Monothematic Delusions
and
The Nature of Belief

Sam Wilkinson

A thesis submitted for the degree of Doctor of Philosophy

Philosophy Department

School of Philosophy, Psychology and Language Sciences

The University of Edinburgh

2013

Declaration

I hereby declare that this thesis is of my own composition and that it contains no material previously submitted for any other degree or professional qualification. The work in this thesis has been produced by myself, except where due acknowledgement is made in the text.

Sam Wilkinson

Contents

Acknowledgements

I would like to express my gratitude, first and foremost, to my two supervisors, Matt Nudds and Matthew Chrisman. Without their guidance and persistent help this thesis would not have been possible. I would also like to thank Julian Kiverstein, who stood in as my secondary supervisor (during the academic year 2010-2011) when Matthew Chrisman was on research leave.

I would like to thank Max Coltheart and Robyn Langdon for hosting me during a very productive month in April 2012, at the Macquarie Centre for Cognitive Science, in Sydney. In addition, I would like to thank Dominic Murphy at the University of Sydney, for inviting me as a visiting speaker, and for several fruitful discussions during my month in Australia.

Thanks to Elisabeth Pacherie and Jérôme Dokic who hosted me at the Institut Jean Nicod in Paris, in November 2010 and March 2011, and who have since been generous with their time and comments.

Financial support was provided by the Arts and Humanities Research Council for the first three years, and the Jacobsen Trust, for the fourth. Funding for my trips to Australia and Paris were provided by the George Scott Travel Scholarship and the Oxford Philosophical Fellowship Fund, respectively. I am extremely grateful for the opportunity to take these wonderfully productive visits.

Finally, I would like to acknowledge all those who have significantly helped me or influenced my thinking over the past four or five years. In alphabetical order; Tim Bayne, Felipe Carvalho, Andy Clark, Allan Hazlett, Pierre Jacob, Eric Mandelbaum, Conor McHugh, Kirk Michaelian, Michael Murez, Joelle Proust, Matthew Ratcliffe and Åsa Wikforss.

Abstract

In this thesis I argue that our philosophical account of the nature and norms of belief should both inform and be informed by our scientific understanding of monothematic delusions. In Chapter 1, I examine and criticise standard attempts to answer the question "What is delusion?" In particular, I claim that such attempts are misguided because they misunderstand the kind of term that "delusion" is. In Chapter 2, I look at the nature of explanation in psychology and apply it to delusions. In particular I look at the constraints on a successful explanation of a person's psychological state in terms of brain damage or dysfunction. I then propose, in Chapter 3, a way of understanding how delusions of misidentification arise. In particular, I criticise the standard view that they are formed via inference (in the relevant sense of "inference") on the basis of anomalous experience. I draw on empirical work on object and individual tracking, on dreams, and on the Frégoli delusion, and argue that inference is not only un-necessary, but is actually often bypassed in humans, for judgments of identification. The result is a non-inferential file-retrieval view. On certain views of belief, this would mean that the Capgras delusion lacks the right functional role to count as a genuine belief. In Chapter 4, I criticise such views of belief, and put forward a "downstream only" view. Roughly, something is a case of believing if and only if it disposes people to act in certain ways. I defend such a view against two serious and influential objections. In Chapter 5, I ask whether this means that the Capgras delusion can therefore safely be called a belief. I argue that there is a risk – even if one accepts the downstream only view of belief – that it still won't count as a belief, as a result of the subject's "incoherence" or "agentive inertia." However, I then distinguish egocentric from encyclopaedic doxastic states. This opens the possibility that one can truly say that the subject has the egocentric belief, "This man is not my father", but may fail to have the encyclopaedic belief, "My father has been replaced by an impostor". It also demonstrates that the question "Are delusions beliefs?" has been approached in an unhelpful way by the main participants in the debate.

This thesis is important because it shows the extent to which real-world phenomena can inform and be informed by central philosophical notions like belief. More precisely, it shows that the most plausible way of accounting for monothematic delusions involves abandoning both a strong normativism, and a discrete representationalism, about belief.

INTRODUCTION

"One *can't* believe impossible things."
"I daresay you haven't had much practice," said the Queen. "When I was your age I always did it for half-an-hour a day. Why, sometimes I've believed as many as six impossible things before breakfast."
Lewis Carroll, *Through the Looking Glass*

When we have a central philosophical concept, which seems to pick out a phenomenon in the world, but we discover that there are instances of (what at least *seem to be*) that phenomenon that are at odds with our philosophical concept, what is it that gives way? Do we alter our philosophical concept to fit the anomalous cases? Or do we claim that the anomalous cases don't really exemplify the phenomenon that the philosophical concept picks out? One major motivation behind this thesis is to address these questions, where the central philosophical concept is *belief*, and the anomalous case is delusion, and, in particular, monothematic delusion.

