deficit to initial adoption of the delusional belief – the second stage in the two-factor / three-stage framework.

### 5.1 Abnormal data and anomalous experience

The framework allows for considerable variation in the second stage. For example, this stage might be hidden from consciousness to a greater or a lesser extent. As we have already seen, Coltheart and colleagues commend the view that the processing that leads up to onset of the delusion is wholly unconscious, so that the delusional belief itself is 'the first delusion-relevant event of which the patient is aware' (2010, p. 264).

On a Bayesian account of the kind that we have been considering, there must be, near the beginning of the second stage, evidence or data from which Bayesian inference leads to initial adoption of the delusional belief. On Coltheart and colleagues' account, the starting point for Bayesian inference leading to the Capgras delusion is abnormal data; namely, the absence of activity in the patient's autonomic nervous system. Importantly, this absence of autonomic activity is assumed *not* to be available to consciousness (2010, p. 264): 'people are not conscious of the activities of their autonomic nervous systems, and so a man would not be conscious of a failure of his autonomic nervous system to respond when he encountered his wife'. There is a good theoretical reason behind this assumption (Coltheart 2005). If the autonomic activity that is normally produced in response to a familiar face were available to consciousness then patients suffering from prosopagnosia, but with covert recognition, should be able to use their experience of autonomic activity to discriminate familiar from unfamiliar faces. But such patients are unable to do this (Tranel and Damasio 1985, 1988).[19]

[19] Garry Young has suggested that the results of a study by Greve and Bauer (1990) lend support to the view that autonomic activity does 'pervade consciousness' in the form of a 'sense of preference' (Young 2008, p. 868; see also 2007). Greve and Bauer presented a prosopagnosic patient with pairs of unfamiliar faces. One face in each pair had previously been exposed five times (for 500 msec each time), following trials of a task in which verbal personality descriptors were presented. When asked to select the face that he liked best, the patient chose the previously exposed face from 70% of the pairs. This is certainly an

We assume, with Coltheart and colleagues (2010), that neither autonomic activity produced by familiar faces, nor the absence of autonomic activity, is available as such to consciousness. But this does not rule out the possibility that the abnormal absence of autonomic activity – itself unavailable to consciousness – should lead to an anomalous experience that would be causally antecedent to the adoption of the delusional belief. Indeed, Coltheart suggests just such a possibility, beginning from an important idea about information processing (2005, p. 155): 'It is a general principle of cognitive life that we are continually making predictions, on the basis of what we currently know about the world, concerning what will happen to us next.' Sometimes we engage in the conscious mental activity of making predictions and then learning whether our predictions are fulfilled. But prediction and the comparison of predictions with actual outcomes are also ubiquitous features of unconscious information processing. Coltheart says (ibid.): 'Only when a prediction fails does consciousness get involved; the unconscious system makes some kind of report to consciousness to instigate some intelligent conscious problem-solving behavior that will discover what's wrong.' Thus, in the case of the Capgras patient, what happens is (ibid.): 'the unconscious system predicting that when the wife is next seen a high autonomic response will occur, detecting that this does not occur, and reporting to consciousness, "There's something odd about this woman"'.

Maher makes a very similar suggestion about the consequences for conscious experience of the unconscious operation of prediction and comparison systems (1999, p. 558):

> Survival requires the existence of a detector of changes in the normally regular patterns of environmental stimuli, namely those that are typically dealt with automatically. The detector functions as a general non-specific alarm, a 'significance generator', which then alerts the individual to scan the environment to find out what has changed.

Maher's (1999) and Coltheart's (2005) suggestions converge on the idea that, as the result of a neuropsychological deficit and the subsequent operation of a comparator

---

interesting finding and Young's proposal that 'there was something-it-was-like … for [the patient] to prefer one face over the other' (2008, p. 868) is plausible. But the finding does not support the proposal that this 'sense of preference' was underpinned by activity in the patient's autonomic nervous system. Greve and Bauer (1990) report no data on skin conductance responses and they explicitly attribute the patient's preference responses, not to autonomic activity, but to perceptual fluency (Mandler 1980; Whittlesea 1993). They also tentatively suggest that the patient may have a deficit in forming new FRUs and that the more fluent processing of the briefly exposed face stimuli may have been independent of their status as faces.

Although Greve and Bauer's (1990) results do not support Young's claim that autonomic activity is available to consciousness as a sense of preference, we speculate that the results may indirectly raise a question about Coltheart's (2005) argument that autonomic nervous system responses to faces are *not* available to consciousness. It is plausible that, when a previously exposed face was presented to the patient in Greve and Bauer's study, information about perceptual fluency was, in some way, available to consciousness (e.g. as a sense of preference) although the patient was not able to make use of information about fluency to identify the face as previously exposed. By analogy, it might be suggested, the fact that prosopagnosic patients are not able to use information about autonomic activity to discriminate familiar from unfamiliar faces does not absolutely exclude the possibility that information about autonomic activity is, in some way, available to consciousness.

system, the Capgras patient's experience of his wife is anomalous, characterised by a sense of significance and change.

In summary: From a starting point of abnormal data, the processing that leads to the initial adoption of the delusional belief might be wholly hidden from consciousness. Alternatively, unconscious processes might lead to an anomalous experience that is causally antecedent to the delusional belief.

### 5.2 Explanation or endorsement

There are two importantly different ways in which an anomalous experience might figure in the aetiology of a delusion (Davies et al. 2001). On the first (explanation) option, the content of the delusion is not encoded in the anomalous experience. For example, the content of the experience might be: 'This woman looks just like my wife but there is something different or odd about her'. The content of the delusion itself arises first as the content of an explanatory hypothesis that is then adopted as a belief. This is the account given by Maher (1974, p. 103):

> [T]he explanations (i.e. the delusions) of the patient are derived by cognitive activity that is essentially indistinguishable from that employed by non-patients, by scientists, and by people generally. … [A] delusion is a hypothesis designed to explain unusual perceptual phenomena.

It is clear that, on the explanation option as Maher conceives it, the project of explaining the anomalous experience is a project of the person. If explanation proceeds by Bayesian inference, then this is personal-level inference. We shall understand the explanation option in Maher's way.

On the second (endorsement)[20] option, the content of the delusion is encoded in the anomalous experience. Thus, a state in which the content of the delusion is present is 'the first delusion-relevant event of which the patient is aware'. The anomalous experience is not the delusional belief itself, but the belief would arise as the prepotent doxastic response to the experience (or, on the Spinozan view, would be adopted automatically on the basis of the experience). Although the endorsement option is distinct from Coltheart and colleagues' (2010) view that the delusional belief *itself* is the first conscious delusion-relevant event, both agree that no conscious personal-level reasoning is involved in the stage leading to onset of the delusion (see Turner and Coltheart 2010, p. 368).

