Delusional Inference

Ryan McKay

ARC Centre of Excellence in Cognition and its Disorders
Department of Psychology, Royal Holloway, University of London

---------- To appear in *Mind & Language* ----------

Please address correspondence to:
Dr Ryan McKay
ARC Centre of Excellence in Cognition and its Disorders
Department of Psychology
Royal Holloway, University of London
Egham, Surrey TW20 0EX
United Kingdom
ryantmckay@mac.com

Acknowledgements:
I thank Peter Brugger, Phil Corlett, Zoltan Dienes, Charles Efferson, Robyn Langdon, Frédéric Schneider and Mark Wildon for helpful discussions. I am indebted to Martin Davies and an anonymous reviewer for detailed feedback on a draft of this manuscript. Special thanks to Max Coltheart for valuable comments, advice and encouragement.

Abstract

Does the formation of delusions involve abnormal reasoning? According to the prominent 'two-factor' theory of delusions (e.g., Coltheart, 2007), the answer is yes. The second factor in this theory is supposed to affect a deluded individual's ability to evaluate candidates for belief. However, most published accounts of the two-factor theory have not said much about the nature of this second factor. In an effort to remedy this shortcoming, Coltheart, Menzies and Sutton (2010) recently put forward a Bayesian account of inference in delusions. I outline some criticisms of this important account, and sketch an alternative account of delusional inference that, I argue, avoids these criticisms. Specifically, I argue that the second factor in delusion formation involves a systematic deviation from Bayesian updating, a deviation that may be characterized as a bias towards 'explanatory adequacy'. I present a numerical model of this idea and show that my alternative account is broadly consistent with prominent prediction error models of delusion formation (e.g., Corlett, Murray et al., 2007).

1. Introduction

The notion that the formation of delusions involves an inferential step is not new, and is enshrined in the definition of delusion provided by the *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision* (DSM-IV-TR; American Psychiatric Association, 2000): delusions are 'based on incorrect inference about external reality' (p. 821). The claim that the inference concerned is an inference about 'external reality' is easily disputed[1], but the stipulation that the inference be 'incorrect' is controversial. To be sure, delusions can be extravagantly at variance with reality, but this does not prove that faulty inference is involved in their formation. After all, a valid inference from unsound premises may well yield a false conclusion.[2]

Maher (e.g., 1992; 1999; Maher & Ross, 1984) has prominently argued that the inferential reasoning of deluded individuals is not significantly different to that of healthy individuals. Delusions, for Maher, are not a product of defective inferential reasoning, but reflect an attempt to explain intense and unusual phenomenological experiences. In Maher's view, what is pathological in delusions is not the reasoning that is brought to bear on experiences, but the experiences themselves: 'the locus of the pathology is in the neuropsychology of experience' (Maher, 1999).

Coltheart and colleagues (e.g., Coltheart, 2005; 2007; Coltheart, Langdon, & McKay, 2007, 2011; Davies & Coltheart, 2000; Davies, Coltheart, Langdon, & Breen, 2001; Langdon & Coltheart, 2000) agree that abnormal cognitive data[3] may comprise a necessary factor in delusion formation. Indeed, for many delusions abnormal cognitive data have been identified or hypothesised that plausibly furnish the relevant delusional content (see Table 1). These authors argue, however, that abnormal cognitive data cannot be sufficient for delusion formation, because there are cases where the data in question appear to be present but where a delusion is absent (see fourth column of Table 1). Following important work by Ellis, Stone and Young (e.g., Ellis & Young, 1990; Stone & Young, 1997), Coltheart and colleagues thus postulate a second factor in delusion formation, a factor that affects a deluded individual's ability to evaluate beliefs.[4] Although most published accounts of the two-factor theory have not said much about the nature of this second factor[5], one might assume that the second factor proposal is at odds with Maher's contention that delusions arise via normal inferential responses to highly abnormal data. In a

[1] Certain delusions - delusions of thought insertion, for instance – clearly pertain to *internal* reality.

[2] Conversely, an invalid inference might yield a belief that is accidentally or serendipitously true (see Coltheart, 2009; Jaspers, 1946/1963; Spitzer, 1990).

[3] Coltheart et al. (2010; cf. Young, 2011) eschew the use of the word 'experience', noting that the cognitive abnormalities that initially prompt delusional hypotheses are not always consciously accessible (whereas the term 'experience' is used by many to refer only to mental events of which a person is conscious).

[4] Coltheart et al. (2010) adopt a 'doxastic' conception of delusions - they assume that delusions are beliefs. I also make this assumption (see Bayne & Pacherie, 2005, and Bortolotti, 2009, for defences of the doxastic conception).

[5] Aimola Davies and Davies (2009) recently suggested that the second factor is an impairment of working memory or executive function. See Coltheart (2007) regarding the possible neural basis of the second factor.

recent paper, however, Coltheart et al. (2010) put forward a Bayesian account of delusional inference that, they suggest, *vindicates* Maher's contention. In my opinion the account they offer marks a real advance in thinking about the second factor in delusion formation, but incorporates some problematic assumptions. In what follows I will summarise their account (Section 2), highlight the problems that I see with it (Section 3), and then sketch an alternative account – drawing on Stone and Young (1997) and Aimola Davies and Davies (2009) - that, I will argue, solves these problems (Section 4). In particular, I will show that the alternative account, if correct, vindicates not Maher's view but the view that delusions involve abnormal inference. I will also show that my alternative account is broadly consistent with prominent prediction error models of delusion formation (e.g., Corlett, Honey, & Fletcher, 2007; Corlett, Murray et al., 2007; Corlett, Krystal, Taylor, & Fletcher, 2009; Fletcher & Frith, 2009).

A word about terminology: The Bayesian conception of belief is a probabilistic conception - one's belief that *p* is represented by a value between 0 and 1, reflecting one's degree of confidence in the proposition *p*. This conception can be contrasted with the binary, all-or-nothing conception of belief that standard accounts of the two-factor theory are predicated upon. On the binary conception, one either believes that *p* or doesn't believe that *p*.[6] Coltheart et al. (2010) employ both conceptions, but do not indicate how they are to be reconciled. This is problematic because there are deep inconsistencies between the two conceptions. For example, on the binary conception the notion of contradictory beliefs is straightforward – if one believed both *p* and its negation one would be caught in such a contradiction. On the probabilistic conception, however, it is normal to 'believe' two or more contradictory propositions, provided that the respective probabilistic beliefs sum to one.

In what follows I will unify the two conceptions by treating binary belief as a special case of probabilistic belief. In particular I will take what Christensen (2004, p. 32) calls a 'sub-certainty threshold approach' to unification, by equating binary belief with a probabilistic belief that exceeds 0.5. To avoid confusion I will in general reserve the term 'belief' for those cases where the probability assigned to a hypothesis is greater than this arbitrary threshold. Thus I will speak of 'adopting', 'retaining' and 'rejecting' hypotheses as beliefs. I will consider a hypothesis (and, by extension, a belief) to be true if the content of the proposition corresponds to reality and false if its content doesn't correspond to reality.

[6] Given that 'not believing that *p*' can encompass both cases where one actively *dis*believes that *p* (i.e., where one believes that not-*p*), and cases where one withholds judgement, some might prefer to see discrete belief as a *trinary* notion (Christensen, 2004).