We call people "delusional", or say that they "have delusions", when they behave in certain ways, and more particularly when they make certain claims, and therefore seem to believe certain things. Although the paradigm cases occur in clinical contexts, we do sometimes talk of perfectly healthy people being "delusional" in everyday life. We might say to a friend, "You're delusional if you think the lunar landings were faked!" Although this use is clearly a non-literal exaggeration, it is illustrative insofar as it is intended to highlight the supposed implausibility or unreasonableness of the matter or opinion in question, namely, that the lunar landings were faked. Delusions, whatever they are, seem intuitively closely tied to notions of implausibility and unreasonableness.

More literal use of the word “delusion” appears in clinical and pathological contexts. In these contexts, patients appear to believe very strange things, sometimes very strange things *indeed*. Here is a sample of the sorts of things these patients might say.

1. “The man who lives with me and who looks just like my father is not my father.”
2. “I am being followed by an old friend in constantly changing disguises.”
3. “The person I see in the mirror is not me.”
4. “This arm [the speaker’s arm] is not my arm, it is yours.”
6. “I am dead.”
7. “Someone else is controlling my actions.”
8. “I am Napoleon.”

Clinical delusions like these can occur in the context of psychiatric disorders such as bipolar disorder, schizophrenia and Alzheimer’s dementia, but they can also occur as the defining characteristic of delusional disorders caused by brain damage (examples 7 and 8 above are of the former category, 1- 6 are generally of the latter). Delusions therefore have this interesting, if somewhat confusing, status of being both, at times, the symptoms of underlying pathology (as in schizophrenia or Alzheimer’s dementia), and at others, of single-handedly characterising delusional disorders. To add to this confusion, these delusional disorders are often themselves called “delusions”: “the Capgras delusion” (example 1 above), “the Frégoli delusion” (example 2), “the Cotard delusion” (example 6), these can all be used to label delusional *disorders* rather than belief-like states. Sometimes this confusion is averted by using “syndrome” as in “Capgras syndrome”, but this is seen by many as unsatisfactory (e.g. Persons 1986, Stone and Young 1997) because a syndrome technically involves a cluster of symptoms, whereas the Capgras “syndrome” only has one defining symptom, namely, the delusional misidentification itself.

In this introduction, I want to introduce the reader to the variety of delusional phenomena that are found in the clinic, and the ways in which they have been categorised. It is important to see that, like many other philosophers who work on delusions, I will be focusing primarily on a rare and narrow subset of delusional phenomena. I try, here, to justify this narrow focus. I will end by giving a chapter-by-chapter synopsis.

1. Classifying Delusional Phenomena: Descriptive and Aetiological taxonomies

The variety that is to be found among delusional phenomena is truly daunting. How can we classify different cases of delusion? There are two different kinds of taxonomy that we could use. One, which we might call a *descriptive taxonomy*, describes the surface features of the delusional behaviour; in other words, how the patient behaves both physically and in terms of utterances. The other, which we might call an *aetiological taxonomy*, divides cases in terms of the causes or mechanisms that give rise to them. The data that will contribute to such a taxonomy may be provided by a wide variety of techniques and disciplines, from cognitive psychology, to electrophysiology, to imaging techniques such as fMRI and PET, to computational modelling and so on. We will see in Chapter 2 that one major explanatory challenge involves the fitting together of findings from different disciplines and techniques.

It is important to see that the former, descriptive, taxonomy does not require this, since it is interested in surface features. Two patients who behave, or are disposed to behave, in terms of their utterances and physical behaviour, in relevantly the same way, are taken to have the same kind of delusion even if what is going on at a neural level is different. If such cases were to present themselves then this underlying difference would have serious consequences for treatment. One subject might respond well to different kinds of treatment than the other. Therefore a diagnosis on purely descriptive grounds ought to be considered incomplete in not only a theoretical but also a very practical sense.

1.1. Monothematic vs. Polythematic delusions

A common descriptive taxonomy is one that distinguishes between monothematic and polythematic delusions. As these labels suggest, monothematic delusions are restricted to one ("mono") "theme", whereas polythematic delusions cover many ("poly") "themes". More concretely, this means that a patient suffering from polythematic delusions will hold delusions covering a range of subject matter. Note that here the word "delusion" is used in the broad clinical sense, as something that a patient suffers from, rather than as a mental state, otherwise it would make little or no sense to speak of "*a* polythematic delusion". A polythematic delusion, in the first sense, is a symptomatic cluster comprised of a number of different delusions (belief-life states) in the second sense. An example of a polythematic delusion is a persecutory delusion (see e.g. Freeman 2008), where a patient will appear to have a variety of strange inter-related beliefs, such that the government is out to "get her", that they are stealing her mail, and controlling her actions, and that they have assassinated various key historical figures. Polythematic delusions tend to occur in cases of, and be primary symptoms of, schizophrenia. They can be fleeting and constantly evolving. They also tend to be mood congruent, which is to say, in keeping with the general mood of the subject. Thus delusions of persecution will occur against a background of an overall mood of paranoia.

