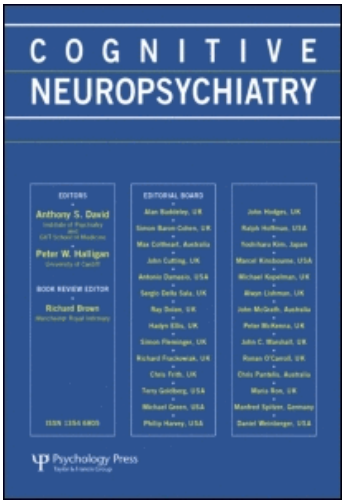

Cognitive Neuropsychiatry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713659088

Abductive inference and delusional belief

Max Coltheart [a]; Peter Menzies [b]; John Sutton [ab]
[a] Macquarie Centre for Cognitive Science, [b] Department of Philosophy, Macquarie University, Sydney, Australia

First published on: 15 December 2009

To cite this Article Coltheart, Max, Menzies, Peter and Sutton, John(2010) 'Abductive inference and delusional belief', Cognitive Neuropsychiatry, 15: 1, 261 — 287, First published on: 15 December 2009 (iFirst)
To link to this Article: DOI: 10.1080/13546800903439120
URL: http://dx.doi.org/10.1080/13546800903439120

COGNITIVE NEUROPSYCHIATRY
2010, 15 (1/2/3), 261–287

Psychology Press
Taylor & Francis Group

Abductive inference and delusional belief

Max Coltheart[1], Peter Menzies[2], and John Sutton[1,2]

[1]*Macquarie Centre for Cognitive Science,* [2]*Department of Philosophy, Macquarie University, Sydney, Australia*

Delusional beliefs have sometimes been considered as rational inferences from abnormal experiences. We explore this idea in more detail, making the following points. First, the abnormalities of cognition that initially prompt the entertaining of a delusional belief are not always conscious and since we prefer to restrict the term "experience" to consciousness we refer to "abnormal data" rather than "abnormal experience". Second, we argue that in relation to many delusions (we consider seven) one can clearly identify what the abnormal cognitive data are which prompted the delusion and what the neuropsychological impairment is which is responsible for the occurrence of these data; but one can equally clearly point to cases where this impairment is present but delusion is not. So the impairment is not sufficient for delusion to occur: a second cognitive impairment, one that affects the ability to evaluate beliefs, must also be present. Third (and this is the main thrust of our paper), we consider in detail what the nature of the inference is that leads from the abnormal data to the belief. This is not deductive inference and it is not inference by enumerative induction; it is abductive inference. We offer a Bayesian account of abductive inference and apply it to the explanation of delusional belief.

Keywords: Delusion; Abductive inference; Cognitive neuropsychiatry; Bayesian inference; Experience.

INTRODUCTION

William James—that uncannily prescient forefather of cognitive science—had this to say about delusional beliefs: "The delusions of the insane are apt to affect certain typical forms, very difficult to explain. But in many cases they are certainly theories which the patients invent to account for their bodily sensations" (1890/1950, chap. XIX). We think this is correct. So did one of the most influential recent thinkers about delusions, Brendan Maher: "A delusion

Correspondence should be addressed to Max Coltheart, Macquarie Centre for Cognitive Science, Macquarie University, Sydney NSW 2019, Australia. E-mail: max@maccs.mq.edu.au

We thank Ryan McKay and an anonymous reviewer for very helpful critiques of earlier drafts of this paper, and Robyn Langdon for her editorial work.

http://www.psypress.com/cogneuropsychiatry DOI: 10.1080/13546800903439120

is a hypothesis designed to explain unusual perceptual phenomena'' (1974, p. 103).

Our paper is about how deluded people go about generating the ''theories'' (James) or ''hypotheses'' (Maher) by which they can explain the unusual ''bodily sensations'' (James) or ''perceptual phenomena'' (Maher) with which they are confronted.

James was also correct in his view that delusions come in certain typical forms, and these forms have begun to be well-characterised since the birth, a century after the publication of his *Principles* (1890/1950), of the discipline of cognitive neuropsychiatry. The forms of delusion include the Cotard delusion (the belief that you are dead), the Frégoli delusion (the belief that you are constantly being followed around by people you know, whom you can't recognise because they are in disguise), somatoparaphrenia (the belief that some part of your body—your arm, for example—belongs to someone else rather than you), mirrored self-misidentification (the belief that the person you see when you look in the mirror is not you, but some stranger who happens to look like you), and others. What we have to say in this paper about how patients generate hypotheses that give rise to delusional beliefs is intended to apply to all of these forms of delusion. But we will mainly expound these ideas in relation to just one form of delusion, the one that has been most extensively studied.

THE CAPGRAS DELUSION

This is the belief that someone who is emotionally close to you—often a spouse, for example, let's say your wife—has been replaced by someone else, someone whom you consider a stranger even while acknowledging that this person does look very like your wife.

What could possibly suggest to someone the idea that the woman he is looking at (who is in fact his wife) is not his wife but some complete stranger? We believe the answer to this question is: a neuropsychological impairment that has disconnected the face recognition system (itself intact) from the autonomic nervous system (itself intact).

The evidence for this view is now very strong. In healthy control subjects, and in nondelusional psychiatric subjects, familiar faces evoke much larger responses of the autonomic nervous system, indexed by the skin conductance response (SCR) evoked by the presentation of a face, than do unfamiliar faces. But four different studies of Capgras patients (Brighetti, Bonifacci, Borlimi, & Ottaviani, 2007; Ellis, Lewis, Moselhy, & Young, 2000; Ellis, Young, Quayle, & de Pauw, 1997; Hirstein & Ramachandran, 1997) have shown that this differential autonomic responsivity to familiar faces compared to unknown faces is absent in people with Capgras delusion. In Capgras sufferers, the

autonomic responses to faces are small, and are no greater for familiar than for unfamiliar faces. Hence, for the Capgras patient the autonomic response to his wife's face is just as it would be if the face he is looking at *were* the face of a stranger; and the patient does indeed say—and, we believe, *believe*—that this is the face of a stranger.

Imagine that you suffered a sudden stroke whose only effect on your brain was to disconnect your face recognition system (itself intact) from your autonomic nervous system (itself intact). Consider what will happen the next time you encounter your wife, in some context where you are expecting to see her (in the kitchen of your house, for example) and see her face. When you do encounter her, you will expect the usual consequence—a response by your autonomic nervous system. But this does not happen. Why not?—and see her face. How are you to explain this violation of expectation?

James, and Maher, would say that the task confronting the deluded person here is to formulate a hypothesis, which, if true, would account for the failure of an autonomic response to occur. Maher was very explicit concerning how this happens. His view was that the types of inferences from observations to the explanations of those observations that are performed by the deluded patient in the genesis of the delusion are no different from the types of inferences routinely made by the folk, and indeed those made by the scientist ("the 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"; 1974, p. 103).