TABLE 1

Neuropsychological impairments with and without delusion (from Coltheart et al., 2010, pp. 267-8; reprinted by permission of the publisher, Taylor & Francis Group, http://www.informaworld.com)

Neuropsychological impairment	*Abnormal data generated by neuropsychological impairment*	*Hypothesis abductively inferred to explain the abnormal data*	*Cases where the neuropsychological impairment is present but there is no delusion*
Disconnection of face recognition system from autonomic system	Highly familiar faces (e.g. wife's face) no longer produce autonomic response	'That's not my wife; it is some stranger who looks like her' (Capgras delusion)	Tranel, Damasio, & Damasio, 1995.
Autonomic system over-responsive to faces	Even the faces of strangers produce autonomic responses	'People I know are following me around; I don't recognize them; they are in disguise' (Frégoli delusion)	Patient experiences faces of strangers as highly familiar (Vuilleumier, Mohr, Valenza, Wetzel, & Landis, 2003).
Autonomic system under-responsive to all stimuli	No emotional response to one's environment	'I am dead' (Cotard delusion)	Pure autonomic failure: patient is not autonomically responsive to any stimuli (Heims, Critchley, Dolan, Mathias, & Cipolotti, 2004; Magnifico, Misra, Murray & Mathias, 1998)
Mirror agnosia (loss of knowledge about how mirrors work)	Mirrors are treated as windows	'This person can't be me (because I am looking at him through a window, so he is in a different part of space than I)' (Mirrored-self misidentification)	Other cases of mirror agnosia (Binkofski, Buccino, Dohle, Seitz, & Freund, 1999)
Impaired face processing	When the patient looks into a mirror, the mental representation constructed of the face that is seen there does not match the representation of the patient's face that is stored in long-term memory	'The person I am looking at in the mirror isn't me' (Mirrored-self misidentification)	Many cases of prosopagnosia
Failure of computation of sensory feedback from movement	Voluntary movements no longer are sensed as voluntary	'Other people can cause my arm to move' (Delusion of alien control)	Haptic deafferentation: patient gets no sensory feedback from any actions performed (Fourneret, Paillard, Lamarre, Cole, & Jeannerod, 2002)
Damage to motor area of brain controlling arm	Left arm is paralysed; patient can't move it.	'This arm [the patient's left arm] isn't mine; it is [some specified other person]'s' (Somatoparaphrenia)	Many cases of left-sided paralysis

2. A Bayesian Account of Inference in Delusions

Coltheart et al. (2010) develop their account with reference to Capgras delusion, which involves the belief that a close friend or relative has been replaced by a physically identical impostor. Coltheart et al. endorse a prominent cognitive neuropsychological theory of this delusion (see Ellis & Young, 1990; Stone & Young, 1997). According to this theory, the delusion stems from a neuropsychological disconnection of the face recognition system (itself intact) from the autonomic nervous system (itself intact). Individuals with this disconnection would lack the ordinary autonomic response to familiar faces. Put another way, faces that are visually familiar would be rendered autonomically unfamiliar (McKay & Kinsbourne, 2010). Coltheart et al. offer a Bayesian account of the inference from these discrepant familiarity data to the belief that a stranger is impersonating one's friend or family member.

Imagine that a man suffers a minor stroke on his way home from work one day and that his face recognition system becomes disconnected from his autonomic nervous system as a result. Let *o* represent the incongruous data available to him when encountering his wife at home that evening:

> *o*: visual familiarity data in the absence of autonomic familiarity data

How is this gentleman to explain *o*? Assume that there are just two mutually exclusive hypotheses under consideration:[7]

> h_w: That woman who looks like my wife and claims to be my wife is in fact my wife
> h_s: That woman who looks like my wife and claims to be my wife is a stranger

Let $P(h_w)$ represent the initial ('prior') probability assigned to the 'wife hypothesis' h_w *before* he encounters her that evening, and let $P(h_s)$ represent the prior probability assigned to the 'stranger hypothesis' h_s. These two priors, $P(h_w)$ and $P(h_s)$, are *prior to* the data *o*. Having observed *o*, which hypothesis is more probable? Another way of putting this is to ask which *posterior* probability (revised so as to take the data into account) is greater, $P(h_w \mid o)$ or $P(h_s \mid o)$?

Bayes' theorem (Bayes, 1763) specifies the optimal procedure for updating the probabilities assigned to hypotheses in the light of new evidence. The posterior probability of the wife hypothesis using Bayes' theorem is:

$$P(h_w \mid o) = \frac{P(h_w) \times P(o \mid h_w)}{P(o)} \qquad (1)$$

[7] As noted in fn. 3, the abnormal data that initially prompt delusional hypotheses are not always accessible to consciousness. Although I speak here of the 'gentleman', I make no assumptions about the extent to which the generation and evaluation of candidate hypotheses are conscious, person-level processes. In reality Bayesian inference is likely to be occurring simultaneously at many levels of a hierarchy (see Fletcher & Frith, 2009).

Likewise, the posterior probability of the stranger hypothesis is:

$$P(h_s \mid o) = \frac{P(h_s) \times P(o \mid h_s)}{P(o)}\text{8} \tag{2}$$

In each case the posterior probability *P*(*h* | *o*) is proportional to the product of the prior probability *P*(*h*) and another term known as the *likelihood*, *P*(*o* | *h*). The likelihood denotes the probability of observing the data if the hypothesis in question were true – it represents how likely *o* would be under *h*. This term can be viewed as a measure of the explanatory power of the hypothesis with respect to *o* (Coltheart et al., 2010).[9,10]

Coltheart et al. (2010) note that the posterior probability of the stranger hypothesis *P*(h_s | *o*) can be much higher than the posterior probability of the wife hypothesis *P*(h_w | *o*), *provided* that the relative explanatory power of the former offsets its relatively low prior probability. If this condition obtains, a rational individual (one who updates probabilities in accordance with the Bayesian prescription) will adopt the belief that the woman in his home is a stranger. A good way to see this is to consider the ratio of the two posterior probabilities, the *posterior odds*. The posterior odds in favour of the stranger hypothesis h_s are:

$$\frac{P(h_s \mid o)}{P(h_w \mid o)} = \frac{P(h_s)}{P(h_w)} \times \frac{P(o \mid h_s)}{P(o \mid h_w)} \tag{3}$$

The first and second terms to the right of this equation are known respectively as the

[8] Given that there are just two mutually exclusive hypotheses under consideration here, *P*(h_s | *o*) is also equal to 1 - *P*(h_w | *o*).

[9] According to Coltheart et al's 'probabilistic account of explanation' (2010, p. 271), the value of the likelihood *P*(*o* | *h*) quantifies the degree to which a hypothesis *h* explains an observation *o*. Coltheart et al. thus equate explanatory power with likelihood, and in this paper I also adopt this conception of explanation (later I will speak of explanatory 'adequacy'). There are, however, other views. For example, although Lipton (2004) notes that explanatory considerations may inform our estimates of likelihood because 'lovelier explanations tend to make what they explain likelier', he adds that 'high likelihood is no guarantee of good explanation' (p. 114; see also Aimola Davies & Davies, 2009). The observation *o* that my wife has two legs would be very likely under the hypothesis *h* that she is having an affair, so the likelihood *P*(*o* | *h*) is extremely high. But it is hard to see how the hypothesis that she is having an affair provides any explanation whatsoever (where explanation is understood in the sense of providing understanding) for the observed fact that she has two legs.

[10] The term *P*(*o*) denotes the prior probability of encountering the data *o*. In Bayes' theorem this term functions as a normalising constant, ensuring that the posterior probabilities of all hypotheses under consideration sum to 1.

prior odds and the *likelihood ratio*.[11] Coltheart et al's (2010) point is that if the likelihood ratio is sufficiently high, the posterior odds in favour of the stranger hypothesis can be high even if the prior odds are low.

So far the account of Coltheart et al. (2010) is consistent with Maher's approach. Provided that the relative power of the stranger hypothesis to explain the discrepant familiarity data offsets the low prior probability of that hypothesis, the adoption of the impostor[12] belief in the presence of these very abnormal data is 'a perfectly rational response' (Coltheart et al., 2010, p. 281).

The notion that discrepancies between visual and autonomic familiarity data figure in the aetiology of Capgras delusion has received empirical support. A series of studies have now shown that whereas control participants show a pattern of autonomic discrimination (indexed by skin conductance response) between familiar and unfamiliar faces, Capgras patients do not (e.g., Brighetti, Bonifacci, Borlimi, & Ottaviani, 2007; Ellis, Young, Quayle, & de Pauw, 1997; Hirstein & Ramachandran, 1997). Other work, however, suggests that such discrepancies are not *sufficient* for the delusion to develop. Tranel, Damasio and Damasio (1995) have shown that although patients with damage to ventromedial frontal regions of the brain also fail to show a pattern of autonomic discrimination between familiar and unfamiliar faces, these patients do not evince Capgras delusion. Coltheart et al. thus need an account of why these patients are different to the Capgras patients. If the impostor inference is the rational inference to make in the wake of discrepant familiarity data, why do the ventromedial frontal patients not make it?

Coltheart et al. (2010) argue that, as time passes, individuals with the neuropsychological disconnection we have been considering will be confronted with new data relevant to the stranger hypothesis they have adopted as a belief – data that should undermine it:

> For example, the subject might learn that trusted friends and family believe the person is his wife, that this person wears a wedding ring that has his wife's initials engraved in it, that this person knows things about the subject's past life that only his wife could know, and so on. (Coltheart et al., 2010, p. 279).