Monothematic delusions, on the other hand, are restricted to a specific domain. Delusions of misidentification tend to fall under this category. A well-known example of such a delusion is the Capgras delusion, where the patient claims that a loved one (e.g. Lucchelli and Spinnler 2007) (or a small selection of loved ones (e.g. Hirstein and Ramachandran 1997), or sometimes a wider group of people (Ellis et al. 1997)) looks like the loved one in question but isn't in fact them. It is often called (perhaps not optimally, as we shall see in Chapter 5) “the delusion that a loved one has been replaced by an impostor”. Unlike polythematic delusions, monothematic delusions (in their purest forms) tend to occur in the context of localized brain damage, rather than in schizophrenia. Also, unlike polythematic delusions, they tend to be mood incongruent. Indeed it is sometimes hard to think what mood they might be congruent with: what kind of mood might lead you to misidentify yourself in a mirror? However, there are some monothematic delusions that seem to allow for mood congruent interpretations, but we need to be very careful. Delusional misidentifications that occur in schizophrenics (in patients who have polythematic delusions) seem to be more of a paranoid nature (“You’re an impostor, trying to hurt me” (cf. de Pauw and Szulecka 1988)) whereas those that occur in the context of brain damage can seem to be more like a calmer cognitive disruption (“He’s a nice guy doctor, but he’s not my father” (Hirstein and Ramachandran 1997, p.438)). Monothematic delusions also tend to be more fixed and stable. Both kinds of delusions, the (sometimes) fleeting, schizophrenic ones, and the stable monothematic ones, tend to be resistant in the face of contrary evidence (they are “tenacious”). As we will see in Chapter 1, this is closely related (although perhaps not strictly essential) to what makes us classify them as delusions in the first place.

1.2. Circumscribed vs. Elaborated

When a patient doesn’t draw consequences from her delusion, that delusion is said to be *circumscribed*. When such consequences are drawn or even extrapolated (and these consequences will themselves tend to be delusional) they are said to be *elaborated*. So, someone with persecutory delusions, who (for example) has a whole web of beliefs about the CIA stealing her mail and so on, is suffering from elaborated delusions. It is therefore unsurprising that monothematic delusions are circumscribed (if they were elaborated, they would be polythematic) and that polythematic delusions tend to be elaborated. It is worth noting that whereas the monothematic/polythematic distinction describes the apparent doxastic commitments of the agent, the circumscribed/elaborated distinction describes how the subject goes on to reason from her delusions: it seems to label dispositions to reason. It would be *in principle* possible for there to be polythematic delusions that weren't the result of a process of elaboration (i.e. arose independently), even if in practice this may not happen.[1]

1.5. Degrees of psychosis

Some theorists talk of degrees of psychosis (e.g. Heckers 2009). A delusional patient is more or less psychotic to the extent that she has lost "contact with reality". Thus, typically, the Capgras delusion in the context of brain damage is a delusion with a low degree of psychosis, since the patient is generally in touch with reality. She has a fairly normal take on how things stand in the world and the laws of nature that govern it: she just has this one very peculiar belief or set of beliefs (often) about a family member or two. Schizophrenic patients, on the other hand, living in solipsistic delusional alternate realities (often in realities where the laws of nature are flexible or broken) are suffering from paradigm cases of psychotic delusions.

1.4. Bizarre vs. Mundane

The bizarre/mundane distinction is interesting insofar as it alerts us to the fact that delusions can be about a variety of different types of subject-matter: they can range from thinking that your wife is being unfaithful (when evidence clearly doesn't support this), to thinking, as the Nobel Prize-winning mathematician John Nash did, that you are the left foot of God. Roughly the pattern here is that bizarre delusions tend to occur in the context of schizophrenia, and not generally in cases of brain damage. However, fairly mundane delusions can also occur in schizophrenics. The bizarreness of the delusion is usually thought

[1] A very similar distinction exists between systematic and unsystematic delusions. As the terms suggest, a systematic set of delusions all fit together as part of a system. Monothematic delusions are paradigm cases of unsystematic delusions.