The explanation and endorsement options are distinct. But it is important to observe that the distinction between them is *not* that one allows, while the other excludes, the possibility that Bayesian inference figures in the aetiology of the delusion. It is entirely consistent with the endorsement option that the anomalous experience encoding the content of the delusion is the product of perceptual processes that are well described as unconscious subpersonal-level Bayesian inference.

---

[20] The terminology of 'endorsement' comes from Bayne and Pacherie (2004). Descriptions of the explanation *versus* endorsement options, with varying terminology, are also given by Davies and Coltheart (2000), Fine, Craigie and Gold (2005), Aimola Davies and Davies (2009), Langdon and Bayne (2010), and Turner and Coltheart (2010).

5.3 Bayesian inference in a perceptual module

In our earlier discussion of the processes leading up to initial adoption of the delusional belief (section 2), we noted that it is somewhat unclear which probabilities are supposed to figure in the worked examples of Coltheart and colleagues (2010) and McKay (2012). Are they supposed to be realistic probabilities or the patient's credences, for example? McKay challenges Coltheart et al.'s account on the grounds that their proposed prior probability for the stranger hypothesis, $P(H_S) = 0.01$, is 'unrealistically high' (2012, p. 339) and offers a much lower estimate. But he accepts their estimate of the likelihood ratio as favouring $H_S$ in the ratio 999:1. This acceptance is puzzling because it seems clear that realistic values for the likelihoods, $P(D|H_S)$ and $P(D|H_W)$, would be identical. The seen face of a woman who looked just like the patient's wife would activate $FRU_W$ in the same way as the seen face of the patient's wife would; and it would cause the same activity in the autonomic nervous system. So a realistic likelihood ratio would be 1:1 and no amount of discounting prior probabilities (which favour $H_W$) would produce posterior probabilities favouring $H_S$.

It may seem more natural to interpret the probabilities in the worked examples as the patient's own credences but this, too, is problematic. First, the worked examples (so interpreted) are far from intuitive if, as Coltheart and colleagues explicitly maintain, activity in the patient's autonomic nervous system is not available to the patient's consciousness. How are we to understand a credence such as $P(D|H_W)$ if the patient knows nothing of autonomic activity as such and does not even experience autonomic activity? Second, waiving this first concern, suppose that the patient assigns a low conditional credence to there being no activity in his autonomic nervous system given that the woman he sees in front of him is his wife and a high conditional credence to there being no activity in his autonomic nervous system given that the woman is not his wife but looks just like his wife. The only way for the patient's credences $P(D|H_W)$ and $P(D|H_S)$ to come apart in the way described is for the patient to assign high credences to hypotheses according to which the autonomic nervous system is not just responsive to qualitative inputs from visual perception via the FRUs, but also directly sensitive to facts about the identity of the person perceived.

Our suggestion is that the probabilities in Bayesian inference that starts from the abnormal data, D (the absence of activity in the patient's autonomic nervous system) should be conceived as probabilities assigned by a subpersonal-level unconscious information-processing system or perceptual module.[21] The module's probabilities may not be realistic, because a module has only limited information about how the world (including the module itself) really works. And the module's probabilities may not be

[21] Attributions of credences to persons are standardly explained in terms of dispositions to accept (or regard as fair) bets at specified odds. Information-processing systems or modules do not engage in betting behaviour but simply produce outputs that are available to other systems. A description of a system as engaging in Bayesian inference or updating is to be understood as a putative account of the computations being carried out by the system. As such it is answerable to the system's input-output dispositions and to such evidence as may be available concerning intermediate stages of the computation (Shea, this volume, 2012).

coherent with personal-level credences because modules are, at least to some extent, informationally encapsulated (Fodor 1983). Processes in a module do not draw on all that the person knows or believes and so we sometimes experience what we, at the personal level, know to be perceptual illusions. It is also because of this relative independence of the probability distributions of perceptual systems from a person's credences that we can learn from experience (Fodor 1989).

Let us return to Coltheart and colleagues' (2010) and McKay's (2012) worked examples, now conceived as examples of Bayesian inference within a perceptual module. Fodor says that the function of modules is 'to so represent the world as to make it accessible to thought' (1983, p. 40) and that the outputs of modules 'are typically phenomenologically salient' (ibid., p. 87). So we assume that, although the Bayesian inference begins from data that are not available to personal-level consciousness, the results of the inference will determine the content of an experience that presents or represents the world as being a certain way.

A module that is dedicated to processing information about faces will have a limited representational 'vocabulary' and the hypotheses that it can 'consider' will be correspondingly restricted. We assume, for the purposes of the example, that these include, not only hypotheses about whose face a presented face qualitatively looks like, but also hypotheses about who the presented person *is*; for example, that the person is the patient's wife ($H_W$), or aunt Agatha ($H_A$), or Bob the bank teller ($H_B$), or that the person is somebody unfamiliar ($H_S$). From some initial distribution, the probabilities of these hypotheses are updated on the basis of input information about the qualitative appearance of the presented face; specifically, the face looks just like the face of the patient's wife. After this updating, $P(H_W)$ is much greater than $P(H_S)$. Since these are probabilities assigned by a module performing a specific computational task on the basis of limited information, we do not assume that they are wholly realistic.

It is important that, at the processing stage we are considering, the module generates predictions about levels of activity in the autonomic nervous system from hypotheses about the identity of the person whose face is presented. From the hypothesis $H_W$, the module generates the prediction that there will be a high level of autonomic activity. From $H_A$ or from $H_B$, the prediction is that there will be a low level of autonomic activity, somewhat above baseline. And from $H_S$, the prediction is that there will be no autonomic activity above baseline. We assume that the generation of these predictions is based on past associations – perhaps between activation of one or another PIN, or none, and levels of autonomic activity.[22] So we do not expect that the predictions will be based on information about how the module itself actually works. Specifically, we do not expect that the predictions will take into account that autonomic activity is caused by activation

[22] In the model of face processing proposed by Ellis and Lewis (2001), an integrative device has access to information from PINs and from affective responses to familiar stimuli in the autonomic nervous system (see also Breen, Caine and Coltheart 2000; Breen, Coltheart and Caine 2001; Lewis and Ellis 2001). Such a device could, in principle, learn associations, generate predictions, and compare predicted and actual autonomic activity. In a more comprehensive account, we should also consider predictions about the voice and the gait, for example, of the person putatively identified.

of a FRU and that if a stranger's face looks just like a known face then it will activate the same FRU, and cause the same autonomic activity, as the known face would.