Coltheart et al. specify that these further data be represented by o^*. They argue that whereas the original data, o, were more likely to be observed if the stranger hypothesis were true than if the wife hypothesis were true (i.e., $P(o|h_s) > P(o|h_w)$), the new data, o^*, should be more likely to be observed if the wife hypothesis were

[11] Because the term $P(o)$ is a constant denominator in equations (1) and (2) it is cancelled out in the multiplication process.

[12] Here I conflate the hypothesis that the woman who looks like one's wife is a *stranger* with the hypothesis that one's wife has been replaced by an *impostor*. When considering Capgras delusion it is often important to keep these hypotheses distinct; after all, there is an important difference between resembling one's wife and *posing as* one's wife. In Coltheart et al's (2010) paper, however, h_s is the hypothesis that the woman who looks like one's wife and *claims to be* one's wife is a stranger. I have kept this formulation of the stranger hypothesis and so use 'stranger' and 'impostor' interchangeably.

true than if the stranger hypothesis were true (i.e., $P(o^*|h_w) > P(o^*|h_s)$). Given o^*, the explanatory power of the stranger hypothesis is relatively weak. In the wake of the new evidence o^*, a rational participant should therefore reject h_s and adopt h_w. According to Coltheart et al., the patients with ventromedial frontal lesions update their beliefs in a rational manner, and end up adopting h_w (the non-delusional belief). Of course, this claim raises a new question – in the wake of o^*, why don't the Capgras patients also reject h_s and adopt h_w? This is where Coltheart et al. invoke a second factor. In both the ventromedial frontal and the Capgras patients the face recognition system has become disconnected from the autonomic nervous system. This is Factor One. Coltheart et al. argue, however, that the Capgras patients – but not the ventromedial frontal patients - suffer a second impairment, an impaired ability to consider new evidence so as to revise current beliefs. This is Factor Two. Because of this second impairment, the Capgras patients reject or discount the evidence o^*, and retain the delusional stranger belief (see Figure 1).

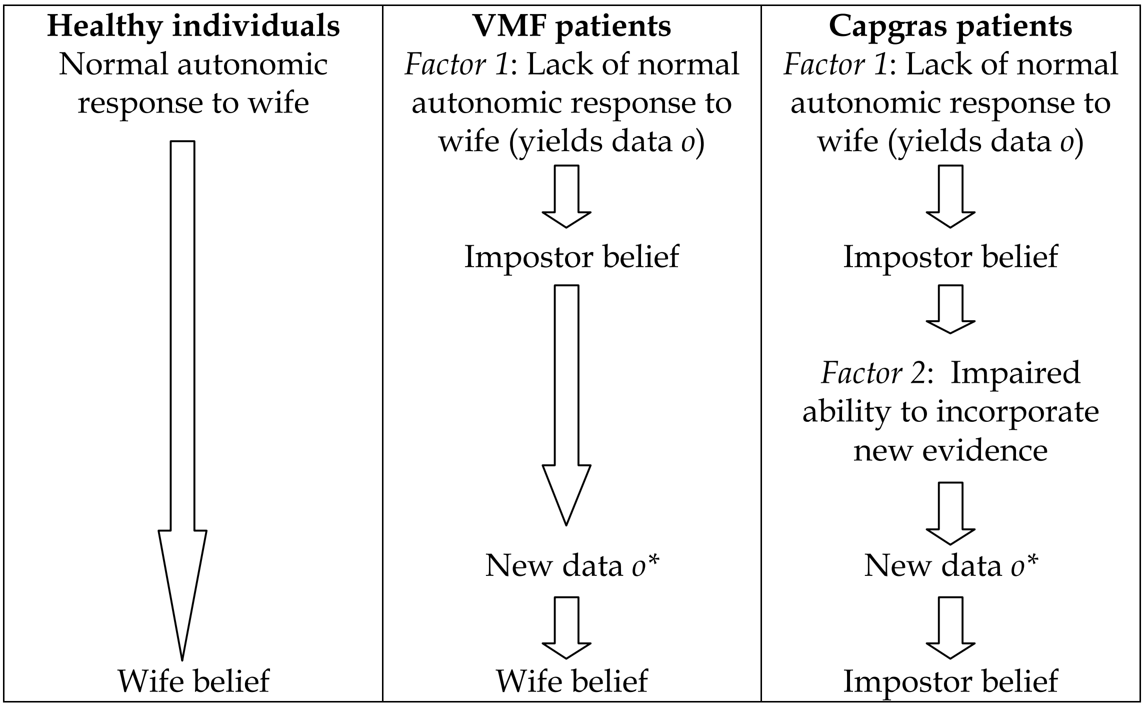
Healthy individuals
Normal autonomic response to wife
Wife belief
VMF patients
Factor 1: Lack of normal autonomic response to wife (yields data o)
Impostor belief
New data o*
Wife belief
Capgras patients
Factor 1: Lack of normal autonomic response to wife (yields data o)
Impostor belief
Factor 2: Impaired ability to incorporate new evidence
New data o*
Impostor belief

Figure 1: Summary of the account of Coltheart et al. (2010)

In summary, Coltheart et al. (2010) propose that the formation of a delusion requires two cognitive impairments. The first of these impairments (Factor One) will vary from delusion to delusion. In the case of Capgras delusion the first impairment involves the disconnection of the face recognition system from the autonomic nervous system, and yields discrepancies between visual and autonomic familiarity data (see Table 1 for a list of first factor neuropsychological impairments that putatively underpin a series of other delusions). The second impairment (Factor Two), however, represents a failure to adequately incorporate relevant evidence in belief revision, and is the same across different delusions.

3. Problems with the Account of Coltheart et al. (2010)

The account of delusional inference put forward by Coltheart et al. (2010) is vivid and detailed. There are, however, a number of potential problems with the story they tell. In this section I detail three such problems, in order of increasing seriousness; in the next section I sketch an alternative account that, I argue, avoids these problems.

3.1 Is the 'Stranger Inference' Reasonable?

As I noted above, Coltheart et al. (2010) consider that the 'Stranger Inference', the inference to h_s, is a rational inference to make in the wake of discrepant familiarity data. Again, they show that the posterior probability of h_s can be much higher than the posterior probability of the wife hypothesis h_w, provided that the relative explanatory power of h_s offsets its low prior probability. The numbers chosen by Coltheart et al. (2010) to illustrate this state of affairs are as follows: $P(h_s) = 0.01$, $P(h_w) = 0.99$, $P(o \mid h_s) = 0.999$ and $P(o \mid h_w) = 0.001$ (see second column of Table 2).

TABLE 2
Models of inference and resulting beliefs

	Coltheart et al. (2010) input, Bayesian inference*	Relatively realistic priors, Bayesian inference*	Relatively realistic priors, bias towards explanatory adequacy**
$P(h_s)$	0.01	0.00027	0.00027
$P(h_w)$	0.99	0.99973	0.99973
$P(o \mid h_s)$	0.999	0.999	0.999
$P(o \mid h_w)$	0.001	0.001	0.001
x	0 (Bayesian inference)	0 (Bayesian inference)	0.02 (adequacy bias)
$P(h_s \mid o)$	0.91	0.21	0.91
$P(h_w \mid o)$	0.09	0.79	0.09
Capgras Delusion?	Yes	No	Yes

h_s: 'That woman who looks like my wife and claims to be my wife is a stranger'.
h_w: 'That woman who looks like my wife and claims to be my wife is in fact my wife'.
o: Visual familiarity data in the absence of autonomic familiarity data.
x: A value in the interval (-1, +1). When $x = 0$ the individual has Bayesian beliefs; when $x < 0$ the individual is doxastically conservative; when $x > 0$ the individual is biased towards explanatory adequacy. See Appendix A.
* See section 3.1 for workings.
** See Appendix B for workings.