Proponents of the view that delusions arise via a normal inferential process applied to abnormal data, a view originating with James and Maher and currently being developed further (see, e.g., Coltheart, 2007; Coltheart, Langdon, & McKay, 2007; Langdon & Coltheart, 2000) have not considered what particular form of inference is involved; but it is clear that what's involved is not deductive or (enumerative) inductive inference—it is a third type of inference, one known as *abductive* inference. So this view of delusional belief might benefit from a consideration of the philosophical conception of abductive inference. The main aim of our paper is to offer such a consideration. Before pursuing that theme, however, we need to elaborate upon what might be meant by the term "abnormal data" used at the beginning of this paragraph, and the corresponding terms used by James ("bodily sensations") and Maher ("unusual perceptual phenomena"). What's particularly important here is the use of the word "experience" in this context.

DELUSIONS AND EXPERIENCES

The James-Maher idea about delusional belief is often expressed by proposing that delusions arise via the application of normal reasoning

processes to abnormal experiences. Maher himself often expressed matters in this way, e.g., "these experiences demand explanation which the patient develops through the same cognitive mechanisms that are found in the normal and scientific theory-building" (1974, p. 111).

The problem we see here is that it is common to treat experiences as by definition conscious, so that nothing of which a person is not conscious can be called an "experience", and the term "conscious experience" is a tautology. If that is how the term "experience" is being used when one proposes that delusions are rational responses to abnormal experiences, then that theory of delusion is false for many kinds of delusions—the Capgras delusion, for example. The reason is that 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. Hence, what happens here is not an abnormal experience, because it is not an experience. It is an abnormality; and the application of processes of abductive inference so as to generate a hypothesis to explain this abnormality is, we consider, the source of the delusion; but we don't want to say that it is an abnormality of experience, because if we do then either (a) we must allow that experiences can be unconscious (which we don't want to do) or else (b) we must conclude that the James-Maher explanation of delusional belief fails to account for the Capgras delusion and kindred delusions (which we also don't want to do).

Of course, there is no doubt that people with Capgras delusion *do* have abnormal (conscious) experiences: Having a stranger living in your house without any explanation is an abnormal experience. So is having some stranger pretending to be your wife. But these experiences are not the cause of the delusion; they are *consequences* of the delusion.

Hence, we want to take the view that the abnormality in Capgras delusion, which prompts the exercise of abductive inference in an effort to generate a hypothesis to explain this abnormality, is not an abnormality of which the patient is *aware*. Nor is the patient aware of the application of abductive inference that occurs here. In the sudden-stroke example we gave earlier, we argue that the first delusion-relevant event of which the patient is aware is the belief "That isn't my wife". Everything that preceded the occurrence of that belief and was responsible for the belief having come about—the stroke, the neuropsychological disconnection, the absence of an autonomic response when the wife is next seen, the invocation of a process of abductive inference to yield some hypothesis to explain this, and the successful generation of such a hypothesis—all of these processes are unconscious. What's conscious is only the *outcome* that this chain of processes generated: the conscious belief "This person isn't my wife."

We consider that this kind of analysis holds true for many forms of delusion, not just for Capgras delusion. That is, in many forms of delusion,

something abnormal occurs of which a person is not conscious, unconscious processes of abductive inference are invoked to seek a hypothesis which, if true, would explain that abnormality, and a hypothesis is found which is judged satisfactory by these unconscious inferential processes. After all of that unconscious processing has been completed, the hypothesis is accepted as a (delusional) belief, and enters consciousness.

Consider, for example, the Frégoli delusion—the belief that people whom you know are constantly following you around, and that the reason you don't recognise which of your acquaintances are doing this is that they wear disguises. One speculative explanation of the Frégoli delusion (Ramachandran & Blakeslee, 1998) is that it is due to neuropsychological damage that has caused abnormal hyperresponsivity of the face recognition system. That would cause the faces of unfamiliar people to activate representations of known faces in the face recognition system, and hence to generate the degree of autonomic response normally occurring only when a familiar face has been seen. This in turn generates the hypothesis that strangers you encounter in the street are often people with whom you are familiar, in disguise. All of this cognitive processing would be unconscious; the hypothesis thus formed unconsciously would subsequently arise in consciousness as a (delusional) belief.

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). The delusional belief "I am dead" is known as Cotard delusion.

These specific ideas about the genesis of delusional beliefs are consistent with the general theory of unconscious processing developed in considerable detail by Mandler (1975). For example, "The analysis of situations and appraisal of the environment ... goes on mainly at the unconscious level" (p. 241); "There are many systems that cannot be brought into consciousness, and probably most systems that analyse the environment in the first place have that characteristic. In most cases, only the products of cognitive and mental activities are available to consciousness" (p. 245); "unconscious processes ... include those that are not available to conscious experience [such as] affective appraisals" (p. 230).

We do not mean to suggest that in *every* type of delusion the abnormality whose detection triggers processes of abductive inference is always an abnormality of which the deluded person is unconscious. Delusions of reference and delusions of persecution may be examples of forms of delusion in which the initial abnormality, which triggers an abductive-inference process, is a conscious experience. Maher (1974, p. 110) gives some persuasive examples of delusions of reference which, from the patients' own accounts,

seem to have been triggered by abnormal perceptual experiences such as a recurrent crackling sound in the wall, or the light coming through a window seeming to be abnormally bright. Our point is that we wish to make it clear that, when we are discussing the role of abductive inference in the genesis of delusional belief, we will be allowing that the entire chain of processes, from the initial abnormality up to the identification via abductive inference of a hypothesis to explain why that abnormality has occurred, may all be unconscious. That is the reason why we will not be using the term "abnormal experience" to refer to the initial event that triggers the chain of processes that culminates in the formation of a delusional belief. Instead, we will use the neutral term "abnormal data".

ABNORMAL DATA WITHOUT DELUSIONAL BELIEF

It seems difficult to deny that the failure of autonomic response to familiar faces that occurs in Capgras delusion plays a causal role in the genesis of the delusion. Surely it can't be the case that it is just an accident that Capgras patients exhibit this failure? We assume it is not, and so we take the view that this failure is necessary for the Capgras delusion to occur. But is it *sufficient* to cause the delusion? The answer seems to be "No". Patients with damage to ventromedial regions of the frontal lobe also fail to show greater autonomic responsivity to familiar than to unfamiliar faces (Tranel, Damasio, & Damasio, 1995); but these patients still correctly identify spouses and other family members (D. Tranel, personal communication, 6 May 2008)—they are not delusional. Why not?

This conundrum is not confined to Capgras delusion; it arises in connection with a number of other delusions. Just as is the case with Capgras delusion, in other forms of delusion one can identify (or in some cases hypothesise) forms of abnormal data to which the patient is being exposed and which by their very nature would seem highly likely to be causally responsible for the content of the delusional belief. But for all of these other delusions, patients exist in whom the relevant form of abnormal data is present and yet there is no delusional belief, implying that even if the occurrence of the abnormal data is necessary for the delusion to arise, it is not sufficient.