We can regard the module's predictions as manifesting its estimates of conditional probabilities. If D is the absence of autonomic activity then the module's likelihoods, $P(D|H_W)$ and $P(D|H_S)$, will favour ($H_S$) over ($H_W$) – just as in Coltheart et al.'s (2010) and McKay's (2012) examples. These likelihoods are unrealistic but they are not the likelihoods of the person within whom the module resides (the patient).

The main point of disagreement between Coltheart and colleagues and McKay was over the question whether the disconnection of the patient's face processing system from the autonomic nervous system is the only pathology or departure from normality that need be implicated in the processes leading up to the onset of the Capgras delusion (section 2.2). McKay argued that, given realistic estimates of prior probabilities, an updating bias in favour of likelihoods at the expense of prior probabilities was needed in order to answer the adoption question. He proposed that this bias is the second factor in the two-factor framework. When we consider Bayesian inference in a perceptual module, the question whether prior probabilities – or, indeed, likelihoods – are realistic is less pressing. Also, because modules are, to some extent, informationally encapsulated, it is of their nature to discount personal-level prior probabilities and to be biased in favour of likelihoods. Earlier, we recalled Hemsley and Garety's (1986) remark about normal deviation from the normative model. Now, as we consider information processing in a perceptual module, this remark seems even more to the point. In perceptual processing, a bias in favour of likelihoods may well be entirely normal.

### 5.4 From anomalous experience to delusional belief

We have described, in a speculative spirit, how information processing in a perceptual module could lead from abnormal data (absence of autonomic activity as the consequence of a neuropsychological deficit) to the stranger hypothesis, $H_S$, being favoured over the wife hypothesis, $H_W$. We assumed that the favoured hypothesis would determine the content of an experience that presents or represents the world as being a certain way. This would be an experience as of a woman who qualitatively looked just like the patient's wife but was not really her (was really a stranger). An anomalous experience is not yet a belief, but the delusional belief, 'This woman looks just like my wife (and says that she is my wife) but is not my wife', would arise as the prepotent doxastic response to this experience. There are two points to notice about this answer to the adoption question. First, no conscious personal-level reasoning is involved in the stage leading to adoption of the delusional belief. Second, the original neuropsychological deficit seems to be the only pathology or departure from normality that is essential to the answer – in line with option (A) (see section 1.1).

It might be objected against this answer that it assumes a controversial position in the philosophy of perception; namely, that properties of being numerically identical to, or distinct from, a particular individual (e.g. the property of being numerically distinct from the patient's wife) are represented in perceptual experience. It is true that the question of which properties of objects are represented in perceptual experience is important and much debated (e.g. Bayne 2009; Hawley and Macpherson 2010; Siegel 2010) and that this has sometimes been offered as an objection to an endorsement account of the

Capgras delusion (Coltheart 2005). We find the endorsement option attractive but, to conclude this section on the adoption question, we briefly consider the possibility that hypotheses like $H_W$ and $H_S$ do not figure in perceptual processing and that the content of the delusion is not encoded in an anomalous experience.

This returns us to the explanation option described by Maher (1974, 1999) and Coltheart (2005). The Capgras patient's experience presents or represents a woman who qualitatively looks just like his wife. There is nothing anomalous about that. But, on the basis of past associations between the qualitative appearance of a presented face and the level of activity in the autonomic nervous system, an unconscious subpersonal-level mechanism predicts a high level of autonomic activity. Because the patient's primary face processing system has been disconnected from the autonomic nervous system, this prediction is not fulfilled and a prediction error signal is generated. Consequently, the patient's experience of the woman who is, in fact, his wife is suffused with a feeling of heightened significance, it cries out for explanation in terms of something different or odd in the immediate environment and particularly in the woman perceived.

The prepotent doxastic response to this experience is to believe something like, 'This woman looks just like my wife but there is something different or odd about her'. But this underdescribes the way in which the belief is permeated by the sense of significance and the urgent demand for explanation and interpretation. We assume that the cognitive processes that are engaged by a situation like that of the Capgras patient are not reflective and unbiased processes of Bayesian inference to the most probable hypothesis all things considered, but more encapsulated and biased processes of inference to the first (most accessible) hypothesis to predict the anomalous experience well enough (that is, with a high enough likelihood on the evidence of the anomalous experience).[23] The accessible hypotheses are not restricted by the limited representational vocabulary of a perceptual module. But the sense that there is something different or odd about the woman that the patient sees in front of him demands attention and explanation. What is it that has changed? In this situation, it may be that the most accessible hypothesis that answers this question and predicts the sense that there is something different or odd about the woman is that, although she looks just like the patient's wife and says that she is the patient's wife, she is not really his wife but a stranger, and so an impostor.[24]

[23] Coltheart, Langdon and McKay (2011, p. 283–4) observe, in effect, that C.S. Peirce's notion of abduction was close to this idea of inference to the first hypothesis with a high enough likelihood. Peirce's characterisation of abduction considered as an inference was: 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. (See Psillos 2009, for a recent review.)

[24] This somewhat encapsulated inference to the first hypothesis to predict or explain a salient piece of evidence is analogous to somewhat encapsulated inference to the first relevant enough interpretation of an utterance, as proposed by Sperber and Wilson's relevance theory of pragmatics (Sperber and Wilson 1986/1995, 2002; Wilson and Sperber 2012). We do not commit ourselves to the thesis that the mind is modular through and through (Sperber 2001; Carruthers 2006) but we acknowledge that informational encapsulation does not provide a straightforward criterion for separating perception from cognition (see Shea, forthcoming).

In a Cartesian – rather than Spinozan or near-Spinozan – spirit, Coltheart, Langdon and McKay (2011, p. 284) say:

> The propositions yielded by abductive inference are not beliefs, but rather are hypotheses or candidates for belief. For any such proposition to be adopted as a belief, it must be submitted to, and survive, a belief-evaluation process, and it is here that plausibility has a critical role.