of in terms of the outlandishness of the content, rather than the plausibility of holding it. Thus the Cotard delusion, the delusion that one is dead, occurs in the context of brain damage, and, although implausible to the degree of impossibility (since it is a self-defeating utterance, given that the person is alive enough to be saying I am dead), does not count as a bizarre delusion (although mundane-ness perhaps comes in degrees: it is plausibly less mundane than the Othello delusion, which is the delusional belief that your spouse is being unfaithful). This is because, although there is nothing mundane about *having* the delusion, the truth conditions are easy enough to grasp: we know what it would be for this person to be dead. He just obviously isn't because he's alive enough to be making such claims. The content lacks the florid weirdness of those delusions, typical of schizophrenia patients, that are labelled bizarre. With these bizarre delusions, it is sometimes not only hard to see how they might turn out true (i.e. the possible world in which they are true is far removed from the actual world) (for example, "Aliens are controlling my actions") but also sometimes hard to imagine what it would be for them to be true (i.e. hard to envisage what such a world might look like) (for example, "I am the left foot of God"). Some have suggested (e.g. Jaspers, Berrios, and others) that such delusional displays are better understood as empty speech acts or exhibits of word play rather than expressions of belief (we will see more on this in Chapters 4 and 5).

Non-bizarre delusions are considered easier to explain. Many, as we shall see, are considered straightforwardly rational responses to certain experiences, rather than the elaborate products of imaginative freewheeling. The Cotard delusion, for example, is often hypothesised to arise from reduced affective response to all stimuli, as caused by localised brain damage. Indeed sometimes patients with brain damage switch from one monothematic delusion to another, suggesting that there may be overlap in one of the aetiological elements of the delusion. McKay and Cipolotti (2007), for example, take the Capgras and Cotard delusions to have the same "first factor", namely experiential deficit or abnormality, but a different "attributional style". We'll see more on this in Chapter 3.

1.5. An Aetiological Dichotomy: Organic vs. Functional

Aetiologically speaking, delusions can arise in a number of different contexts, including, for example, bipolar disorder and dementia (or even, as we will see, towards the end of Chapter 1, in the "secondary" subject in a case of shared delusions or "*folie a deux*", without any discernible brain pathology at all), but the two that are the most extreme and most common respectively, and hence tend to be focused on, are brain damage and schizophrenia. These two distinct aetiological contexts tend to divide delusions into the category of the *organic* and the *functional*: delusions caused by brain damage were taken to have an "organic" basis, whereas delusions arising in the context of schizophrenia were taken to have a "functional" basis. The label "organic" indicates a biological basis, whereas the label "functional" indicates a lack of biological cause, taken to be something high-level, at the level of conceptual thought and reasoning, possibly even triggered by an experience in the patient's life. In crude terms, organic delusions were taken to arise from brain pathology, functional delusions from mind pathology.

Before schizophrenia was understood as having an underlying physical cause, it was common to try to understand the schizophrenic delusions in psychodynamic and motivational terms (which we will learn more about in Chapter 2). Since, firstly, it is now known that schizophrenia has a physical basis (the big breakthrough being the realisation that it can, at least sometimes, be successfully treated with medication), and secondly, psychodynamic accounts have been largely discredited, the organic/functional distinction has been more or less dispensed with. As Bortolotti 2009, puts it: "Today, the received view is that there is a biological basis for all types of delusions, but that in some cases it has not been identified with precision yet." This has lead many philosophers to treat delusions occurring in the context of brain damage and in the context of schizophrenia in the same way. For reasons that will become clear as we progress, although I agree that the original characterization of the distinction is misinformed, I think that lumping together the sorts of delusions that the distinction separated is dangerous. We should avoid the temptation to simplify in the face of such variety and complexity.

In spite of the fact that the organic/functional distinction has been largely dispensed with, it is important to distinguish a biological cause in a broad sense, which would incorporate neurotransmitter deficiency (which is widely taken to be present in schizophrenia), from an anatomical cause, which would be restricted to permanent changes in brain structures, viz. brain damage from strokes, traumatic brain injury or atrophy.[2] Even if schizophrenia has a biological cause, it is one that is different in *kind* from localised brain damage. That is abundantly reflected in the differing cognitive and behavioural profiles of brain-damaged and delusional schizophrenia patients.