Table 1 lists examples. The third column lists the forms of delusional belief being considered here. The first column lists the neuropsychological impairment documented as being present when the delusion is present (except for Frégoli and Cotard delusions, where the impairment has been presented as hypothesised by Ramachandran & Blakeslee, 1998; no empirical studies of autonomic responsivity in these two delusions have been reported). The second column describes the form of abnormal data that would be present as a consequence of the neuropsychological impairment.

TABLE 1
Neuropsychological impairments with and without delusion

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" (Fregoli 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)

Table 1 (*Continued*)

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*
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

We wish to make two points about the contents of this table. The first point is that for all seven delusions the relationship between the specific form of the neuropsychologically generated abnormal data (Column 2) and the specific content of the delusional belief (Column 3) is such as to make highly plausible the view that the nature of the abnormal data is causally responsible for the content of the belief. So our claim is that each delusion results from applying a process of abductive inference to the abnormal data in an effort to explain why these data are occurring.

The second point is that for all seven delusions the neuropsychological impairment (Column 1) that we consider causally implicated in the delusion can nevertheless also be present in people who are not delusional (Column 4).

Thus, we have two things to explain. The first is the nature of the inferential processes which, when applied to the task of explaining the abnormal data, yield the delusional belief. The second is why it is the case that there are patients in whom the same forms of abnormal data are present but who are nevertheless not delusional.

So we turn first to a consideration of the abductive inference process itself, and exactly what role it might play in the generation of delusional beliefs.

ABDUCTIVE INFERENCE

Abductive inference, or inference to the best explanation, is the kind of inference that is used in selecting a hypothesis from a range of hypotheses on the basis of observations. Its governing idea is that explanatory considerations are a guide to inference: People infer from observations to the available hypothesis that best explains those observations. Many inferences in science and in everyday life are naturally characterised in this way. For example, when Newton inferred the truth of his theory of mechanics, he did so on the basis of the fact that it provided the best explanation of a vast range of terrestrial and celestial phenomena. When Sherlock Holmes inferred that it was Moriarty who committed the crime, he did so because this hypothesis best explained the fingerprints, the bloodstains, and other forensic evidence. Notwithstanding Holmes' claims to the contrary, this was not a matter of deduction. The evidence did not logically imply that Moriarty was the culprit, since it remained a logical possibility, formally consistent with the evidence, that someone else committed the crime. Nor was this a matter of enumerative induction that involves reasoning from observed instances of some phenomenon ("These observed swans are white") to the next instance of the phenomenon ("The next swan will be white"), or a generalisation about the phenomenon ("All swans are white"). Holmes did not infer Moriarty's guilt by inferring from Moriarty's commission of many crimes in

the past to the conclusion that he has committed this crime, for this may have been the first crime that Moriarty ever committed. In any case, the evidence took the form of observations that were specific to this particular crime. Indeed, what is typical of abductive inference, unlike enumerative inductive inference, is that the conclusion introduces new concepts that are not present in the reports of the observational evidence. For example, Holmes arrived at a conclusion ("Moriarty is the culprit") that employed concepts not present on the reports of the observational data ("The murderer left fingerprints on the knife; and has a bloodstained shirt.")

As just remarked, explanation plays a key role in abductive inference. There are two standard accounts of abductive inference in the literature, drawing on different understandings of explanation. (In the philosophy of science literature, these accounts are sometimes called accounts of hypothesis or theory confirmation since they are accounts of when a hypothesis is confirmed or disconfirmed by observational evidence.) One account is the logical empiricist account and the other is the Bayesian account. Although differing in the way they understand explanation, the accounts can be understood as normative accounts of how subjects *should* reason and as descriptive accounts of how subjects *actually* reason.

We shall sketch both these accounts before turning to consider how deluded subjects might be seen as employing abductive inference to form their delusional beliefs.

The logical empiricist account of abductive inference turns on its distinctive understanding of explanation as logical implication. On this account, a hypothesis or theory H explains some observation O just in case H logically implies O. (This is consistent with the earlier remark that abductive inference is not a matter of deduction, for there we were saying that the inference of H from O is not a matter of deductively inferring H from O.) The logical empiricists concentrated on abductive inference or theory confirmation within an account of scientific theories that construed them as formal axiomatic structures. They maintained that while the scientific theories actually used by scientists might not have a formal axiomatic structure, all proper scientific theories could be reconstructed in terms of a characteristic set of axioms and rules of inference. So, within this account, a theory explains some observed phenomenon when a sentence describing the phenomenon can be deduced from the axioms of the regimented theory. For example, Newton's classical mechanics explains the orbit of a planet around the sun by virtue of the fact that the laws of the theory (in conjunction with sentences describing the initial conditions of masses and distances) logically imply a sentence describing the orbit of the planet around the sun. The first part of abductive inference or theory confirmation on the logical empiricist model consists in checking to see whether sentences describing observed phenomena can be deduced from the

theory. However, this is not the whole story; there may be several hypotheses or theories that logically imply the same set of observation sentences. So the second part of abductive inference or theory confirmation consists in comparing the competing theories that explain some observational phenomena to see which of these theories provides the best explanation of the phenomena. Unfortunately, there was not a lot of consensus among the logical empiricists about which explanatory virtues make a theory the best explanation of some phenomena. It was said, for example, that of two rival hypotheses one is a better explanation than the other if it is simpler, or less ad hoc, or more coherent with a wider body of theory. It turned out to be surprisingly difficult to explicate these crucial explanatory virtues without introducing some subjective elements into theory choice.

This account of abductive inference or theory confirmation makes sense in the context of scientific theories that have a highly articulated formal structure enabling hypotheses to logically imply observational sentences. However, the theories or hypotheses that we are interested in do not have this kind of structure; and indeed they are not the kind of theories or hypotheses that logically imply observation sentences. For example, we want to be able to say that the hypothesis that Moriarty committed the crime explains the observed evidence to do with fingerprints, blood stains, etc., even though this hypothesis does not, strictly speaking, logically imply the evidence. Similarly, we want to be able to say that a deluded subject's hypothesis that his wife has been replaced by a stranger explains certain observations even though it does not, strictly speaking, logically entail the observations.