In contrast, we conjecture that there is a prepotent doxastic tendency towards acceptance of a hypothesis that arises in the way that we have just described; that is, as the first hypothesis to predict (and perhaps also explain) a highly salient piece of evidence.[25] This prepotent tendency would correspond to a cognitive imperative of explanatory adequacy, as the prepotent response to perceptual experience corresponds to the cognitive imperative of observational adequacy. On the explanation option, as on the endorsement option, the delusional belief arises as a prepotent doxastic response and, in both cases, the processes leading to initial adoption of the belief discount personal-level prior probabilities and are biased in favour of likelihoods. The difference between the two cases is just that, while the endorsement option involves encapsulated and biased subpersonal-level perceptual processes of Bayesian inference, the explanation option involves encapsulated and biased post-perceptual processes of Bayesian inference. Compartmentalisation of the newly adopted belief allows prior probabilities to be taken into account in subsequent belief evaluation, in accordance with the third cognitive imperative of conservatism (Stone and Young 1997; Aimola Davies and Davies 2009).

## 6. Why does the belief, once adopted, persist rather than being rejected?

Coltheart's second question is, 'Why does the patient not reject the belief?' (2007, p. 1044). As we have noted (section 1.1), this question can be interpreted in two ways. On one interpretation, the question asks why the patient adopts the hypothesis as a belief, rather than rejecting it (Coltheart et al. 2011). On a second interpretation, it asks why the patient, having initially adopted the belief, does not subsequently reject it (Coltheart 2007, 2010). On the first interpretation, Coltheart's second question is the adoption question; on the second interpretation, it is the persistence question.

The distinction between these two interpretations of Coltheart's second question led us to two options for the role of the second factor in the two-factor framework. According to option (A), no pathology or departure from normality beyond the first factor (neuropsychological deficit) is needed to answer the adoption question and the second factor answers the persistence question. Thus, the role of the second factor is to explain a failure of post-adoption belief evaluation. According to option (B), two factors are needed to answer the adoption question and no additional pathology or departure from normality is needed to answer the persistence question. Thus, the role of the second factor is to explain the transition from the first neuropsychological deficit to the initial adoption

[25] In section 1.3, we allowed that explanatory virtue might influence the psychological accessibility of hypotheses.

of the delusional belief. On our reading, Coltheart et al. (2010) adopt option (A) while McKay (2012) adopts option (B) – as, it seems, do Coltheart et al. (2011).

On option (B), it is plausible to propose that the cognitive nature of the second factor is that of a bias; specifically, a bias in favour of likelihoods at the expense of prior probabilities. (What is not so clear is whether the bias amounts to a pathology or departure from normality. Might it just be a bias within the normal range, so that option (B) would appeal to only one pathological factor?) We were not able, however, to find a role for a second factor of the kind specified by option (A). This would be an impairment that would explain the persistence of the delusional belief once it had been adopted. Given the implausibility of the belief, why is it not rejected, even before new evidence becomes available? On a Bayesian approach, there seems to be no role for such an impairment because, once a belief has been adopted, persistence is the normal, and even normatively correct, case.

We do not regard that difficulty as revealing a problem for the two-factor framework, for a Bayesian approach, or for option (A). Rather, it reflects the fact that, given the standard idealising assumption of a single coherent assignment of credences, Bayesian belief evaluation can only be updating of credences by conditionalising on new (that is, post-adoption) evidence. Our view (see section 4) is that we should relax the standard idealising assumption and allow that the assignment of a high credence to a proposition in accordance with the cognitive imperative of observational or explanatory adequacy can be compartmentalised, so that prior credences remain available to figure in subsequent belief evaluation. We suggest that compartmentalisation should normally be triggered whenever belief formation is near-Spinozan. The reason for this suggestion is that near-Spinozan belief formation ensures that belief adoption precedes evaluation. In such cases, compartmentalisation allows the prior assignment of credences to be protected, at least until belief evaluation has been undertaken. Without compartmentalisation, the prior credences would be automatically updated by conditionalisation on the proposition newly assigned a high credence, with the result that critical belief evaluation would be impossible.[26]

Now that we have seen how post-adoption belief evaluation could be possible in principle, we can ask how it could be impaired.

### 6.1 Bayesian belief evaluation and cognitive resources

The Capgras patient has initially adopted the belief that the woman who looks just like his wife and claims to be his wife is not really his wife but a stranger. The content of the delusional belief has arisen in one or other of two plausible ways (endorsement or explanation) from an anomalous experience that arose, in turn, from a neuropsychological deficit by processes of Bayesian inference in a perceptual module. On either option, the processes that led up to initial adoption of the delusional belief were

[26] It might be that compartmentalisation should not be triggered when it is not required; for example, when the newly assigned credence coincides with the prior credence assigned to that proposition. But equally, if compartmentalisation were triggered in such a case, it would be short-lived as belief evaluation and subsequent integration would be trivial.

(we assume) normal, but were at least somewhat encapsulated and correspondingly biased in favour of likelihoods at the expense of prior probabilities. Thanks to compartmentalisation of the delusional belief, the patient's prior credences have been preserved and are, in principle, available for reflective, critical evaluation of the newly adopted belief. But belief evaluation is demanding of cognitive resources (Gilbert 1991).

In their seminal contribution to the philosophy and psychology of delusion, Stone and Young describe (1997, p. 349): 'a tension between forming beliefs that require little readjustment to the web of belief (conservatism) and forming beliefs that do justice to the deliverances of one's perceptual systems' and they propose that, in delusion, the balance between these cognitive imperatives 'goes too far towards observational [or explanatory] adequacy as against conservatism'. Our account is broadly consistent with that proposal and we agree that belief evaluation will involve control or management of these potentially competing imperatives. More specifically, critical evaluation of the delusional belief in the light of prior credences will require some inhibition of the prepotent tendencies corresponding to the imperatives of observational and explanatory adequacy. The patient will need to step back from his initial adoption of the delusional belief and say (paraphrasing patient HS, described by Chatterjee and Mennemeier 1996, p. 227), 'Sometimes I feel like my wife has been replaced by an impostor and I have to tell myself "no she hasn't"'.

Critical evaluation of the delusional belief will also require that the patient achieve a better understanding of his real situation. As the patient described by Turner and Coltheart said (2010, p. 371):

> I've started going through it, and seeing what could possibly happen and what couldn't happen. … there must be an explanation for it. … The lady knows me way back. She could say things that happened 40 years go, and I wonder where she gets them from. And then I worked it out and I've wondered if it's Mary all the time.'

This 'going through it' and 'working it out' requires, not only executive processes of control and inhibition, but also working memory resources for the maintenance and manipulation of information. Impaired executive function or working memory could figure in an answer to the persistence question (Aimola Davies and Davies 2009).