2. What is to come

This thesis, overall, will focus mainly on monothematic delusions that occur in the context of brain damage, rather than those that occur in the context of schizophrenia. This is partly because they are more stable and tractable than the delusions that occur in schizophrenia: the behavioural and neuropsychological profiles of these patients are comparatively well established (although, as we will see, there is a great deal of variety among cases, and there is still much more work to be done to understand each case fully). More important, however, are the following two characteristics of monothematic delusions caused by brain damage. First, they are prime examples of something physical/physiological (i.e. brain damage) having a more or less direct impact on something (traditionally)

[2] There are in fact anatomical irregularities in schizophrenic patients. However, as Broome (2004, p.36) puts it: "That is not to say that patients with schizophrenia have normal cerebral anatomy and neuropsychology, but rather that the Capgras' and Cotard's patients are likely to have neurological quality to their presentation that is clinically relevant."

paradigmatically mental (i.e. belief).[3] This will be of particular importance in Chapter 2, and will be picked up in Chapter 3. Second, and in a way that relates to the first characteristic, they are the most problematic for our standard philosophical understanding of what belief is (which we will explore in Chapters 4 and 5). This is partly because, unlike patients with schizophrenia, patients suffering from these monothematic delusions exhibit a striking degree of normality beyond the delusion's domain (which is often some form of delusional misidentification, the mechanisms behind which we will explore in Chapter 3). As a result, we can't just dismiss them as completely irrational or "crazy." We will see that in other respects, too, cases of monothematic delusions do not resemble the paradigms of belief we find in mainstream analytic philosophy. Indeed, on some theories, the delusions in question aren't even formed on the basis of reasons: they are simply caused in a brute way by the brain damage. In Chapter 4 we will see various reasons why this is widely held to be problematic and unlike the paradigmatic philosophical picture of what beliefs are like.

The view I will seek to motivate throughout the thesis is that *some* central characteristics of a philosophical theory of belief must stand firm in the face of phenomena that appear to be beliefs but which fail to exhibit those characteristics. When this happens, the philosopher must insist, that, contrary to appearances, what we are presented with is not belief, in the relevant central philosophical sense (which will be elaborated in Chapter 4). However, on my view, the characteristics that must stand firm (namely, dispositions to coherent actions) are not *all* of those that are held by the most widely held philosophical views of belief. So, as I asked at the start of this introduction: Do we alter our concept of belief to accommodate the anomalous cases or claim that the anomalous cases don't really exemplify belief? My answer is: we need to carefully do a bit of both.

This introduction serves to introduce the subject matter of delusions generally, and to present the overarching aim of the thesis; an aim that I hope is clearly reflected in the title,

[3] This is, of course, not to deny that schizophrenia has physiological underpinnings. However, the interaction seems less direct. Delusional thinking seems to arise more gradually in schizophrenia, as a result of a more global breakdown in cognition.

Monothematic Delusions and the Nature of Belief. There are various sub-questions and sub-conclusions along the way. So, although there is a progression throughout the thesis, each chapter has a potentially freestanding take-home message. I list them in the following synopsis.

SYNOPSIS

Chapter 1 – What is Delusion?

“Definitions” of delusion are bound, in my view, to fail, since as I shall argue delusion is neither a natural kind, nor a fundamental theoretical concept, but rather a folk-evaluative concept. In clinical contexts, it has important diagnostic utility, but does not, in itself, do any scientific work. One consequence is that you can’t neatly define “delusion.” A further consequence of this is that, contrary to several theorists, “you can’t take the delusion out of delusion”. By this I mean that you can’t take a paradigmatically delusional phenomenon and discover that, in the light of a discovery about its true nature, it fails to be a delusion in light of your “definition.” The word “delusion” gets its meaning from the way it is used. Schematically put, suppose your definition requires that something have feature F (e.g. irrationality) in order to be a delusion. Then you discover that a paradigmatic case like the Capgras delusion lacks that feature. Whereas several theorists claim that this means that the Capgras delusion is not really a delusion, I claim that it shows instead that the definition is flawed. It demonstrates that delusions don’t necessarily have feature F.

Chapter 2 – Explaining Delusions

There are two very different kinds of explanation in psychology (broadly construed). These are personal and subpersonal explanations (a.k.a. personal and subpersonal “levels” of explanation). Within subpersonal explanation, there are many different “levels” that operate at different relevant “fineness of grain”. Some of these levels are characterised functionally, and as such are not reducible to “lower levels”. However, unlike what some theorists seem to think, personal explanation is not simply another level (e.g. the “top” level). It’s an altogether different, and vitally important, kind of explanation. In particular, it is a very special way of explaining belief and action. However, it is not always available, even in cases where it seems as though it might be, and it is very important to be aware of when it is or it isn't available. Different scientific aetiologies of delusions (e.g., of the Capgras delusion) have different consequences for the availability of personal explanation.

Chapter 3 – A Proposed Aetiology of Delusional Misidentification

Delusional misidentification is commonly understood as the product of an inference on the basis of evidence present in the subject's experience. For example, in the Capgras delusion, the patient sees someone who looks like a loved one, but who feels unfamiliar, so they infer that they must not be the loved one. I question this by presenting a distinction between "recognition" and "identification". Identification does not always require recognition for its epistemic justification, nor does it need recognition for its psychological functioning. Judgments of identification are often the result of a malfunctioning mechanism that tracks individuals. This tracking mechanism can be usefully thought of in terms of mental files. The aetiology I propose is therefore called "The non-inferential file-retrieval view".