In this case, it is true that the likelihood ratio in favour of the stranger hypothesis exceeds the prior odds in favour of the wife hypothesis. In fact, the former (0.999/0.001 = 999) exceeds the latter (0.99/0.01 = 99) by a factor of ten, and the posterior probability of the stranger hypothesis, $P(h_s \mid o) \approx 0.91$, is more than ten times greater than the posterior probability of the wife hypothesis, $P(h_w \mid o) \approx 0.09$:

$$\frac{P(h_s \mid o)}{P(h_w \mid o)} = \frac{P(h_s)}{P(h_w)} \times \frac{P(o \mid h_s)}{P(o \mid h_w)} \approx 10.09 \tag{4}$$

However, are these numbers plausible? $P(h_s)$ = 0.01 seems unrealistically high - astonishingly high, in fact. A prior of 0.01 for h_s means that the probability the individual assigns to the hypothesis that his wife has been replaced by an impostor at any given point - *before he is confronted with any anomalous data* - is one in a hundred. If this individual encounters his wife only once a day, he will expect her to be replaced by an impostor more than three times a year on average. It seems doubtful that any normal individual would have such a prior. Although there are cases in real life where people are fooled by confidence tricksters posing as their loved ones (see, for example, Grann, 2008), such cases are exceptionally rare. The stranger hypothesis h_s thus represents an exceedingly unlikely – almost miraculous - state of affairs:

> If a miracle is defined as a violation of a law of nature, then a person should assign the hypothesis that a law has been violated a very low prior probability. Given this low prior probability, the posterior probability will be low too even though the likelihood function has a high value. (Coltheart et al., 2010, p. 273).

The replacement of one's wife by a physically identical impostor is a fantastically unlikely occurrence. $P(h_s)$ should thus be vanishingly small and the prior odds in favour of the wife hypothesis should be enormous.

Let's look at what a Bayesian with a more realistic (although still unusual) prior distribution over the two hypotheses would conclude, having observed *o*. I suggested above that a realistic prior for h_s would be vanishingly small, but if we set $P(h_s)$ = 0.00027, in which case our individual will expect his wife to be replaced by an impostor about once every ten years on average (again assuming that he encounters his wife only once a day), then his posterior for the stranger hypothesis, $P(h_s \mid o) \approx$ 0.21, will be just 0.27 of that for the wife hypothesis, $P(h_w \mid o) \approx$ 0.79, and he will reject h_s (see third column of Table 2):[13]

$$\frac{P(h_s \mid o)}{P(h_w \mid o)} = \frac{P(h_s)}{P(h_w)} \times \frac{P(o \mid h_s)}{P(o \mid h_w)} \approx 0.27 \tag{5}$$

In this section I have questioned the claim that adoption of the impostor belief in the presence of discrepant familiarity data is 'a perfectly rational response' to those data (Coltheart et al., 2010, p. 281). It's a rational response given certain values of the relevant priors and likelihoods, but some of those values seem highly implausible. There are, however, at least two responses that Coltheart et al. might make to the objections I have raised here. First, they might acknowledge that the priors they

[13] Here I have retained Coltheart et al's figures for the two likelihoods, i.e., $P(o \mid h_s)$ = 0.999 and $P(o \mid h_w)$ = 0.001.

assigned to h_s and h_w were unrealistic, yet maintain that the relative explanatory power of h_s might nevertheless offset its low prior probability. After all, the prior odds in favour of the wife hypothesis might well be enormous, but if the likelihood ratio in favour of the stranger hypothesis is even more enormous then a rational individual would adopt h_s. If the data are especially salient, a reasonable individual might consider that the stranger hypothesis more than makes up in relative explanatory power what it lacks in prior plausibility:

$$\frac{P(o \mid h_s)}{P(o \mid h_w)} > \frac{P(h_w)}{P(h_s)} \tag{6}$$

A second response Coltheart et al. might make is to highlight a weaker sense in which delusions can be said to arise via inferential responses that are, if not quite rational, then at least normal or legitimate (see Coltheart et al., 2011). This sense involves distinguishing between the inferential processes that *generate* hypotheses - candidates for belief - and the inferential processes by which such candidates are *evaluated* (and 'adopted' or 'rejected', in binary shorthand). As Coltheart et al. (2010) note (see section 4.1 below), the hypotheses adopted as beliefs in many cases of delusion *explain* the unusual data that stimulate them – these data are what one would expect to observe if the hypotheses in question were true (i.e., the likelihoods of the respective data are high). In this sense, therefore, the hypotheses generated are reasonable or legitimate *candidates* for belief, given the data in question.

The question of how hypotheses are generated is an important one, and the notion that considerations of explanatory power (again, see section 4.1 below) inform this process is compelling.[14] However, although Coltheart et al. (2010) do repeatedly mention the *generation* of hypotheses that give rise to delusional beliefs, the most straightforward reading of their paper is as claiming that hypothesis *evaluation* is initially rational in delusion formation, and that is the position I have critiqued in this section.[15]

3.2 Ventromedial Frontal Patients

As Coltheart et al. (2010) acknowledge, their account 'involves the conjecture that even the ventromedial patients initially experience the belief "This is not my wife" and only later abandon this belief in the face of contradictory evidence' (p. 281). Coltheart et al. admit that they know of no evidence that ventromedial frontal

[14] In a related suggestion, Hohwy (2010) notes that in generating candidate hypotheses people are guided by a property transmission heuristic, such that the properties of hypothesized causes mirror the properties of the data these hypotheses are being generated to explain. For example, if the datum is an indentation in some clay with a particular shape, it would be natural to generate the hypothesis that the indentation was caused by an object with the same shape (White, 2009).

[15] Coltheart et al. (2010) state quite explicitly that hypothesis generation is not something that the Bayesian account applies to, and not something that can be assessed as rational or not: '[T]he account says nothing about where hypotheses come from. In the account, the provenance of hypotheses is not a matter for rational assessment. The account simply tells us when it is rational to accept a hypothesis on the basis of evidence, however the hypothesis is generated' (2010, p. 274).

patients ever adopt the stranger belief. Nevertheless, this remains an empirical possibility. Another point that counts against this conjecture, however, stems from Coltheart et al's account of how ventromedial frontal patients might give up h_s once confronted with the new data represented by o^*. The problem is that the new data o^* are presumably coupled with the existing data o. After all, as Coltheart et al. say, 'the failure of autonomic response to a spouse's face is omnipresent' (p. 283) – so those autonomic data would still be present once family and clinicians start offering their conflicting testimony. It may be true that, for ventromedial frontal patients, 'the wife hypothesis explains the new data o^* much, much better than the stranger hypothesis and hence $P^*(o^*|h_w)$ is much higher than $P^*(o^*|h_s)$' (p. 279). It's not at all clear, however, that the wife hypothesis would explain *the conjunction of o and o** better than the stranger hypothesis. If the ventromedial frontal patients had initially adopted the impostor belief, then the initial data o must have been salient enough to override the testimony of the wife herself, because $P(h_s)$ represented the prior probability assigned to the hypothesis that the woman who looked like the wife *and claimed to be the wife* was a stranger. Would these data not also override the testimony of friends and family (and the other new data represented by o^*)?

3.3 Implausible Chronology

As discussed above, Coltheart et al. (2010) need an account of why Capgras sufferers, exposed to the same new data as the ventromedial frontal patients, respond differently to those patients. 'Why do they not reject the stranger belief on the basis of these new data that disconfirm it?' (p. 279). Why do they 'disregard or reject all evidence that is inconsistent with [the stranger] hypothesis' (p. 280)? To answer these questions, Coltheart et al. invoke the second factor. On their account, Capgras patients have Factor One (the neuropsychological disconnection), and adopt the delusional impostor belief as a result; but they also have Factor Two, which means they cannot subsequently revise this belief.

One serious limitation of this proposal is that it requires a precise chronology (see third column of Figure 1). The second factor cannot predate the first, nor can it be acquired at the same time as the first. If the second factor were to take effect prior to (or simultaneous with) the first, it would render the patient unable to update his belief system on the basis of the new evidence constituted by the first! In order for Coltheart et al's story to work, the individual in question would need to acquire Factor One first, and to adopt the delusional impostor belief. Then – *before* wife, friends, family and clinicans have any time to supply contrasting testimony – the gentleman would need to acquire Factor Two, rendering him unable to revise the delusional impostor belief in the light of subsequent testimony.[16]

[16] A further point is that having acquired Factor Two, the deluded patient should be rendered doxastically inert. If he is unable to update his belief system on the basis of new evidence, then he should not be able to revise his beliefs to accommodate, say, a change of government or a change in the weather. As far as I know there is no evidence for this kind of all-encompassing doxastic rigidity in deluded patients. One way to avoid this problem is to limit the scope of Factor Two, perhaps by suggesting that this factor is a domain-specific anomaly that is restricted to processing in a particular module or modules (see fn. 27).