These suggestions about the cognitive resources that are demanded by the task of belief evaluation do not yet amount to a proposal about the nature of the evaluation. The obvious proposal on a Bayesian approach would be that it is an assessment based on prior probability and likelihoods.[27] The question for such an evaluation is whether there is a

[27] This evaluation would assess whether the probability of the committed Capgras delusion proposition, $C^+$, given the evidence on which its adoption as a belief was based, is high or low (see again footnote 15).

> $C^+$ *that the woman who looks just like the patient's wife and claims to be his wife is not really his wife but a stranger*.

The evidence proposition would be a more cautious proposition describing the patient's experience. On the endorsement option, this would be the proposition:

> $C^-_{end}$ *that the patient has a perceptual experience as of the woman who looks just like his wife and claims to be his wife not really being his wife but being a stranger*.

proposition, C*, that is inconsistent with the Capgras delusion proposition and is more probable than the delusion proposition given the anomalous experience on the basis of which the delusional belief was initially adopted (by endorsement or explanation). Informally, and departing somewhat from the Bayesian approach, one might ask whether there is an alternative to the impostor hypothesis that provides a better explanation of the patient's anomalous experience.

There is, of course, an obvious candidate for such a proposition, C*. The patient has suffered a stroke, resulting in disconnection of his primary face processing system from his autonomic nervous system; and this neuropsychological deficit has led to the anomalous experience of his wife. This proposition can, in principle, be recognised as having a higher posterior probability than the Capgras delusion proposition because the patient's prior credences remain available. They have not been updated by conditionalisation on the compartmentalised delusion proposition. Nevertheless, for several reasons, the patient's path to a correct understanding of his situation may be difficult.

First, even if the patient accepts that he has had a stroke and appreciates that a stroke can have many debilitating consequences, the probability of an anomalous experience of the relevant kind given only that the patient has had a stroke is not especially high. The probability of such an experience given that the patient has had a stroke resulting in disconnection of his primary face processing system from his autonomic nervous system is much higher. But the patient may not realise that a stroke can have this consequence and, even if the patient knows about the possibility of disconnection, it may be far from self-evident to him how such disconnection would be manifested in experience.[28] Second, from the point of view of the Capgras patient, once the delusional belief has been initially adopted, the question of interest is not why he is having this anomalous experience (the question answered by C*), but why his wife has been replaced by a stranger (Hohwy and Rosenberg 2005, p. 155):

> The brain pathology hypothesis would only be relevant if the patient could accept that what needs explaining is the mere experience that it is as if the spouse looks like a stranger; it is not relevant if what needs explaining is the real occurrence of a stranger looking like the spouse.

Third, on the explanation option, the feeling of heightened significance that suffuses the experience of the patient's wife cries out for explanation in terms of change in the environment, not change in the patient's brain (such as a stroke).

---

On the explanation option, it would be the proposition:

> $C^-_{exp}$ *that the patient has a perceptual experience (an experience suffused with a feeling of significance and change) as of the woman who looks just like his wife and claims to be his wife having something different or odd about her*.

The probability of $C^+$, given the evidence on which its adoption as a belief was based, will be low if there is a proposition, C*, inconsistent with $C^+$, such that $P'(C^*|C^-_{end}) > P'(C^+|C^-_{end})$ in a case of endorsement, or such that $P'(C^*|C^-_{exp}) > P'(C^+|C^-_{exp})$ in a case of explanation.

[28] Here we are indebted to Ryan McKay.

It is plausible that, if the patient suffered from impaired executive function or working memory, these difficulties in reaching a correct understanding of his situation would be exacerbated. Even without such impairments, it seems plausible that the patient might need some assistance in understanding his situation.

### 6.2 Failure of compartmentalisation

Compartmentalisation of a newly adopted belief allows post-adoption belief evaluation and we have suggested that it should normally be triggered whenever belief formation is near-Spinozan. It is of some interest to compare our suggestion about compartmentalisation with a proposal by Turner and Coltheart that an unconscious checking system 'tags' thoughts 'that require extra conscious checking' (2010, p. 357). The presence of a tag has phenomenal and functional consequences. It 'gives rise to the experience of doubt' and the tagged thought is 'referred to the conscious evaluation system for further work' (ibid.). Failure of the unconscious checking system results in 'absence of doubt' and 'a subjective feeling of conviction' (pp. 358, 360). The untagged thought is able to 'bypass the conscious evaluation system, and … directly affect speech and other behaviour' (p. 358).

There are certainly points of similarity between the compartmentalisation and tagging ideas. Compartmentalisation is primarily functional and we have not suggested any phenomenology associated with it. But we do not rule out the possibility that it might be accompanied by a sense that evaluation is called for. Also, although belief evaluation is demanding of cognitive resources and not automatic, it is plausible that, in healthy individuals whose cognitive resources are not engaged elsewhere, newly adopted and compartmentalised beliefs are normally subject to evaluation – as on Gilbert's (1991) Spinozan view. However, unlike Turner and Coltheart's tagging, compartmentalisation does not preclude guidance of action. A notable advantage of Spinozan or near-Spinozan belief formation – adoption before evaluation – is that it allows early action, though at the cost of occasional action on the basis of unreliable information (Egan 2008). When the near-Spinozan view is combined with compartmentalisation, we expect beliefs that are adopted, but not yet evaluated, to guide action in some, but not all, contexts.

On Turner and Coltheart's (2010) account, absence of a tag allows beliefs to guide action without having been evaluated but, even when the unconscious checking system fails, conscious belief evaluation is still possible and may be externally prompted or initiated. The consequences of failure of compartmentalisation are more dramatic, for it is compartmentalisation that allows post-adoption belief evaluation even without new evidence. Without compartmentalisation, initial adoption of a delusional belief would, given consistency and coherence, eliminate the very considerations in the light of which the belief should be rejected. There would be no basis for reflective, critical evaluation of the delusional belief. This is particularly clear in the case of antecedently available considerations but, as we have seen (section 3.2), the value of new evidence that is similar to evidence that has already been accepted is also diminished. This is the basis for McKay's (2012) view that no additional pathology or departure from normality is needed to explain the delusional belief's persistence once it has been adopted.