The second model of abductive inference we wish to discuss, the Bayesian model, is better suited to explaining this form of inference in the context of theories or hypotheses with little or no formal structure. For this reason we shall focus on this model in the rest of our discussion. The crucial mark of the model is that it marries a natural probabilistic account of explanation to the standard Bayesian model of rational belief systems. Let us start with the probabilistic account of explanation. As remarked earlier, we need to be able to say that hypotheses can explain observational data even when these hypotheses do not logically imply these data. It is possible to capture this idea of explanation if we allow that explanation is not an-all-or-nothing matter and that it can come in degrees. Let us say that the hypothesis H explains observations O to the degree x just in case the probability of O given H is x, i.e., $P(O|H)$ is x. The limiting case is one where H logically implies O, in which case $P(O|H)$ is 1. This definition links degree of explanation to the value of a conditional probability function. It follows from the definition that one hypothesis H_1 explains observations O better than another hypothesis H_2 just in case $P(O|H_1) > P(O|H_2)$. For example, the hypothesis that Moriarty committed the murder explains the observed phenomena better than an

alternative hypothesis just in case the observed phenomena are more probable under the first hypothesis than under the second. Similarly, the deluded subject's hypothesis that his wife has been replaced by a stranger is a better explanation of the observed phenomena than an alternative hypothesis just in case these observed phenomena are more probable under the stranger hypothesis than under the alternative hypothesis.

The probabilistic definition of explanation has the virtue of providing a quantitative measure of a hypothesis' explanatory power (*vis-à-vis* some observations). It also has the virtue of being combinable with the standard Bayesian account of belief systems. The Bayesian account is supposed to be a model of the way a subject organises her belief system and the way she revises her belief system in the light of observational data. There are two basic representational assumptions made in the account. First, a subject's belief system is represented by the degrees of belief she has in certain propositions; and these degrees of belief can be represented as subjective probabilities that satisfy the axioms of the probability calculus. Second, a subject's degrees of belief or subjective probabilities change over time by the process of conditionalisation. Let us represent a subject's degrees of belief at an initial starting point in terms of the probability function P. The Bayesian model represents how this subject's belief system, as represented by her subjective probabilities, change after she has accepted a new observation O into her belief system. Let us suppose that her new degrees of belief continue to satisfy the axioms of the probability calculus; and let us symbolise them in terms of the *new* probability function P_n. Then the Bayesian thesis that belief change occurs by conditionalisation states that for any hypothesis H:

$$P_n(H) = P(H|O)$$

In other words, the value assigned by the new probability function P_n to any hypothesis H after the observational data O have been accepted is the same as the value assigned by the initial probability function P conditional on O.

The latter probability function can be expanded by Bayes' theorem into the following

$$P(H|O) = P(H).\ P(O|H)/P(O)$$

This equation tell us that a rational subject's conditional probability P(H|O) (the *posterior probability* of H conditional on the observational data O) is a function of P(H) (the *prior probability* of H), the probability P(O|H) (the *likelihood* of O given H), and P(O) (the initial probability of the evidence O).

The previous Bayesian formula breaks down the calculation of the conditional probability P(H|O) into several steps. One step is the calculation of the likelihood function P(O|H). According to the definition of explanation we have proposed, the value of this function should be given by how well the

hypothesis H explains the observational data O. If it explains them well, then the likelihood function will be high; and if it doesn't explain them well, the likelihood function will be low. But the likelihood function does not by itself determine the posterior probability P(H|O). It may be that the likelihood function P(O|H) is high, but because the prior probability of H, P(H), is low, the posterior probability is also relatively low. David Hume's remarks about miracles illustrate this point. The hypothesis that a miracle has occurred (H) may explain, in the sense of being a good fit with, the fact that a witness has reported seeing the miracle (O). Indeed the likelihood function P(O|H) may be very high. But this does not mean that a person should give a high probability to the hypothesis that the miracle has occurred on the basis of the report of the witness. For the prior probability of the hypothesis that the miracle has occurred, P(H), is very low indeed: 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. In Hume's view, a person should give minimal credence to the hypothesis that a miracle has occurred on the basis of a witness report. The Bayesian formula for calculating posterior probabilities vindicates Hume's plausible view.

In this exposition we have made no mention of the probability of the evidence O. In many Bayesian models of hypothesis confirmation, this probability does not play a central role. It is a commonplace of such accounts that hypothesis confirmation is almost always comparative in character: We are seldom interested in the question whether one hypothesis is confirmed by observational data in absolute terms, and almost always interested in the question which of two or more available hypotheses the observational data best confirm or support. When comparing the posterior probability of the hypothesis H with that of another hypothesis H′, Bayes' theorem implies the following equality holds

$$\frac{P(H|0)}{P(H'|0)} = \frac{P(H).\ P(O|H)}{P(H').\ P(O|H')}$$

In words, this says that the ratio of the posterior probabilities of H and H′ is a function of the ratios of the prior probabilities of these hypotheses and ratios of the likelihoods of the observational data given these two hypotheses. In this equation the probability of the observational data, P(O), has disappeared because it is cancelled out in the multiplication processes.

We have now assembled all the pieces needed to give the Bayesian account of abductive inference. As we noted at the outset, an account of abductive inference must supply an answer to the question: When is it reasonable to infer a hypothesis H on the basis of observational data O? Or perhaps the

question may be more naturally expressed in comparative form: When is it reasonable to prefer one hypothesis H over another hypothesis H′ on the basis of the observational data O? The Bayesian model answers this question by saying that it is rational to prefer the hypothesis H to hypothesis H′ on the basis of the observed data O just when the conditional probability P(H|O) is higher than the conditional probability P(H′|O). When this condition holds, the model states that H is *better supported or confirmed* by these data than H′. An abductive inference to the hypothesis H from O is justified or reasonable, then, just when H is better supported by O than any other available hypothesis H′.

In this Bayesian account of abductive inference, we can see the two-step procedure evident in the logical empiricist account. The first step consists in the assessment of how well one or more hypotheses explain the observational data. In the logical empiricist account, this is decided by whether each hypothesis logically implies the observational data, whereas in the Bayesian account this is determined by a comparison of the likelihood functions associated with each hypothesis. The second step of the assessment procedure consists in comparing rival hypotheses on the basis of their general plausibility. In the logical empiricist account, the second step is effected by comparing rival hypotheses in terms of their simplicity, their level of "adhocness", and their coherence with background assumptions. In the Bayesian account, this assessment is made by comparing the prior probabilities of the rival hypotheses, where these probabilities express a measure of the general plausibility of the hypotheses based on their simplicity and overall coherence with background assumptions. The second step in each account seems to involve an irremediable element of subjectivity: There does not seem to be any objective measure of the general plausibility of a hypothesis in either account. The virtue of the Bayesian theory is that it gives us a quantitative measure of each element in the assessment procedure and a way of combining them (via Bayes' theorem) to yield an overall measure of the extent to which it is reasonable or justified to accept a hypothesis on the basis of certain observations.

It is worthwhile noting two important restrictions in scope of the Bayesian account of abductive inference. One restriction is that the 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. The other restriction is that the Bayesian account says nothing systematic about the nature of observations, or the conditions of their rational acceptance into belief systems. The account is silent about questions such as what counts as an observation and to what extent observations are theory laden and so on. The account characterises observations as the starting points for the process of abductive

reasoning: Observations are whatever propositions are taken for granted and treated as givens in the process of testing the explanatory adequacy of hypotheses and theories. However, the Bayesian account does presuppose it is rational for subjects in suitable circumstances to trust their senses and the testimony of others and to incorporate such observational data into their belief systems. If this was not rational, the Bayesian account of abductive reasoning could never explain belief change or revision: The procedure would simply describe the hypothetical process by which hypotheses would be confirmed if certain observations were made without ever implying that subjects actually revised their hypotheses on the basis of what they observed. Clearly, it is normal and usually rational to accept the deliverances of one's senses and the testimony of others; and the Bayesian account presupposes that this is rational even though it does not provide a systematic account of the specific circumstances in which it is rational. This second point will prove to be significant in our discussion of when the reasoning of Capgras subjects lapses into irrationality.