4. An Alternative Account

I've identified three problems with the account of delusional inference put forward by Coltheart et al. (2010). In view of these difficulties, it seems doubtful that both Capgras patients and ventromedial frontal patients adopt the stranger belief via rational (approximately Bayesian) inference from highly unusual data, and that whereas ventromedial frontal patients subsequently update further – ending up with the non-delusional 'wife' belief – Capgras patients are unable to update further. I've argued that the stranger hypothesis is so implausible that any normal individual should give it minimal initial credence. Also, I've suggested that 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. Finally, I've shown that the story of Coltheart et al. requires a rather elaborate and implausible chronology. Here I outline an alternative account that avoids these problems.

4.1 Doxastic Conservatism vs. Explanatory Adequacy

Coltheart et al. (2010) state that 'In delusional subjects… the balance between the whole background belief system and the particular evidence base is not successfully achieved' (p. 276). This remark chimes with discussions in certain earlier papers. In particular, Stone and Young (1997; see also Davies & Coltheart, 2000; Aimola Davies & Davies, 2009) suggested that healthy belief formation involves a balance between two principles, one a principle of doxastic conservatism (whereby existing beliefs are maintained), the other a principle of observational adequacy (whereby the evidence of the senses is accommodated). Stone and Young suggested that delusions might be explained in terms of a bias towards observational adequacy.

In the account of Coltheart et al. (2010), however, the second factor in delusion formation looks like a bias towards doxastic conservatism (although they do not use this term). After all, individuals with their second factor are said to be 'impaired at revising preexisting beliefs on the basis of new evidence relevant to any particular belief' (p. 282). As we have seen, there are a number of problems with this account. I think that these problems largely disappear, however, if we pursue the original suggestion of Stone and Young, although the tendency towards observational adequacy might, following Aimola Davies and Davies (2009), be better characterised as a tendency towards *explanatory* adequacy.[17]

The fact that Coltheart et al. appear to conceive of the second factor as a bias towards doxastic conservatism is surprising, as they are quite aware of the explanatory adequacy of the various delusional hypotheses they consider. In discussing Cotard delusion (which involves the belief that one is dead), for example, they say:

[17] 'The imperative of observational adequacy corresponds to the prepotent doxastic response of treating a perceptual experience as veridical (seeing is believing). The imperative of explanatory adequacy corresponds to a prepotent doxastic tendency towards acceptance of a hypothesis that explains a salient piece of evidence and is thereby confirmed' (Aimola Davies & Davies, 2009, p. 293). Individuals with a bias towards explanatory adequacy are unduly influenced by the relative explanatory power of hypotheses.

> Suppose you suffered a form of brain damage that impaired the autonomic nervous system itself. Now you would not respond autonomically to any form of stimulation. What hypothesis might abductive inference yield that would be *explanatorily adequate* here? The hypothesis 'I am dead' is one such (because dead people are autonomically unresponsive). (Coltheart et al., 2010, p. 265; my emphasis.)

The hypothesis 'I am dead' would *explain* a lack of autonomic responses to one's environment – the latter data are precisely what one would expect if the hypothesis were true. Similarly, the hypothesis 'Other people can cause my arm to move' (endorsed by patients with delusions of alien control) would explain bodily movements that are not sensed as voluntary. And in Capgras delusion, of course, the stranger hypothesis, h_s, would explain why a face that is visually familiar fails to elicit an autonomic response.[18]

Bayes' theorem can be viewed as a prescription for navigating between excessive tendencies towards explanatory adequacy and doxastic conservatism. It is possible, therefore, to illustrate each of these tendencies as systematic deviations from the Bayesian prescription (see Appendix A for a mathematical model of these deviations):

$$P(h \mid o) = \frac{P(h)P(o \mid h)}{\sum_{h' \in H} P(h')P(o \mid h')} \quad (7)$$[19]

An individual with a bias towards explanatory adequacy will update beliefs *as if* ignoring the relevant prior probabilities of candidate hypotheses.[20] To the extent that he has this bias, his posterior for h, having observed data o, will approximate the ratio of the likelihood $P(o|h)$ to the sum of the respective likelihoods of the explanatory data given the $|H|$ mutually exclusive hypotheses under consideration:

[18] Coltheart et al's account of Capgras delusion appears to rely on both the principles of explanatory adequacy *and* doxastic conservatism: they appeal to the former to account for why the stranger hypothesis is initially adopted (although, notably, they do not suggest that there is any departure from Bayesian inference here), and they invoke the latter to explain why, in Capgras patients (but not ventromedial frontal patients), the stranger belief is maintained in the face of subsequent disconfirmatory evidence.

[19] This version of Bayes' theorem is derived from the version in (1) and (2) using the principle of *marginalization* (see Griffiths, Kemp, & Tenenbaum, 2008). H denotes the set of all hypotheses under consideration ($|H|$ denotes the cardinality of the set H, i.e., the number of hypotheses considered by the agent). This version of Bayes' rule makes it clear that $P(h|o)$ is directly proportional to the product of $P(h)$ and $P(o|h)$, relative to the sum of these same scores – products of priors and likelihoods – for all alternative hypotheses considered by the agent (Griffiths et al., 2008).

[20] I make no claims about the actual computational processes producing these biases. One possibility is that the relevant probabilities are mis*assessed* (in the case of a bias towards explanatory adequacy, for example, the individual might have a uniform prior distribution over the hypotheses in H). Another possibility is that the components of the Bayesian model are mis*aggregated*, i.e., not combined according to Bayes' rule (Fischhoff & Beyth-Marom, 1983; see also Hemsley & Garety, 1986).

$$P(h \mid o) \rightarrow \frac{P(o \mid h)}{\sum_{h' \in H} P(o \mid h')} \tag{8}$$

An individual with a bias towards doxastic conservatism, on the other hand, will update beliefs as if ignoring the relevant likelihood functions. To the extent that she has this bias, her posterior for *h*, having observed data *o*, will approximate her prior for *h*:

$$P(h \mid o) \rightarrow P(h) \tag{9}$$

4.2 Factor Two as a Bias Towards Explanatory Adequacy

Following Stone and Young (1997), I suggest that the second factor in delusion formation comprises a bias towards explanatory adequacy.[21,22] Capgras delusion, on this story, results when brain damage or disruption causes the face recognition system to become disconnected from the autonomic nervous system, generating the anomalous data *o* (Factor One). This disconnection occurs in conjunction with a bias towards explanatory adequacy (Factor Two), such that the individual updates beliefs as if ignoring the relevant prior probabilities of candidate hypotheses. He thus adopts as a belief the hypothesis that best explains the abnormal perceptual data available to him – the stranger hypothesis h_s (see third column of Figure 2).[23]

[21] As I have already noted, Stone and Young (1997) referred to observational adequacy rather than explanatory adequacy. It's worth adding that Stone and Young did not conceive of adequacy and conservatism in Bayesian terms. Indeed, whereas I have argued that Bayes' theorem is a formula for balancing these competing demands, Stone and Young suggested 'that there is no context-free balance to be advocated… no general context-free way to determine how [the tension between these principles] should be resolved' (1997, pp. 349-59).

[22] The notion that delusions involve a bias towards explanatory adequacy is broadly consistent with evidence that deluded individuals 'jump to conclusions' on probabilistic reasoning tasks (e.g., see Huq, Garety, & Hemsley, 1988; Fine, Gardner, Craigie, & Gold, 2007). In a seminal study, Huq et al. (1988) drew beads (ostensibly at random) from one of two hidden jars – a mostly red jar (containing 85 red beads and 15 yellow beads) or a mostly yellow jar (85 yellow, 15 red). Participants had to decide which jar the beads were being drawn from, and the experimenter continued to draw beads (to a maximum of 20) until this decision was reached. Huq et al. found that, relative to control participants, deluded participants required fewer draws before making a decision – and nearly half decided on the basis of just a single draw. One problem with this paradigm is that no information is given about the prior distribution over the two available options (e.g., a mostly red jar or a mostly yellow jar). In the absence of such information the most reasonable assumption is (arguably) that both options are equally likely (a uniform prior), and in that case the task cannot distinguish, at least on the basis of the initial draw, between an explanatory adequacy bias and Bayesian updating (in any case it is not at all clear how to infer beliefs from behavior in this task, as participants are typically given no financial incentive to persist). It would be interesting to compare the performance of deluded and control participants on incentivized variants of these tasks where the prior distribution over options is non-uniform (or where participants are informed as much).

[23] An explanatory adequacy bias can yield posteriors for h_s and h_w, given relatively realistic priors, that are equivalent to those in Coltheart et al's (2010) example of Bayesian inference with unrealistic priors (see fourth column of Table 2; see Appendix B for detailed workings).