When we adopt a Bayesian approach, with the standard idealising assumption of a single coherent assignment of credences, persistence seems to be the normal, and

normatively correct, consequence of adoption of the delusional belief. Once we shift to the view of the mind as fragmented, and of compartmentalisation as normal, compartmentalisation of newly adopted beliefs allows post-adoption belief evaluation. So there is a role for a pathology or departure from normality that would distinctively explain the persistence of a delusional belief – even after its initial adoption had been accounted for. We have suggested that impaired executive function or working memory might play that role. Impaired executive function might prevent the patient from stepping back from his initial adoption of the delusional belief; and impaired working memory might not allow the patient to work out the consequences of his prior beliefs. Failure of compartmentalisation takes us closer to the supposed normative ideal of a single coherent assignment of credences. But now it appears as a departure from normality and as another possible explanation of persistence. Failure of compartmentalisation allows updating of credences by conditionalising on the newly adopted belief (that is, on the proposition newly assigned a high credence) and thus eliminates the considerations on which reflective, critical evaluation of that belief should be based.

**Conclusion**

The leading idea of the two-factor framework for explaining delusions is that two factors in the aetiology of delusions provide answers to two questions. First, where did the delusion come from? Second, why does the patient not reject the belief? Answers to the first question have been provided for a range of monothematic delusions, mainly of neuropsychological origin. These answers have a common form: they specify a neuropsychological deficit that is supposed to be related to the content of the delusion in some plausible way. The particular answers inevitably differ from delusion to delusion and may differ between patients with the same delusion.

The second question can be interpreted in more than one way. Interpreted as the adoption question, it asks why a delusional hypothesis that is related to the neuropsychological deficit was adopted as a belief, rather than being rejected. Interpreted as the persistence question, it asks why the initially adopted belief persists, rather than being subsequently rejected. The adoption question and the persistence question both need answers but the fact that there are two interpretations of the second question indicates that there is some unclarity about the role of the second factor in the two-factor framework. In fact, while it has been proposed that the second factor is the same in all cases of delusion of neuropsychological origin, the role, cognitive nature, and neural basis of the second factor have not been well specified (see Coltheart et al. 2010, p. 282).

In sections 2 and 3, we reviewed two Bayesian developments of the two-factor framework (Coltheart et al. 2010; McKay 2012). The two accounts disagree about the processes that lead from the first factor (neuropsychological deficit) to the initial adoption of the delusional belief. According to Coltheart and colleagues, no departure from normality other than the first factor is implicated in the processes leading to adoption of the belief and they take this to be a vindication of 'Maher's basic contention that delusional beliefs *arise* via rational inferential responses to highly unusual data' (2010, p. 277; emphasis added). In contrast, McKay argues that a second factor in the form of an updating bias is required.

The inevitable consequence of this difference is that the two accounts also disagree over the role of the second factor in the aetiology of delusions. Coltheart et al. say that the second factor is not needed to answer the adoption question but is needed to answer the persistence question – option (A). But we were unable to discover a role in a Bayesian account for a pathological factor that would distinctively explain the persistence of a delusional belief after its initial adoption had been accounted for. McKay says that the second factor is needed to answer the adoption question and that no additional departure from normality is needed to answer the persistence question – option (B).

In the central section of the chapter, we showed that, with the standard Bayesian idealising assumption of a single coherent assignment of credences in place, Bayesian belief evaluation could only be updating of credences by conditionalising on new evidence. There is no place for evaluation of a newly adopted belief on the basis of pre-existing considerations. We argued for a relaxation of the idealising assumption in order to allow compartmentalisation of newly adopted beliefs, so that prior credences would remain available to figure in subsequent belief evaluation.

In sections 5 and 6, we returned to the adoption and persistence questions. We proposed an answer to the adoption question in terms of Bayesian inference within an unconscious information-processing system or perceptual module. On our account, the results of the inference process determine the content of an experience. The initial adoption of the delusional belief is a prepotent doxastic response to that experience, on the endorsement option, or the result of a prepotent doxastic tendency to accept an accessible hypothesis that predicts the experience, on the explanation option (see "From anomalous experience to delusional belief" section). We suggested that compartmentalisation should normally be triggered whenever belief formation is near-Spinozan. Thus delusional beliefs would normally be compartmentalised, in line with the oft-remarked circumscription of monothematic delusions. In answer to the persistence question, we argued that evaluation of a delusional belief would normally be a difficult task that might require assistance. The difficulties would be exacerbated by impairments of executive function or working memory (or both) and a more dramatic explanation of persistence would be provided by failure of compartmentalisation itself.

**References**


Aimola Davies, A.M. and Davies, M. 2009: Explaining pathologies of belief. In M.R. Broome and L. Bortolotti (eds), *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*, pp. 285–323. Oxford: Oxford University Press.

Aimola Davies, A.M., Davies, M., Ogden, J.A., Smithson, M. and White, R.C. 2009: Cognitive and motivational factors in anosognosia. In T. Bayne and J. Fernández (eds), *Delusions and Self-Deception: Affective Influences on Belief-Formation*, pp. 187–225. Hove, East Sussex: Psychology Press.

American Psychiatric Association 2000: *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision (DSM–IV–TR)*. Washington, DC: American Psychiatric Association.

Bauer, R.M. 1984: Autonomic recognition of names and faces: A neuropsychological application of the Guilty Knowledge Test. *Neuropsychologia*, 22, 457–69.

Bayne, T. 2009: Perception and the reach of phenomenal content. *Philosophical Quarterly*, 59, 385–404.

Bayne, T. and Pacherie, E. 2004: Bottom-up or topdown? Campbell's rationalist account of monothematic delusions. *Philosophy, Psychiatry, and Psychology*, 11, 1–11.

Bottini, G., Bisiach, E., Sterzi, R. and Vallar, G. 2002: Feeling touches in someone else's hand. *NeuroReport*, 13, 249–52.

Breen, N., Caine, D. and Coltheart, M. 2000: Models of face recognition and delusional misidentification: A critical review. *Cognitive Neuropsychology*, 17, 55–71.

Breen, N., Caine, D. and Coltheart, M. 2001: Mirrored-self misidentifiation: Two cases of focal onset dementia. *Neurocase*, 7, 239–54.

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

Breen, N., Coltheart, M. and Caine, D. 2001: A two-way window on face recognition. *Trends in Cognitive Sciences*, 5, 234–5.

Brighetti, G., Bonifacci, P., Borlimi, R. and Ottaviani, C. 2007: 'Far from the heart far from the eye': Evidence from the Capgras delusion. *Cognitive Neuropsychiatry*, 12, 189–97.

Bruce, V. and Young, A.W. 1986: Understanding face recognition. *British Journal of Psychology*, 77, 305–27.