We must now apply the account of abductive inference we have sketched to our problems about delusional belief systems. But first we offer a more general observation about the very possibility of explaining delusions in this way, prompted by the pessimism of one philosopher of cognitive science about the idea of such explanations. In *The Mind Doesn't Work That Way*, Jerry Fodor argues that "by all the signs, the cognitive mind is up to its ghostly ears in abduction" (2000, p. 78). Fodor identifies holistic abductive reasoning as the characteristic form of belief fixation in human cognition. Computational processes in an autonomous and encapsulated peripheral module access only the information in that module's proprietary database. Such computational processes may in a sense be abductive in nature, but the fact that they draw only on local information makes them tractable to model. In contrast, our "central systems" typically employ a quite different kind of abductive reasoning: "global" or "context-sensitive" thinking, in which potentially any other background beliefs we hold may influence the current reasoning process. Even though such reasoning is, as we have suggested, fallible and defeasible, we rely entirely on it: As Paul Thagard puts it (2007, p. 226), "despite its inherent riskiness, abductive inference is an essential part of human mental life".

Fodor is pessimistic about the scope of cognitive science precisely because of the centrality of this global kind of abductive reasoning. For, he argues, "we do not know how abduction works. So we do not know how the cognitive mind works; all we know anything much about is modules" (2000, p. 78). The problem, in his view, is that the global, holistic forms of abductive reasoning that central systems typically employ seem to involve a paradox. When Sherlock Holmes inferred that the criminal is Moriarty, he drew on a scattered array of background assumptions about fingerprints, blood stains,

human motivation, and so on, as well as his recent observational data: His effective use of this entire system of background beliefs was what gave his abductive reasoning its reliability. But Holmes didn't laboriously search through every single item in memory: To be effective, his abductive inference required the use of only a small subset of his beliefs. The paradox of this global form of abduction, then, is that on the one hand, in principle "the whole background of epistemic commitments [should] be somehow brought to bear in planning and belief fixation", while on the other hand it is only feasible in real time to access only the relevant information (Fodor, 2000, p. 37). Yet somehow, as if by magic, we ordinary reasoners, like Sherlock Holmes, successfully manage this selective access, judging relevance well enough to get by, most of the time. This is one version of the frame problem for artificial intelligence (Copeland, 1993, pp. 91–117; Fodor, 2000, pp. 37–38; Pylyshyn, 1987): Our models and theories offer no good formalism for specifying or delimiting in advance just how many and which other items of background knowledge should be included as relevant for any episode of reasoning. Fodor argues, in particular, that the computational theory of mind can't capture the way that we draw on the right beliefs for the current context (2000, pp. 23–53): Until someone has a good idea about how to model abductive inferences "which are sensitive to global properties of belief systems", we should hold off on studying the "more interesting and less peripheral parts of the mind" (pp. 98–99; cf. Fodor, 1975, pp. 197–205, on the limits of computational psychology).

But perhaps the prospects for computational accounts of global forms of abduction are not as gloomy as Fodor suggests. Some researchers seek to integrate computational and cognitive neuroscientific accounts of the mechanisms of inference (see Thagard, 2007, for a review; also Thagard & Shelley, 1997); and although they are not our topic here, there are also promising attempts to formalise models of inference in Bayesian networks and machine learning (Pearl, 1988). But more to the point in the current context is a direct response to Fodor. If we are right that the Bayesian model of abductive inference can help us both explain and evaluate the rationality of typical inferences made by delusional subjects, then the possibility of scientific study of central systems will be vindicated, and hence Fodor's "First Law of the Nonexistence of Cognitive Science" (Fodor, 1983) will be refuted.

In delusional subjects, we have argued, the balance between the whole background belief system and the particular evidence base is not successfully achieved. Capgras patients, for example, exhibit nonstandard reasoning in the sense that they do not efficiently use the right subset of background beliefs, or check their hypothesis effectively against other information available to them. As a result, their abductive reasoning isn't reliable. An account, within our Bayesian framework, of the peculiarities of their reasoning would go some

way towards countering Fodor's pessimistic assessment. We now attempt to offer such an account.

THE APPLICATION TO DELUSIONAL BELIEF

We are now in a position to explain how the account of abductive inference we have sketched vindicates Maher's basic contention that delusional beliefs arise via rational inferential responses to highly unusual data. We shall evaluate the rationality of typical inferences made by delusional subjects within the Bayesian model of abductive inference.

Let us consider the situation of a person with Capgras delusion who is confronted with some highly unusual data: in seeing the person who claims to be his wife, the heightened activity in his autonomic nervous system that occurred on previous encounters with his wife now does not occur. Let O be the proposition that describes the data that he now observes. How is the subject to explain O? One hypothesis is that the person who claims to be his wife is actually a stranger: Label this hypothesis H_S. Another hypothesis is that this person really is his wife: Label this hypothesis H_W. Which hypothesis is better supported by his observations?

According to the Bayesian model, the first step in working out which hypothesis is better supported is to ask: Which hypothesis better explains the observed data? Or in other words, are the observed data more probable under the stranger hypothesis H_S or under the wife hypothesis H_W? The striking thing about these two particular hypotheses is that the observed data are clearly much more likely under the stranger hypothesis than under the wife hypothesis. It would be highly improbable for the subject to have the low autonomic response if the person really was his wife, but very probable indeed if the person were a stranger. Indeed, the difference in probability becomes more marked the more specific we imagine the observed data to be. If, for example, we suppose that what the subject unconsciously detects is a highly anomalous contrast between the high autonomic response he would expect to his wife and his low autonomic response occurring currently, then the stranger hypothesis will explain his observations much better than the wife hypothesis. So we conjecture that the person's subjective likelihood functions will make the following ratio very high:

$$\frac{P(O|H_S)}{P(O|H_W)} \qquad \text{(A)}$$

This ratio does not by itself determine which hypothesis is better supported by the observed data. To determine this we have to know the prior probabilities of the two hypotheses. It would seem that a subject might give a very low prior probability to the stranger hypothesis H_S and a very high

prior probability to the wife hypothesis H_W in view of the general implausibility of the first and the general plausibility of the second. So it would be reasonable to suppose that the subject might assign prior probabilities to the two hypotheses so as to make the following ratio very low:

$$\frac{P(H_S)}{P(H_W)} \quad \text{(B)}$$

Nonetheless, the crucial feature of the comparative form of Bayes' theorem is that the ratio of posterior probabilities

$$\frac{P(H_S|O)}{P(H_w|O)} \quad \text{(C)}$$

can still be high even when ratio (B) is low provided that the ratio (A) is sufficiently high. For example, if we suppose that the subject's prior probability $P(H_S)$ is 1/100 and $P(H_W)$ is 99/100 then the ratio (B) will be low: 1/99. But if we suppose that $P(O|H_S)$ is very high, say 999/1000, and that $P(O|H_W)$ is very low, say1/1000, then the ratio (A) will be very high, i.e., 999/1. In this case, the ratio of posterior probabilities (C) will also be high: 999/99. In other words, the posterior probability of the stranger hypothesis H_S will be more than 10 times greater than the posterior probability of the wife hypothesis H_W.