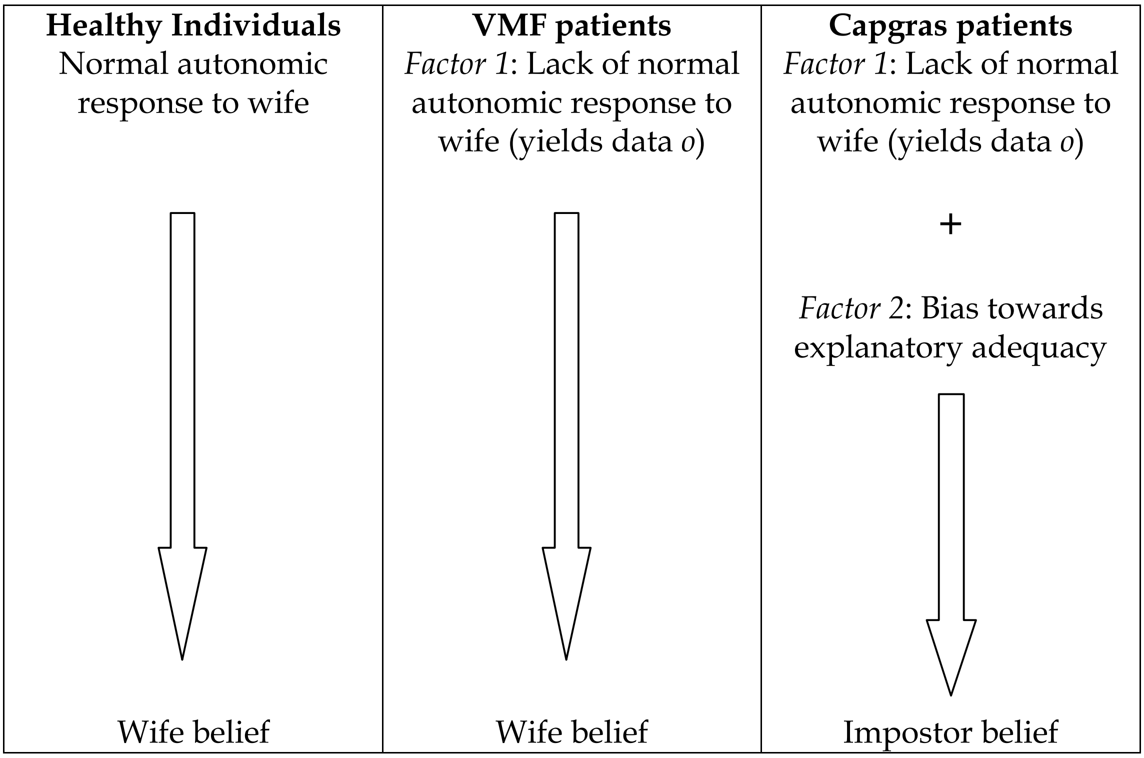
Healthy Individuals
Normal autonomic response to wife
Wife belief
VMF patients
Factor 1: Lack of normal autonomic response to wife (yields data o)
Wife belief
Capgras patients
Factor 1: Lack of normal autonomic response to wife (yields data o)
+
Factor 2: Bias towards explanatory adequacy
Impostor belief

Figure 2: Summary of an alternative account based on Stone and Young (1997) and Aimola Davies and Davies (2009)

This seems quite a parsimonious story – and it neatly solves the problems identified above. First, we no longer need to suggest that any (otherwise) normal individual with the Capgras first factor (the neuropsychological disconnection) will adopt the stranger hypothesis as a belief. The stranger hypothesis is so implausible as to make it virtually miraculous, and any reasonable individual will give it minimal credence. Individuals without Factor Two, therefore, will reject this hypothesis even in the face of data that strongly confirm it. Capgras patients, however, have Factor Two. They are biased towards explanatory adequacy, so are unduly influenced by the explanatory power of the stranger hypothesis and adopt it as a belief. Second, we no longer need to assume that ventromedial frontal patients ever adopt the stranger belief. These patients may well have the Capgras first factor, but they don't have the second, general factor (the bias towards explanatory adequacy) – so for these patients the general implausibility of the impostor scenario outweighs its relative explanatory power and the wife belief is retained. Finally, we can dispense with the elaborate chronology – although two distinct cognitive abnormalities are postulated, we need make no assumptions about the timing of these abnormalities (compare Figures 1 and 2). They may arise simultaneously, independent results of the same neuropathological process; or one may predate the other.

4.3 Prediction Errors

An important line of research into delusions involves the concept of *prediction error*. A prediction error is a discrepancy between what is expected and what actually comes to pass (Corlett, Honey et al., 2007). Such errors drive learning in organisms: they indicate a need to update beliefs about the world (Corlett et al., 2009; Fletcher & Frith, 2009). Although normally prediction error signals are minimised when beliefs best approximate reality, prediction error models of delusion formation implicate a disturbance in the error-dependent updating of beliefs: delusions are conceived as attempts to accommodate inappropriately generated prediction error signals (Corlett, Honey et al., 2007; Corlett et al., 2009; Fletcher & Frith, 2009; see Kapur, 2003).

Recent experiments have provided impressive support for such models. For example, Corlett, Murray et al. (2007) had participants learn associations between certain foods and allergic reactions, and used fMRI to explore the impact on brain activity of subsequent trials that either confirmed or violated the participants' engendered expectancies. Two groups of participants were tested, a group of patients with delusions and a group of healthy controls. Right prefrontal cortex (rPFC) responses[24] in the patient group were indicative of disrupted prediction error processing: whereas for control participants rPFC activation was relatively strong for expectancy violations and relatively weak for expectancy confirmations, the magnitude of rPFC response to expectancy violations in the patient group was not significantly different from that for expectancy confirmations. Further, the more severe the delusions in the patient group, the less likely that rPFC activation distinguished violation and confirmation of expectancy. These findings suggest that, in deluded individuals, 'the prefrontal cortex responds to physiological noise as if it were salient biological signal and drives the attribution of salience and attention to irrelevant and inconsequential environmental events' (Corlett, Murray et al., 2007, p. 2398).

How might the concept of prediction error fit with the two-factor model of delusions? The most obvious suggestion is that aberrant prediction error signaling comprises the second factor. In terms of the scheme I have outlined above, an excess of prediction error signal is what underpins the bias towards explanatory adequacy. Prediction error signals are triggered by discrepancies between the data expected and the data encountered. Such signals render salient the unexpected data and initiate a revision of beliefs to accommodate these data. If there is an excess of prediction error signal, inappropriately heightened salience is attached to the data (Kapur, 2003), and belief revision is excessively accommodatory – biased towards explanatory adequacy. Seen in this light, Factors One and Two play essentially the same role – both factors result in salient data. Factor One generates data that are at odds with prior beliefs. This discrepancy triggers prediction error signals that render the data salient. Factor Two involves the endogenous generation of prediction error signals, 'artificially' heightening the salience of given data. One might think of Factor

[24] This brain region has previously been identified as a reliable marker for prediction error processing (e.g., Corlett et al., 2004).

One as generating the prediction error signal, and of Factor Two as turning up the gain on this signal.

One implication of the above analysis is that, contrary to standard accounts of the two-factor theory, either factor might be sufficient in isolation to produce delusions. As I noted above, there are cases where salient data are presumably present but where a delusion is absent (again, see fourth column of Table 1). It may be that, in these cases, the salience of the data in question has simply not passed a certain threshold. It's worth noting in this connection that prediction error theorists challenge the strict conceptual distinction between perception and belief upon which the two-factor account of delusions is predicated (Fletcher & Frith, 2009; Mishara & Corlett, 2009). Such theorists see no need to implicate separate contributing factors at two levels, the experiential/perceptual and the inferential/doxastic. Rather, they implicate a single factor – aberrant prediction error signaling – that interferes with the updating of hypotheses across a broad perceptual-doxastic continuum.[25,26] Capgras delusion, for example, may result when aberrant prediction error signal results in inappropriate heightened salience being attached to a familiar other, engendering a perceived lack of familiarity with that person (Corlett, D'Souza, & Krystal, 2010).