Chapter 4 – On Failing to Believe: A Downstream-only View

The aetiology put forward in Chapter 3 would make several theorists deny that the Capgras delusion is genuinely a belief. Chapters 4 and 5 address this issue. It is important to distinguish the evaluative and constitutive norms of a phenomenon, i.e. the norms that dictate when something is a good or bad instance of the phenomenon in question (evaluative) and the norms that dictate when it ceases to be that phenomenon altogether (constitutive). When deciding what features of a belief-like phenomenon make it a bad belief, or no longer a belief at all, only downstream considerations are relevant for the latter. The result is a "downstream-only" view of belief, according to which belief can be formed in any old way; it may count as very bad believing, but as long as the subject has the right dispositions to *action*, it counts as belief. Such a view has been heavily criticised and I defend it against the two objections that I consider most threatening.

Chapter 5 – Is the Capgras Delusion a Case of Belief?

The delusional subject sometimes acts in ways that do not seem coherent with her delusional claims. This may suggest that, on the view put forward in Chapter 4, this prevents us from attributing the delusional belief. However, I first argue against the *discrete representationalism* implicit in the standard approach to the question of whether delusions are beliefs, in favour of a *holistic dispositionalist* view. Against this background, the relevant question is not, "Is the delusional state a belief state?" but rather "What does the delusional subject believe?" The really interesting problem with delusional patients is that we find it hard to attribute beliefs to them. It is not only because of the pathological nature of these subjects, that we find this hard. It is also partly because we have been using rather blunt tools in attempting such attributions. In particular, we have been using rather vague and ambiguous sentential attributions (e.g. "The subject believes that a loved one has been replaced by an identical-looking stranger"). Even the rather simple refinement of distinguishing egocentric from encyclopaedic believing, and reflecting on the relationship between the two, sheds valuable light on these cases. The Capgras patient can be more accurately attributed the egocentric belief "This man, here before me, is not my father", than the encyclopaedic belief "My father has been replaced by an identical-looking impostor".

CHAPTER 1

What is Delusion?

Introduction

Although the overall focus of this thesis is on delusions of misidentification caused by brain damage, in this opening chapter I'll look at what delusion is generally and try to see what, if anything, there is in common among all cases of delusion. In other words, I'll explore possible answers to the following question. Under what circumstances should we say (in a literal sense) that any given subject or utterance is delusional? To put it another way: Under what circumstances can we truthfully say of someone that they are delusional? Are there facts about whether people are delusional, and if there are, what do they look like?

The aim of this chapter, taken in isolation, is to characterise the concept of delusion, and to see why previous attempts at characterisation have failed. However, in terms of how this fits in with the thesis as a whole, the most important point is that delusion should not be conceptually tied to epistemic irrationality. This is important for views, which we will examine in more detail in Chapters 2 and 3, about the nature and aetiologies of specific delusions. Some aetiologies (e.g. Maher 1974) take (some) delusions to arise as a rational response to an unusual experience. However, there are theorists who think that aetiologies that don't attribute *irrationality* to the patient are in fact "taking the delusion out of delusion", i.e. showing that these paradigm cases of delusion are not genuinely delusions, since delusion is somehow conceptually tied to epistemic irrationality.[1]

Tying delusion to epistemic irrationality is very common. Thus, we get Moor and Tucker (1979, p.390) claiming that a delusion "is a belief that a person has, although he has

[1] I take the phrase "taking the delusion out of delusion" from Bayne and Fernandez's (2009) illuminating introduction to their co-edited anthology. They writes that "the conjunction of an epistemic conception of delusion and an empiricist-based account of delusions threatens to 'take the delusion out of delusion', for the upshot of the empiricist accounts seems to be that the patient's belief is not held despite what constitutes incontrovertible and obvious proof or evidence to the contrary."

[…] considerable evidence against the belief and comparatively no evidence for it." Similarly, within a rather different tradition, Jaspers distinguished delusions proper from mere "delusion-like ideas", which arise "understandably […] from false perceptions or from experience of derealization in states of altered consciousness" (1963 [1913], p.96). Very recently, and in a way that most clearly expresses the view I wish to oppose, Jennifer Radden has claimed:

> My own view is that "perceptual delusions" [like the Capgras delusion] do not rank as delusional states. […] As reasonable inferences from misleading perceptual experiences, "perceptual delusions" are not epistemic lapses of the sort by which delusional states are identified. (2012, p.28)

I don't think such views are right. In particular, they misunderstand the concept of delusion, and do so largely because they fail to appreciate the *kind* of concept that DELUSION is.[2] What makes someone delusional is not tied to irrationality, on the most common understanding of that term, namely, in the sense of an "epistemic lapse", of going "against the evidence". As a result, contra Radden, "perceptual delusions", like the Capgras delusion, remain, even if epistemically rational, paradigm cases of delusion.