The general point here is that if the stranger hypothesis explains the observed data much better than the wife hypothesis, the fact that the stranger hypothesis has a lower prior probability than the wife hypothesis can be offset in the calculation of posterior probabilities. And indeed it seems reasonable to suppose that this is precisely the situation with the subject suffering from Capgras delusion. 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.

This account of the nature of the inferential processes which yield the delusional belief completes the first explanatory task we set ourselves. But of course everything we have said so far in this section applies to every patient who has suffered neuropsychological damage that has disconnected the face recognition system from the autonomic nervous system; and as we pointed out earlier, only some of these patients exhibit Capgras delusion. Others (those with ventromedial frontal lesions) are not delusional despite the presence of the same abnormal data that is, we argue, causally implicated in the genesis of the Capgras delusion. Why doesn't the presence of these abnormal data in the ventromedial patients result in delusion? How do these patients escape the incipient delusion?

We have argued that any subject with this neuropsychological disconnection will arrive at the belief that he is faced with a stranger through an abductive reasoning process that relies heavily on the fact that this hypothesis provides a better explanation of his highly unusual data than the rival hypothesis that the person really is his wife. However, after he has accepted the stranger hypothesis, new data relevant to this hypothesis will emerge that he should take into account; and these additional data should undermine his belief in the stranger hypothesis. 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. Let us suppose that these further data can be formulated by the proposition O*. Surely, an application of abductive reasoning to these new data should lead the subject to lower his credence in the stranger hypothesis. For example, let P* be the subjective probability of a subject who has already incorporated the data O into his belief system. Then the comparative form of Bayes' theorem states:

$$\frac{P^*(H_S|0^*)}{P^*(H_W|0^*)} = \frac{P^*(H_S).\ P^*(O^*|\ H_S)}{P^*(H_W).\ P^*(O^*|\ H_W)}$$

Here employing the same kind of reasoning as before, one should be able to show that the subject will assign a very low posterior probability to H_S by comparison with H_W because even if the ratio of prior probabilities of the two hypotheses is high (supposing the subject gives a higher initial probability to H_S than to H_W), the ratio of likelihood probabilities will be very low indeed. The reason for this is that whatever the earlier observational data were, 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 $P^*(O^*|\ H_S)$. In this case the wife hypothesis is much better supported by the new data than the stranger hypothesis. Our suggestion is that patients with ventromedial frontal damage reason in this way, and their taking into account the new evidence against H_S is the mechanism via which these patients avoid succumbing to a delusional belief in the longer term, even if this belief was initially entertained.

But how are we to explain the fact that Capgras patients, exposed to just the same data as the ventromedial patients (both the old abnormal data caused by the neuropsychological disconnection and the new data provided by wife, family, friends, and clinicians that contradicts the belief) respond differently to these new data? Why do they not reject the stranger belief on the basis of these new data that disconfirm it? 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.

Rejection of such evidence frequently takes the form of confabulation. Asked to explain how he know that the "stranger" is not his wife, a Capgras patient will often mention differences (which must of course be fictitious) between the physical appearance or characteristic behaviour of his wife and the "stranger" (for many examples see Coltheart & Turner, 2010, and other papers in this Special Issue).

So it seems as if the new information does not even enter the deluded subject's belief system as data that need to be explained (as distinct from needing to be explained away, via confabulation). It is a commonplace that scientists who are wedded to an explanatory theory have an irrational tendency to reject evidence that cannot be explained by their theory. The reaction of the deluded subject in response to the new information O* looks like a highly exaggerated and unreasonable version of the same tendency.

To be sure, it is sometimes reasonable to reject information that cannot be explained by the hypothesis one is committed to. And indeed the Bayesian framework explains how this can be so. Again, let P* be the probability function of a rational subject who has already incorporated the anomalous perceptual data O into his belief system. What credence should this subject give to the proposition O* before he actually learns it? The Bayesian model enables us to answer this question in terms of the theorem on total probability, which implies:

$$P^*(O^*) = P^*(H_S).\ P^*(O^*/\ H_S) + P^*(H_W).P^*(O^*/\ H_W)$$

(Here we assume for the sake of simplicity that H_W is the negation of H_S.) In other words, the credence that the subject should give to the information O* is a weighted average of the probability that the information is true under the two competing hypotheses. Assuming that the rational subject will take the wife hypothesis to be a much better explanation of the data O* than the stranger hypothesis, he'll give a low value to $P^*(O^*/\ H_S)$ and a high value to P*(O*/Hw). For the sake of illustration, suppose he gives the first the value 1/100 and the second the value 90/100. However, when these probabilities are weighted by the probabilities that he assigns the two hypotheses, say $P^*(H_S) = 99/100$ and $P^*(H_W) = 1/100$, the previous formula implies the overall credence assigned to O* will be very low; indeed, it will be just less than 2/100. So even though the data O* would be much better explained by the wife hypothesis than the stranger hypothesis, the low credence that is given to the wife hypothesis at this stage means that the rational subject should not expect the observational proposition O* to be true.

Of course, this is the credence that a rational subject should give to the proposition O* *before* actually learning it. As we noted earlier, the Bayesian

model presupposes that it is reasonable in suitable circumstances to accept the evidence of one's senses and the testimony of others. So, for example, when the ventromedial patient sees the woman wearing his wife's wedding ring and hears his friends and clinicians repeatedly tell him she is indeed his wife, it is reasonable for him to accept this information regardless of the fact that he earlier gave it low credence. It is precisely at this point that the Capgras patients differ from the ventromedial patients: The 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. Although their abductive reasoning is faultless in response to the endogenous data caused by their neuropsychological impairment, they fail at a later stage to accept the exogenous information available to them through their senses and the testimony of others. This is the way in which they differ from the ventromedial patients, who do not suffer from this failure.