Capgras, J. and Reboul-Lachaux, J. 1923: L'illusion des 'sosies' dans un délire systématisé chronique. *Bulletin de la Société Clinique de Médicine Mentale*, 2, 6–16.

Carruthers, P. 2006: *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

Chatterjee, A. and Mennemeier, M. 1996: Anosognosia for hemiplegia: Patient retrospections. *Cognitive Neuropsychiatry*, 1, 221–37.

Coltheart, M. 2005: Conscious experience and delusional belief. *Philosophy, Psychiatry, and Psychology*, 12, 153–7.

Coltheart, M. 2007: Cognitive neuropsychiatry and delusional belief. *Quarterly Journal of Experimental Psychology*, 60, 1041–62.

Coltheart, M. 2010: The neuropsychology of delusions. *Annals of the New York Academy of Sciences*, 1191, 16–26.

Coltheart, M., Langdon, R. and McKay, R. 2007: Schizophrenia and monothematic delusions. *Schizophrenia Bulletin*, 33, 642–7.

Coltheart, M., Langdon, R. and McKay, R. 2011: Delusional belief. *Annual Review of Psychology*, 62, 271–98.

Coltheart, M., Menzies, P. and Sutton, J. 2010: Abductive inference and delusional belief. *Cognitive Neuropsychiatry*, 15, 261–87.

Corlett, P.R., Frith, C.D. and Fletcher, P.C. 2009: From drugs to deprivation: A Bayesian framework for understanding models of psychosis. *Psychopharmacology*, 206, 515–30.

Corlett, P.R., Krystal, J.H., Taylor, J.R. and Fletcher, P.C. 2009: Why do delusions persist? *Frontiers in Human Neuroscience*, 9 (12), 1–9.

Corlett, P.R., Murray, G.K., Honey, G.D. et al. 2007: Disrupted prediction-error signal in psychosis: Evidence for an associative account of delusions. *Brain*, 130, 2387–400.

Cotard, J. 1882: Du délire des négations. *Archives de Neurologie*, 4, 152–70, 282–95.

Courbon, P. and Fail, G. 1927: Syndrome d''illusion de Frégoli' et schizophrenie. *Bulletin de la Société Clinique de Médecine Mentale*, 20, 121–5.

David, A.S. 1993: Cognitive neuropsychiatry. *Psychological Medicine*, 23, 1–5.

Davies, M. 2000: Interaction without reduction: The relationship between personal and sub-personal levels of description. *Mind and Society*, 1, 87–105.

Davies, M., Aimola Davies, A.M. and Coltheart, M. Anosognosia and the two-factor theory of delusions. *Mind & Language*, 20, 209–36.

Davies, M. and Coltheart, M. 2000: Introduction: Pathologies of belief. *Mind & Language*, 15, 1–46.

Davies, M., Coltheart, M., Langdon, R. and Breen, N. 2001: Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, and Psychology*, 8, 133–58.

de Pauw, K.W., Szulecka, T.K. and Poltock, T.L. 1987: Frégoli syndrome after cerebral infarction. *Journal of Nervous and Mental Diseases*, 175, 433–8.

Dennett, D.C. 1969: *Content and Consciousness*. London: Routledge and Kegan Paul.

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

Egan, A. 2008: Seeing and believing: Perception, belief formation and the divided mind. *Philosophical Studies*, 140, 47–63

Egan, A. and Elga, A. 2005: I can't believe I'm stupid. *Philosophical Perspectives*, 19, 77–94.

Ellis, H.D. 1998: Cognitive neuropsychiatry and delusional misidentification syndromes: An exemplary vindication of the new discipline. *Cognitive Neuropsychiatry*, 3, 81–90.

Ellis, H.D. and Lewis, M.B. 2001: Capgras delusion: A window on face recognition. *Trends in Cognitive Sciences*, 5, 149–56.

Ellis, H.D., Lewis, M.B., Moselhy, H.F. and Young, A.W. 2000: Automatic without autonomic responses to familiar faces: Differential components of covert face recognition in a case of Capgras delusion. *Cognitive Neuropsychiatry*, 5, 255–69.

Ellis, H.D., Whitley, J. and Luauté, J.P. 1994: Delusional misidentification: The three original papers on the Capgras, Fregoli and intermetamorphosis delusions. *History of Psychiatry,* 5, 117–46.

Ellis, H. D. and Young, A. W. 1990: Accounting for delusional misidentifications. *British Journal of Psychiatry*, 157, 239–48.

Ellis, H.D., Young, A.W., Quayle, A.H. and de Pauw, K.W. 1997: Reduced autonomic responses to faces in Capgras delusion. *Proceedings of the Royal Society: Biological Sciences*, B264, 1085–92.

Fine, C., Craigie, J. and Gold, I. 2005: Damned if you do, damned if you don't: The impasse in cognitive accounts of the Capgras delusion. *Philosophy, Psychiatry, and Psychology*, 12, 143–51.

Fletcher, P.C. and 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. 1983: *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J.A. 1989: Why should the mind be modular? In A. George (ed.), *Reflections on Chomsky*, pp. 1–22. Oxford: Blackwell.

Fodor, J.A. 2000: *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

Friston, K. 2005: A theory of cortical responses. *Philosophical Transactions of the Royal Society: Biological Sciences*, 360, 815–36.

Friston, K. 2009: The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences*, 13, 293–301.

Friston, K. 2010: The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11, 127–38.

Friston, K. and Stephan, K. E. 2007: Free-energy and the brain. *Synthese*, 159, 417–58.

Frith, C.D. 1992: *The Cognitive Neuropsychology of Schizophrenia*. Hove, East Sussex: Lawrence Erlbaum Associates.

Frith, C.D. 2007: *Making Up the Mind: How the Brain Creates Our Mental World*. Oxford: Blackwell Publishing.

Frith, C.D. and Done, D.J. 1989: Experiences of alien control in schizophrenia reflect a disorder in the central monitoring of action. *Psychological Medicine*, 19, 359–63.

Gilbert, D.T. 1991: How mental systems believe. *American Psychologist*, 46, 107–19.

Gilbert, D.T., Krull, D.S. and Malone, P.S. 1990: Believing the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, 59, 601–13.

Gilbert, D.T., Tafadori, R.W. and Malone, P.S. 1993: You can't not believe everything you read. *Journal of Personality and Social Psychology*, 65, 221–33.

Gold, I. and Stoljar, D. 1999): A neuron doctrine in the philosophy of neuroscience. *Behavioral and Brain Sciences*, 22, 809–69.