I have suggested that the second factor in delusion formation involves the excessive production of prediction error signal, driving a bias towards explanatory adequacy in belief revision. But what if the production of prediction error signal were attenuated rather than heightened? In this case, insufficient salience would be attached to discrepancies between expected and encountered data, and the impetus to revise beliefs would be weakened or removed. This would be a bias towards doxastic conservatism. It's interesting to consider the types of psychopathology that might be associated with such a bias[27]. Potential candidates might include certain negative features of schizophrenia, for example catatonia. Fletcher and Frith (2009) have speculated that the same fundamental deficit (disrupted prediction error signalling) might account both for positive features such as delusions and negative features such as catatonia. My speculation here is more specific – that whereas delusions involve an excess of prediction error signalling, negative features such as

[25] Such models invoke a hierarchical Bayesian arrangement whereby the posteriors computed by lower levels of the system serve as priors for higher levels (and vice versa; see Fletcher & Frith, 2009).

[26] As a two-factor theorist, my view is that the distinction between perception and belief is not easily dispensed with - particularly at the personal level of description. As Martin Davies (personal communication) notes, 'One indication of this is that obliterating the distinction between perception and belief seems to have the consequence that illusory perceptual states that are informationally encapsulated from most of a person's beliefs are liable to be counted as delusions.'

[27] It seems likely that certain domain-specific biases towards doxastic conservatism represent biological adaptations. For example, according to Error Management Theory (Haselton & Buss, 2000; McKay & Efferson, 2010), biased beliefs in males about female sexual interest may have been adaptive in the ancestral past and may persist in modern males – a prediction that has garnered empirical support (Abbey, 1982; Haselton, 2003; Haselton & Buss, 2000). How might natural selection implement such 'adaptive misbeliefs' (McKay & Dennett, 2009)? Perhaps via hard-wired, domain-specific priors, coupled with domain-specific biases towards doxastic conservatism that confer a resistance to disconfirmatory evidence. It would be interesting to investigate whether prediction error signal is attenuated when certain domain-specific expectancies are violated.

catatonia involve a decrement in such signalling (see Hohwy, 2004, for a similar idea).

5. Summary and Conclusion

To what extent does delusion formation involve abnormal inference? A number of distinct positions can be distinguished on this issue. First, there is the single-factor approach of Maher (e.g., 1992; 1999). Maher's position is that delusion formation does not occur because inferential reasoning is abnormal, but rather because normal inferential reasoning is brought to bear on abnormal experience. Then there is the two-factor approach of Coltheart and colleagues (e.g., Coltheart, 2005; 2007). Like Maher, Coltheart and colleagues distinguish between experience and belief (or at least between *data* and belief; see fn. 3). Unlike Maher, however, they implicate abnormalities at both levels: two abnormalities with distinct neuropsychological underpinnings. Finally, there is the approach of prediction error theorists (e.g., Fletcher & Frith, 2009; Mishara & Corlett, 2009). Like Maher, such theorists argue that delusion formation can be explained with a single factor. Unlike Maher, however, these theorists disavow any strict conceptual separation between experience and belief. There is just one basic abnormality - aberrant prediction error signaling - that disrupts inference across the board.

The recent Bayesian account of delusional inference put forward by Coltheart et al. (2010) is an impressive attempt to flesh out the two-factor theory of delusion formation, and in particular to say something specific about the nature of the second factor. I have outlined some criticisms of this account, and have sketched an alternative account of the second factor, based on the important contributions of Stone and Young (1997) and Aimola Davies and Davies (2009). In particular, I have argued that the second factor in delusion formation involves a systematic deviation from Bayesian updating, a deviation that may be characterized as a bias towards explanatory adequacy. I have connected this idea with prominent work on prediction error signalling, and have suggested that the bias in my model (represented by the variable x; see Appendix A) is driven by aberrant prediction error signal. At the default value ($x = 0$), prediction error signalling is normal, and individuals update beliefs in an approximately Bayesian fashion. Departures from this default correspond either to a dysfunctional surge in prediction error signalling ($x \rightarrow 1$), underpinning a bias towards explanatory adequacy (and perhaps leading to bizarre delusions, as per my conception of the second factor); or to a pathological attenuation of such signalling ($x \rightarrow -1$), underpinning a conservative bias (and perhaps contributing to negative features such as catatonia).

References

Abbey, A. (1982). Sex differences in attributions for friendly behavior: Do males misperceive females' friendliness? *Journal of Personality and Social Psychology, 42*, 830-838.

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

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

Bayes, T. R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370-418.

Bayne, T., & Pacherie, E. (2005). In defence of the doxastic conception of delusions. *Mind and Language, 20*(2), 163-188.

Bortolotti, L. 2009: *Delusions and Other Irrational Beliefs*. Oxford: Oxford University Press.

Brighetti, G., Bonifacci, P., Borlimi, R., & Ottaviani, C. (2007). "Far from the heart far from the eye": Evidence from the Capgras delusion. *Cognitive Neuropsychiatry, 12*(3), 189-197.

Christensen, D. (2004). *Putting Logic in Its Place: Formal Constraints on Rational Belief.* Oxford: Oxford University Press.

Coltheart, M. (2005). Delusional belief. *Australian Journal of Psychology, 57*(2), 72-76.

Coltheart, M. (2007). The 33rd Sir Frederick Bartlett Lecture: Cognitive neuropsychiatry and delusional belief. *The Quarterly Journal of Experimental Psychology, 60*(8), 1041-1062.

Coltheart, M. (2009). Delusions and misbeliefs. *Behavioral and Brain Sciences, 32*, 517.

Coltheart, M., Langdon, R., & McKay, R. (2007). Schizophrenia and monothematic delusions. *Schizophrenia Bulletin, 33*(3), 642-647.

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

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

Corlett, P. R., Aitken, M. R., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. A., & al., e. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron, 44*, 877-888.

Corlett, P. R., D'Souza, D. C., & Krystal, J. H. (2010). Capgras syndrome induced by ketamine in a healthy subject. *Biological Psychiatry.*

Corlett, P. R., Honey, G. D., & Fletcher, P. C. (2007). From prediction error to psychosis: Ketamine as a pharmacological model of delusions. *Journal of Psychopharmacology, 21*(3), 238-252.

Corlett, P. R., Krystal, J. H., Taylor, J. R., & Fletcher, P. C. (2009). Why do delusions persist? . *Frontiers in Human Neuroscience, 3.*

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

Davies, M., & Coltheart, M. (2000). Introduction: Pathologies of belief. In M. Coltheart & M. Davies (Eds.), *Pathologies of belief.* (pp. 1-46). Malden, MA, US: Blackwell Publishers.

Davies, M., Coltheart, M., Langdon, R., & Breen, N. (2001). Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, & Psychology, Vol 8*(2-3), 133-158.

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

Ellis, H. D., Young, A. W., Quayle, A. H., & de Pauw, K. W. (1997). Reduced autonomic responses to faces in Capgras delusion. *Proceeding of the Royal Society of London: Biological Sciences, B264*, 1085-1092.

Fine, C., Gardner, M., Craigie, J., & Gold, I. (2007). Hopping, skipping or jumping to conclusions? Clarifying the role of the JTC bias in delusions. *Cognitive Neuropsychiatry, 12*(1), 46-77.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90*(3), 239-260.

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

Grann, D. (2008). The Chameleon: The many lives of Frédéric Bourdin. *The New Yorker, AUGUST 11, 2008*.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational cognitive modeling*: Cambridge University Press.

Haselton, M. G. (2003). The sexual overperception bias: Evidence of a systematic bias in men from a survey of naturally occurring events. *Journal of Research in Personality, 37*(1), 34-47.

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology, 78*(1), 81-91.

Hemsley, D. R., & Garety, P. A. (1986). The formation of maintenance of delusions: A Bayesian analysis. *British Journal of Psychiatry, 149*, 51-56.

Hirstein, W. S., & Ramachandran, V. S. (1997). Capgras syndrome: A novel probe for understanding the neural representation of the identity and familiarity of persons. *Proceedings of the Royal Society B: Biological Sciences, 264*, 437-444.

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

Hohwy, J. (2010). The hypothesis testing brain: Some philosophical applications. In W. Christensen & E. Schier & J. Sutton (Eds.), *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science* (pp. 135-144). Sydney: Macquarie Centre for Cognitive Science.

Huq, S. F., Garety, P. A., & Hemsley, D. R. (1988). Probabilistic judgements in deluded and non-deluded subjects. *Quarterly Journal of Experimental Psychology A, 40*(4), 801-812.

Jaspers, K. (1946/1963). *General psychopathology* (J. Hoenig & M. W. Hamilton, Trans. 7th ed.): The Johns Hopkins University Press.