So we have arrived at an account of the phenomena we set out to explain. We have proposed that Capgras patients and ventromedial patients who are subject to the same neuropsychological impairment both initially favour the stranger hypothesis over the wife hypothesis. This piece of reasoning is a perfectly rational response to very abnormal data. In both sets of patients this reasoning results in the conscious belief "This is a stranger who looks like my wife". However, as time goes on, both kinds of subjects receive additional information, much of it in the form of the evidence of their senses and the testimony of others. The ventromedial subjects respond rationally to this extra information by revising the credence they give to the two hypotheses and settling on the wife hypothesis as the true hypothesis; so the delusional belief is rejected. In contrast, the Capgras subjects depart from rationality by rejecting this evidence: despite its seeming reliability, they irrationally ignore or discount the evidence on the basis of its incompatibility with the hypothesis to which they have become committed; so the delusional belief persists.

This 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. We know of no studies that have provided evidence in support of or conflicting with this conjecture; but future research could certainly provide such evidence (for example, by confronting patients with their spouses immediately after the occurrence of the brain insult, and assessing spouse identification).

What we have done here is to flesh out a little a two-factor theory of delusional belief, which has been developing over the past decade (Coltheart, 2007; Coltheart et al., 2007; Langdon & Coltheart, 2000). In that theory, Factor 1 is the abnormality, which is responsible for the content of a delusional belief (and hence differs for different beliefs; the first factors for seven delusions are listed in Column 1 of Table 1; the first factor is what

prompts the delusional belief. A delusional belief needs to be prompted in this way if it is to come to mind (something which occurs, we have suggested, via a process of Bayesian abductive inference); but when this happens the belief will ultimately be rejected unless Factor 2 is also present. This factor is, according to the two-factor theory, common to all forms of delusional belief. It has been rather vaguely described as "defective belief evaluation". We have been more specific about its nature: After Factor 1 has prompted an hypothesis about the endogenous data which prompts a delusional belief, new exogenous evidence emerges that is inconsistent with that belief, and the second factor is a failure of the system whose job it is to consider new evidence of this type so as to revise current beliefs.

The idea here is that Factor 1 is present in both Capgras and ventromedial patients, whereas Factor 2 is present only in Capgras patients. There is a little neuroanatomical support for these claims. In the case of Capgras delusion, Factor 1 is a disruption at some point in the neural pathway from the face recognition system to the autonomic nervous system. Nothing is known about which point in that pathway is damaged in Capgras patients, but the data from skin conductance studies of autonomic responding to familiar and unfamiliar faces by Capgras patients shows that it must be damaged at some point. For the same reason this pathway must also be damaged in the patients with ventromedial frontal lesions, and here we do have some evidence as to where this pathway is damaged. Tranel et al. (1995, pp. 431–432) consider that the pathway from face recognition to autonomic nervous system involves "triggering of central autonomic control nuclei from the ventromedial frontal cortices" and so if there is ventromedial frontal damage this triggering "would be precluded, and electrodermal skin conductance responses would not occur in response to the types of stimuli used in these experiments" (i.e., familiar and unfamiliar faces).

What about the neuroanatomical basis of Factor 2 in Capgras delusion? Coltheart (2007, pp. 10–11) reviewed a number of neuropsychological studies of patients with Capgras or other delusions, and pointed out that damage to the right frontal lobe has very commonly been reported in such studies. Hence, the ability to update the belief system on the basis of new evidence relevant to any particular belief may depend on the integrity of some region in the right frontal lobe. The ventromedial patients studied by Tranel et al. (1995) did not have damage to lateral regions of the right frontal lobe; their damage was to medial regions of both frontal lobes. We speculate therefore that what all seven delusions represented in Table 1 (and other delusions also) have in common is damage to some lateral region of right frontal cortex, a consequence of which is that subjects are impaired at revising preexisting beliefs on the basis of new evidence relevant to any particular belief.

An objection to this sweeping claim that immediately springs to mind is this. Surely odd beliefs of all kinds occur to all of us from time to time. If so,

wouldn't someone who could not evaluate beliefs become delusional over time about all kinds of things? But that is not what the kinds of patients referred to in Table 1 are like: Their delusions are monothematic (confined to a single belief or set of related beliefs).

A response to this objection can be developed from the observations of Bisiach, Rusconi, and Vallar (1991). Their delusional patient AR, who had suffered a right hemisphere stroke that had paralysed her left arm, exhibited somatoparaphrenia in relation to that arm: She believed it to be her mother's. In the course of a neurological investigation, the patient received a standard assessment for vestibular dysfunction, irrigation of the left external ear canal with ice water. That abolished the delusion; after receiving this cold caloric left vestibular stimulation, the patient acknowledged that her left arm was hers. Two hours later, the delusion had returned.

Brain imaging studies have shown that cold caloric left vestibular stimulation produces activation of the right hemisphere, including regions of the right frontal lobe (Fasold et al., 2002; Lobel, Kleine, Le Bihan, Leroy-Willig, & Berthoz, 1998). If AR's ability to counter delusional belief via revising preexisting beliefs had been completely destroyed, nothing should be able to overcome the delusion. Instead, it seems plausible to argue here that the right frontal system responsible for such updating was impaired rather than abolished, and so could be recruited when there is abnormally high activation of the right frontal lobe, as is produced by the left cold caloric vestibular stimulation.

This hypothesis allows us to offer an explanation for why patients with Capgras and other delusions are delusional about only one topic. The odd thoughts that all of us have come and go, and because of their fleeting nature even a deficient belief updating system can prevent them from turning into delusional beliefs. But the failure of autonomic response to a spouse's face is omnipresent, and so the evidence for the stranger hypothesis is permanently powerful—too powerful to be overcome by a weakened belief updating system.

The hypothesis also allows us to say something about another feature of Capgras and other delusions: that they wax and wane. There are reports in the Capgras literature of patients who at some times believe the spouse to be a stranger and at other times have the correct belief about the spouse's identity. When patients first express some delusional belief, it is common for family, friends, and clinicians to challenge the belief, to assure the patient that it is false, and to offer various lines of evidence against it: These are the O* data we discussed earlier. If this campaign against the delusional belief is strong enough, and if the right frontal belief updating system is not too seriously impaired, the result may be that the patient can respond to the O* data by rejecting the delusional belief. Now there is no longer any need for others to campaign against the belief, and so the O* data, data inconsistent

with the delusional belief, cease to be presented to the patient. But data consistent with the delusion—the data listed in Column 1 of Table 1—remain present. So the delusion returns.

A NOTE ON HELMHOLTZIAN UNCONSCIOUS INFERENCE

The idea that stimulus recognition and identification involves unconscious processes of abductive inference is of course far from new; it goes back at least as far as Helmholtz (1867/1910), and figures prominently in the theoretical work of Irvin Rock (1983). The information collected by the senses is inherently ambiguous. For example, when an object has an elliptically shaped retinal image, that might be because what is being viewed is an ellipse in the frontoparallel plane; but the observation is equally compatible with this object being circular but tilted. There are thus effectively an infinite number of questions to the answer: Given that I have this particular image on my retina, what must the world be like? Hatfield (2002, p. 116) commented on Helmholtz's approach to such issues as follows: "Helmholtz maintained that perception draws on the same cognitive mechanisms as do ordinary reasoning and scientific inference (1867/1910, 3: 28–29)", and he commented on Rock's approach as follows:

> "The most explicit recent analysis of unconscious inferences in perception is due to Irvin Rock (1983). Rock identified four sorts of cognitive operations at work in perception, [two of which were] (1) unconscious description, in the case of form perception (1983, ch. 3); (2) problem solving and *inference to the best explanation*, in the case of stimulus ambiguity or stimulus features that would yield unexplained coincidences if interpreted literally (1983, chs. 4–7)." (p. 125, emphasis added)

The similarity between these ideas about how perception works and the ideas about the genesis of delusional belief which we introduced with our quotations from William James and Brendan Maher at the beginning of this paper and which we have elaborated upon in our paper could hardly be clearer.