As Jones (1999) points out, "three general surveys of delusion concluded that very little consensus has been reached on its nature, taxonomy and origins (Winters and Neal 1983; Butler and Braff 1991; Roberts 1992)." The view I will defend renders this completely unsurprising. DELUSION is not a natural kind concept. Nor is it a *scientifically* important theoretical construct. Many of the phenomena that we call delusions are extremely interesting in and of themselves, but that is very different to claiming that their classification *as delusional* does important scientific work. "Delusion" is in fact, firstly, a folk term, and it

[2] Following convention, I write the concept "delusion" as "DELUSION", but don't use small caps when writing about "the concept *of* delusion".

is, secondly, an evaluative term. The former means that one ought to be conservative with regards to it. Unlike natural kind or theoretical terms, the correctness conditions of the application of folk terms are determined by how the linguistic community uses those terms (these terms lack something I will call "answerability"). Characterising the concepts expressed by such terms requires one to pay attention to, rather than regiment, current usage. That it is an evaluative term suggests that it isn't straightforwardly descriptive and therefore doesn't admit of a characterisation in terms of necessary and sufficient conditions.

You can rephrase my point as follows. Any theorist, of course, can *stipulate* that they will use the word "delusion" in such and such a way. But if people have their own intuitions about the application conditions of the term, and these conflict with the stipulated conditions, such a theorist will have a hard time convincing others that this is how the word *should* be used, if, firstly, delusion is not a natural kind and, second, it does no vital theoretical work.

The structure of this chapter will be as follows. I will start by looking at two different approaches to defining delusion. These different approaches correspond to clinical concerns of diagnosis and treatment, on the one hand, and philosophical concerns of "getting the concept right" on the other. I focus on the latter, and oppose a common way in which theorists have attempted to achieve this. In particular, I look at various reasons to oppose views that tie delusion to epistemic irrationality, principally by arguing that epistemic irrationality is neither sufficient, nor is it necessary, for something to count as a delusion. I then motivate and present an alternative approach that sees delusions as an infringement of folk norms, and, in particular, a folk norm of "understandability". I end by looking at some illustrative cases of delusional phenomena.

1. Defining Delusion: Psychiatric and Philosophical Approaches

We will now look at ways of isolating the conditions in which we can truthfully (or appropriately) say that a person or utterance is delusional. We will see that various attempts at doing this fail, and that the problem is not with the details of the proposals, but the nature of the attempts themselves.

1.1. Delusion, Mental Illness and Clinical Practice

As Jaspers put it, "Since time immemorial, delusion has been taken to be the basic characteristic of madness" (1963, p.93). And as Jones, somewhat less strongly, puts it (1999, p.1) delusions "continue to guide clinicians in identifying psychosis. Delusion, therefore, has remained an important psychiatric concept". The clinician faces a wide variety of different psychopathological cases and has to decide how to proceed, if possible, with treatment of some sort. It is therefore common practice to have a checklist of symptoms for the purposes of diagnosis. For example, with schizophrenia you have a list of five symptoms, and if the patient presents with any two of the five symptoms then you diagnose them with schizophrenia.[3] [4]

Now, the point that Jones is making is that one of the symptoms on such a check-list (say, a check-list for schizophrenia) might be a delusion.[5] In order to know how to recognise a delusion, the psychopathologist might turn to the official manual, The Diagnostic and Statistical Manual Mental Disorders, or DSM for short. The DSM characterises a delusion as follows:

[3] The symptoms in question are: Delusions; Hallucinations; Disorganized speech (which is a manifestation of formal thought disorder); Grossly disorganized behaviour or catatonic behaviour; "Negative symptoms" which lump together lack of emotional response, alogia (lack or decline in speech), or avolition (lack or decline in motivation).

[4] These diagnoses are clearly not picking out a unified natural kind, something already actually out there in the world with a pre-existing nature (like water or gold, or perhaps belief) (cf. Bentall 1990). But when your job is to help the sick and troubled, this is not a primary concern.

[5] Being able to do this would presumably also allow the psychopathologist to diagnose not simply disorders that have delusions among their symptoms, but also delusional disorders that have a delusion as their one and only symptom or, better, as their defining characteristic.