Kapur, S. (2003). Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *American Journal of Psychiatry, 160*(1), 13-23.

Langdon, R., & Coltheart, M. (2000). The cognitive neuropsychology of delusions. *Mind & Language, 15*(1), 183-216.

Lipton, P. (2004). *Inference to the best explanation* (Second edition ed.). London: Routledge.

Maher, B. A. (1992). Delusions: Contemporary etiological hypotheses. *Psychiatric Annals, 22*, 260-268.

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

Maher, B. A., & Ross, J. S. (1984). Delusions. In H. E. Adams & P. B. Sutker (Eds.), *Comprehensive handbook of psychopathology* (pp. 383-409). New York: Plenum Press.

McKay, R., & Dennett, D. (2009). The evolution of misbelief (Target article); Our evolving beliefs about evolved misbelief (Response to commentaries). *Behavioral and Brain Sciences, 32*(6), 493-561.

McKay, R., & Efferson, C. (2010). The subtleties of error management. *Evolution & Human Behavior, 31*, 309-319.

McKay, R., & Kinsbourne, M. (2010). Confabulation, delusion, and anosognosia: Motivational factors and false claims. *Cognitive Neuropsychiatry, 15*(1), 288-318.

Mishara, A. L., & Corlett, P. (2009). Are delusions biologically adaptive? Salvaging the doxastic shear pin. *Behavioral and Brain Sciences, 32*(6), 530-531.

Spitzer, M. (1990). On defining delusions. *Comprehensive Psychiatry, 31*(5), 377-397.

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

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

White, P. A. (2009). Property transmission: An explanatory account of the role of similarity information in causal inference. *Psychological Bulletin, 135*(5), 774-793.

Young, G. (2011). On abductive inference and delusional belief: Why there is still a role for patient experience within explanations of Capgras delusion. *Cognitive Neuropsychiatry, 16*(4), 303-325.

Appendix A: A Model of Biased Inference

Let us use the variable *x* to model biases towards explanatory adequacy and doxastic conservatism. Let *x* be a value in the interval (-1, +1). Now modify Bayes' theorem as follows, incorporating the two piecewise-defined functions α(*x*) and β(*x*):

$$P(h \mid o) = \frac{\left[P(h) - \alpha \times \left[P(h) - \tfrac{1}{|H|}\right]\right] \times \left[P(o \mid h) - \beta \times \left[P(o \mid h) - \tfrac{1}{|H|}\right]\right]}{\sum_{h' \in H} \left[P(h') - \alpha \times \left[P(h') - \tfrac{1}{|H|}\right]\right] \times \left[P(o \mid h') - \beta \times \left[P(o \mid h') - \tfrac{1}{|H|}\right]\right]}$$

$$\alpha(x) = \left\{\begin{matrix} x, x > 0 \\ 0, x \leq 0 \end{matrix}\right\} \qquad \beta(x) = \left\{\begin{matrix} -x, x < 0 \\ 0, x \geq 0 \end{matrix}\right\} \tag{10}$$

Let us assume that the default value of *x* is 0 - this state reflects unbiased Bayesian inference. When *x* is 0, both α and β are 0, and the rather formidable theorem above is simply Bayes' theorem:

$$P(h \mid o) = \frac{P(h)P(o \mid h)}{\sum_{h' \in H} P(h')P(o \mid h')} \tag{11}$$

As *x* increases, however, the posterior $P(h \mid o)$ will increasingly approximate the ratio of the likelihood $P(o \mid h)$ to the sum of all relevant likelihoods, yielding an explanatory adequacy bias (see worked example in Appendix B); conversely, decreasing *x* will cause the posterior $P(h \mid o)$ to increasingly approximate the prior *P*(*h*), reflecting a doxastic conservatism bias.

Appendix B: Worked Example of a Bias Towards Explanatory Adequacy

Here we implement a bias towards explanatory adequacy (setting x = 0.02) in a situation where there are two hypotheses under consideration, h_s and h_w, i.e., $H = \{h_s, h_w\}$ and $|H| = 2$. Relatively realistic priors are assigned for the two hypotheses and Coltheart et al's figures for the two likelihoods are retained: $P(h_s) = 0.00027$, $P(h_w) = 0.99973$, $P(o|h_s) = 0.999$ and $P(o|h_w) = 0.001$ (see section 3.1, and fourth column of Table 2).

The model of biased inference (see Appendix A) is as follows:

$$P(h|o)=\frac{\left[P(h)-\alpha\times\left[P(h)-\tfrac{1}{|H|}\right]\right]\times\left[P(o|h)-\beta\times\left[P(o|h)-\tfrac{1}{|H|}\right]\right]}{\sum_{h'\in H}\left[P(h')-\alpha\times\left[P(h')-\tfrac{1}{|H|}\right]\right]\times\left[P(o|h')-\beta\times\left[P(o|h')-\tfrac{1}{|H|}\right]\right]}$$

$$\alpha(x)=\begin{Bmatrix}x,x>0\\0,x\le 0\end{Bmatrix}\qquad \beta(x)=\begin{Bmatrix}-x,x<0\\0,x\ge 0\end{Bmatrix} \tag{12}$$

Substituting x = 0.02 into the piecewise-defined functions α(x) and β(x) gives the following values of α and β: α = 0.02 and β = 0. Substituting these values into the equation for $P(h|o)$ gives:

$$P(h|o)=\frac{\left[P(h)-0.02\times\left[P(h)-\tfrac{1}{|H|}\right]\right]\times\left[P(o|h)\right]}{\sum_{h'\in H}\left[P(h')-0.02\times\left[P(h')-\tfrac{1}{|H|}\right]\right]\times\left[P(o|h')\right]} \tag{13}$$

The equation for the posterior probability of the stranger hypothesis is thus:

$$P(h_s|o)=\frac{\left[P(h_s)-0.02\times\left[P(h_s)-\tfrac{1}{|H|}\right]\right]\times\left[P(o|h_s)\right]}{\left[\left[P(h_s)-0.02\times\left[P(h_s)-\tfrac{1}{|H|}\right]\right]\times\left[P(o|h_s)\right]\right]+\left[\left[P(h_w)-0.02\times\left[P(h_w)-\tfrac{1}{|H|}\right]\right]\times\left[P(o|h_w)\right]\right]} \tag{14}$$

Likewise, the equation for the posterior probability of the wife hypothesis is:

$$P(h_w|o)=\frac{\left[P(h_w)-0.02\times\left[P(h_w)-\tfrac{1}{|H|}\right]\right]\times\left[P(o|h_w)\right]}{\left[\left[P(h_s)-0.02\times\left[P(h_s)-\tfrac{1}{|H|}\right]\right]\times\left[P(o|h_s)\right]\right]+\left[\left[P(h_w)-0.02\times\left[P(h_w)-\tfrac{1}{|H|}\right]\right]\times\left[P(o|h_w)\right]\right]} \tag{15}$$

Substituting the above values for $|H|$, $P(h_s)$, $P(h_w)$, $P(o|h_s)$ and $P(o|h_w)$ gives the following solution for $P(h_s|o)$:

$$P(h_s|o) = \frac{\left[0.00027 - 0.02 \times \left[0.00027 - 0.5\right]\right] \times \left[0.999\right]}{\left[\left[0.00027 - 0.02 \times \left[0.00027 - 0.5\right]\right] \times \left[0.999\right]\right] + \left[\left[0.99973 - 0.02 \times \left[0.99973 - 0.5\right]\right] \times \left[0.001\right]\right]} \quad (16)$$

$$\therefore P(h_s|o) \approx 0.91 \quad (17)$$

…and this solution for $P(h_w|o)$:

$$P(h_w|o) = \frac{\left[0.99973 - 0.02 \times \left[0.99973 - 0.5\right]\right] \times \left[0.001\right]}{\left[\left[0.00027 - 0.02 \times \left[0.00027 - 0.5\right]\right] \times \left[0.999\right]\right] + \left[\left[0.99973 - 0.02 \times \left[0.99973 - 0.5\right]\right] \times \left[0.001\right]\right]} \quad (18)$$

$$\therefore P(h_w|o) \approx 0.09 \quad (19)$$

Thus with x = 0.02 and relatively realistic priors, the posterior probability of the stranger hypothesis ($P(h_s|o) \approx 0.91$) is more than ten times greater than the posterior probability of the wife hypothesis ($P(h_w|o) \approx 0.09$), and the individual will adopt the stranger hypothesis as a belief.