CONCLUSIONS

We have offered a general account of the seven delusions described in Table 1—which perhaps also applies to other forms of delusional belief too. The first step in the genesis of a delusional belief is the introduction into a person's unconscious (or conscious) mental life of some new and at present inexplicable observation O. These are the observations listed in Column 3 of

Table 1, and, at least for the seven delusions considered there, the source of the abnormal observation is the neuropsychological impairment given in Column 1.

The next step involves the patient seeking to explain the novel observation. This search involves discovering some hypothesis H about the world, which, if true, would make O very probable. (Of course, it is the facts described by the observation O rather than the mere occurrence of the observation that the hypothesis H must make probable.) There will be many possible potential states of the world such that, if any of these were actually true of the world, the facts described by O would be likely to obtain. The choice between these hypotheses that is part of the process of abductive inference is, we have suggested, done in a Bayesian manner that depends crucially on the likelihood function P(O|H), the (subjective) probability of O given the truth of H. No matter how unlikely a particular H is, this probability function may be higher for that H than for any other Hs no matter how much more plausible these other Hs are than the H which is adopted. So an H for which the conditional probability P(O/H) is particularly high can be assigned a very high value of the posterior probability P(H|O) even when P(H) is very low; and it is these posterior probabilities that control the adoption of hypotheses as beliefs.

But we have been at pains to emphasise that even when everything described in the previous paragraph occurs to someone, that person will not necessarily become delusional; these counterexamples are listed in Column 4 of Table 1. What is needed to create a full-blown delusional case is a second cognitive impairment, one that prevents new observations occurring after the initial adoption of the hypothesis as a belief from causing the rejection of that belief, as they ought to because they conflict with that belief. Our present account implies, though, that even for patients without this second impairment—patients in Column 4 of Table 1—the delusional belief will initially occur, but then be rejected. As we have noted, this is speculative: We would welcome further studies of the behavioural and phenomenological differences between these two groups of patients—for example, between Capgras patients and ventromedial patients—in the very early aftermath of neuropsychological damage.

We have largely used the Capgras delusion to work through this account, but the application of this account to at least some of the other delusions in Table 1—the Cotard and Frégoli delusions, for example—is readily apparent. What needs to be done next is to work through all these seven forms of delusion—and others—to determine whether or not a plausible two-factor account involving a Bayesian process of abductive inference can be spelled out for all.

REFERENCES

Binkofski, F., Buccino, G., Dohle, C., Seitz, R. J., & Freund, H. J. (1999). Mirror agnosia and mirror ataxia constitute different parietal lobe disorders. *Annals of Neurology*, *47*, 553–554.

Bisiach, E., Rusconi, M. L., & Vallar, G. (1991). Remission of somatoparaphrenic delusion through vestibular stimulation. *Neuropsychologia*, *10*, 1029–1031.

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

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

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

Coltheart, M., & Turner, M. (2009). Confabulation and delusion. In W. Hirstein (Ed.), *Confabulation: Views from neurology, psychiatry, neuroscience and philosophy* (pp. 173–188). Oxford, UK: Oxford University Press.

Copeland, J. (1993). *Artificial intelligence: A philosophical introduction*. Oxford, UK: Blackwell.

Ellis, H. D., Lewis, M. B., Moselhy, H. F., & 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*, 255–269.

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

Fasold, O., von Brevern, M., Kuhberg, M., Ploner, C. J., Villringer, A., Lempert, T., et al. (2002). Human vestibular cortex as identified with caloric stimulation in functional magnetic resonance imaging. *NeuroImage*, *17*, 1384–1393.

Fodor, J. A. (1975). *The language of thought*. New York: Crowell.

Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A. (2000). *The mind doesn't work that way*. Cambridge, MA: MIT Press.

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

Hatfield, G. (2002). Perception as unconscious inference. In D. Heyer & R. Mausfeld (Eds.), *Perception and the physical world: Psychological and philosophical issues in perception* (pp. 113–143). New York: Wiley.

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

Helmholtz, H. L. (1867). *Handbuch der physiologischen Optik*. Leipzig: L. Voss. Reprinted 1910, with extensive commentary, in A. Gullstrand, J. von Kries, & W. Nagel (Eds.), *Handbuch der physiologischen Optik* (3rd ed.). Hamburg and Leipzig: L. Voss.

Hirstein, W., & 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. *Series B: Biological Science*, *264*, 437–444.

James, W. (1950). *Principles of psychology* (Vol. 2). New York: Henry Holt & Co. (Original work published 1890).

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

Lobel, E., Kleine, J. F., Le Bihan, D., Leroy-Willig, A., & Berthoz, A. (1998). Functional MRI of galvanic vestibular stimulation. *Journal of Neurophysiology*, *80*, 2699–2709.

Magnifico, F., Misra, V. P., Murray, N. M. F., & Mathias, C. J. (1998). The sympathetic skin response in peripheral autonomic failure—evaluation in pure autonomic failure, pure cholinergic failure and dopamine-hydroxylase deficiency. *Clinical Autonomic Research*, *8*, 133–138.

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

Mandler, G. (1975). Consciousness: Respectable, useful and probably necessary. In R. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 229–254). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufman.

Pylyshyn, Z. W. (Ed.). (1987). *The robot's dilemma: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Ramachandran, V. S., & Blakeslee, S. (1998). *Phantoms in the brain: Human nature and the architecture of the mind*. London: Fourth Estate.

Rock, I. (1983). *The logic of perception*. Cambridge, MA: MIT Press.

Thagard, P. (2007). Abductive inference: From philosophical analysis to neural mechanisms. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Cognitive, mathematical, and neuroscientific approaches* (pp. 226–247). Cambridge, UK: Cambridge University Press.

Thagard, P., & Shelley, C. P. (1997). Abductive reasoning: Logic, visual thinking, and coherence. In M. L. Dalla Chiara, K. Doets, D. Mundici, & J. van Benthem (Eds.), *Logic and scientific methods* (pp. 413–427). Dordrecht, The Netherlands: Kluwer.

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

Turner, M., & Coltheart, M. (2010). Confabulation and delusion: A common monitoring framework? *Cognitive Neuropsychiatry*, *15*(1/2/3), 346–376.

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