Greve, K.W. and Bauer, R.M. 1990: Implicit learning of new faces in prosopagnosia: An application of the mere-exposure paradigm. *Neuropsychologia*, 28, 1035–41.

Halligan, P.W. and David, A.S. 2001: Cognitive neuropsychiatry: Towards a scientific psychopathology. *Nature Reviews Neuroscience*, 2, 209–15.

Halligan, P.W., Marshall, J.C. and Wade, D.T. 1995: Unilateral somatoparaphrenia after right hemisphere stroke: A case description. *Cortex*, 31, 173–82.

Hasson, U., Simmons, J.P. and Todorov, A. 2005: Believe it or not: On the possibility of suspending belief. *Psychological Science*, 16, 566–71.

Hawley, K. and Macpherson, F. 2010: *The Admissible Contents of Experience*. Oxford: Wiley-Blackwell.

Hemsley, D.R. and Garety, P.A. 1986: The formation and maintenance of delusions: A Bayesian analysis. *British Journal of Psychiatry*, 149, 51–6.

Hirstein, W. and 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: Biological Sciences*, B264, 437–44.

Hohwy, J. and Rosenberg, R. 2005: Unusual experiences, reality testing and delusions of alien control. *Mind & Language*, 20, 141–62.

Jeffrey, R. 1965: *The Logic of Decision*. New York: McGraw-Hill (Second Edition, Chicago, IL: University of Chicago Press, 1983).

Langdon, R. and Bayne, T. 2010: Delusion and confabulation: Mistakes of perceiving, remembering and believing. *Cognitive Neuropsychiatry*, 15, 319–45.

Langdon, R. and Coltheart, M. 2000: The cognitive neuropsychology of delusions. *Mind & Language*, 15, 184–218.

Levine, J. 1983: Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly,* 64, 354–61.

Lewis, D. 1982: Logic for equivocators. *Noûs*, 16, 431–41.

Lewis, M.B. and Ellis, H.D. 2001: A two-way window on face recognition: Reply to Breen et al. *Trends in Cognitive Sciences*, 5, 235.

Lipton, P. 2004: *Inference to the Best Explanation* (Second Edition). London: Routledge.

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

Maher, B.A. 1988: Anomalous experience and delusional thinking: The logic of explanations. In T.F. Oltmanns and B.A. Maher (eds), *Delusional Beliefs*, pp. 15–33. Chichester: John Wiley and Sons.

Maher, B.A. 1992: Delusions: Contemporary etiological hypotheses. *Psychiatric Annals*, 22, 260–8.

Maher, B.A. 1999: Anomalous experience in everyday life: Its significance for psychopathology. *The Monist*, 82, 547–70.

Mandler, G. 1980: Recognizing: The judgment of previous occurrence. *Psychological Review*, 87, 252–71.

McKay, R. 2012: Delusional inference. *Mind & Language*, 27, 330–55.

Moore, G.E. 1939: Proof of an external world. *Proceedings of the British Academy*, 25, 273–300. Reprinted in *Philosophical Papers*. London: Allen and Unwin, 1959, 127–50.

Nagel, T. 1974: What is it like to be a bat? *Philosophical Review*, 83, 435–50. Reprinted in T. Nagel, *Mortal Questions*, pp. 165–80. Cambridge: Cambridge University Press, 1979.

Psillos, S. 2009: An explorer upon untrodden ground: Peirce on abduction. In D.M. Gabbay, S. Hartmann and J. Woods (eds), *Handbook of the History of Logic, Volume 10: Inductive Logic*, pp. 117–51. Amsterdam: Elsevier BV.

Shea, N, 2012: Reward prediction error signals are meta-representational. *Noûs*.

Shea, N. forthcoming: Distinguishing top-down from bottom-up effects. In S. Biggs, M. Matthen, and D. Stokes (eds), *Perception and Its Modalities*. Oxford: Oxford University Press.

Siegel, S. 2010: *The Contents of Visual Experience*. Oxford: Oxford University Press, 2010.

Stone, T. and Young, A.W. 1997: Delusions and brain injury: The philosophy and psychology of belief. *Mind & Language*, 12, 327–64.

Sperber, D. 2001: In defense of massive modularity. In E. Dupoux (ed.) *Language, Brain, and Cognitive Development: Essays in Honor of Jacques Mehler*, pp. 47–57. Cambridge, MA: MIT Press.

Sperber, D. and Wilson, D. 1986: *Relevance: Communication and Cognition* (Second Edition, 1995). Oxford: Blackwell Publishing.

Sperber, D. and Wilson, D. 2002: Pragmatics, modularity and mind-reading. *Mind & Language*, 17, 3–23.

Stone, T. and Young, A.W. 1997: Delusions and brain injury: The philosophy and psychology of belief. *Mind & Language*, 12, 327–64.

Tranel, D. and Damasio, A.R. 1985: Knowledge without awareness: An autonomic index of facial recognition by prosopagnosics. *Science*, 228, 1453–4.

Tranel, D. and Damasio, A.R. 1988: Non-conscious face recognition in patients with face agnosia. *Behavioural Brain Research*, 30, 235–49.

Tranel, D., Damasio, H. and Damasio, A.R. 1995: Double dissociation between overt and covert recognition. *Journal of Cognitive Neuroscience*, 7, 425–32.

Turner, M. and Coltheart, M. 2010: Confabulation and delusion: A common monitoring framework. *Cognitive Neuropsychiatry*, 15, 346–76.

van Fraassen, B.C. (1980) *The Scientific Image*. Oxford: Oxford University Press.

van Fraassen, B.C. (1989) *Laws and Symmetry*. Oxford: Oxford University Press.

Whittlesea, B.W.A. 1993: Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1235–53.

Wilson, D. and Sperber, D. 2012: *Meaning and Relevance*. Cambridge: Cambridge University Press.

Young, A.W. and Leafhead, K.M. 1996: Betwixt life and death: Case studies of the Cotard delusion. In P.W. Halligan and J.C. Marshall (eds), *Method in Madness: Case Studies in Cognitive Neuropsychiatry*, pp. 147–71. Hove, East Sussex: Psychology Press.

Young, G. 2007: Clarifying 'familiarity': Phenomenal experiences in prosopagnosia and the Capgras delusion. *Philosophy, Psychiatry, and Psychology*, 14, 29–37.

Young, G. 2008: Capgras delusion: An interactionist model. *Consciousness and Cognition*, 17, 863–76.