> A false belief based on incorrect inference about external reality that is firmly sustained despite what almost everyone else believes and despite what constitutes incontrovertible and obvious proof or evidence to the contrary. The belief is not one ordinarily accepted by other members of the person's culture or subculture (e.g. it is not an article of religious faith). (DSM-IVTR, 2000: 821).

Many theorists (e.g. David 1999, Coltheart 2007, Bayne and Fernandez 2009) have noted that, although this definition is the "official" one, it is, if taken to present necessary and sufficient conditions, flagrantly problematic and subject to several counterexamples. For a start, delusions needn't be false; they could be accidentally true. Consider the "madman" who declares on the basis of no evidence that the end of the world is nigh, and (thanks perhaps to a helpful asteroid) just so happens to be right![6] Furthermore, do delusions have to be based on incorrect inference?[7] And what exactly do we mean by that? Do they have to be about external reality? To what extent should the beliefs of others in one's community play a role in determining delusional status? To what extent should religious beliefs, for example, be exempt from delusional status on these grounds? While some psychopathologists have stated these inadequacies in a calm matter-of-fact manner (e.g. Marshall and Halligan 1996), others have actively bemoaned the inadequacy of the official definition. Whether they are correct in seeing the DSM as a strict definition or not, in bemoaning the failure of the DSM to provide such a definition, they are in fact condoning such an attempt. In other words, they wish that they could "define" delusion, rather than seeing their inability to do so as being a necessary (and perfectly acceptable) consequence of the kind of concept that DELUSION is

[6] This may depend on how you characterise belief-contents, however. Characterising belief-sets in a rich and holistic way, there might be room to build into the content of the belief, the grounds on which it was formed. So the madman's delusion is wrong because he believes that the end of the world is nigh, *and* that this is true because of the voices in his head. Or we could say that he is delusional for the belief that the voices in his head are to be trusted.

[7] Radden (2012), and others, seem to answer this in the affirmative. The notion of inference (and absence thereof) will be central to our proposed aetiology in Chapter 3.

(We'll see more on this when I present my own view). For example, Anthony S. David (1999) speaks of the "Impossibility of Defining Delusions" in a paper of that title:

> Most attempted definitions begin with "false belief", and this is swiftly amended to an unfounded belief to counter the circumstance where a person's belief turns out to be true. Then caveats accumulate concerning the person's culture and whether the beliefs are shared. Religious beliefs begin to cause problems here and religious delusions begin to create major conflicts [...] The beleaguered psychopathologist then falls back on the "quality" of the belief - the strength of the conviction in the face of contradictory evidence, the "incorrigibility", the personal commitment, etc. Here, the irrationality seen in "normal" reasoning undermines the specificity of these characteristics for delusions [...] as does the variable conviction and fluctuating insight seen in patients with chronic psychoses who everyone agrees are deluded [...]. Finally we have the add-ons: the distress caused by the belief, its preoccupying quality, and its maladaptiveness generally, again, sometimes equally applicable to other beliefs held by non-psychotic fanatics of one sort or another. In the end we are left with a shambles. (p.17-18)

This lament (and it surely is a lament with strong negatively laden terms like "beleaguered" and "shambles") is revealing. First, David (who, it must be mentioned, is a psychiatrist, and hence from a medical background) seems to be assuming that delusions must, as a matter of conceptual necessity, be pathological. This can be seen from the observation that he takes "irrationality" in the normal, healthy, population to undermine a definition that might be based only on irrationality (i.e. rather than allowing that delusions in healthy people might not be a contradiction in terms).

David seems to be thinking of delusion in two different ways. First, he thinks of it as a diagnostically important psychiatric concept, a Jaspers-style "marker of madness", in the light of which non-pathological delusion is a contradiction is terms. But he also thinks of it and as a concept that has to be carefully defined and free of potential counter-examples. But there is tension here. What provides counter-examples to the DSM definition are cases where delusional status has already been recognised. That's why they are counter-examples. Therefore the DSM isn't guiding our judgements about delusional status.

So what is the DSM definition doing? If it is providing a strict definition in terms of necessary and sufficient conditions (which, to be fair, it superficially *seems* to be), then counter-examples will be damaging. If, however, it is seen as a description of paradigmatic cases of delusions, then it seems to do the job. Most delusions do fit the DSM description. That this is what the "definition" is really doing is, after all, suggested by the acronym DSM: it is a Diagnostic and Statistical Manual. If somebody didn't know how to recognize a delusion, it would be helpful in enabling him or her to do that. It is highly unlikely that such a person would encounter one of the counter-examples first. Similarly, the delusions that clinicians are likely to encounter will fit the DSM "definition". As we will see, my view is that delusional status is attributed on the basis of a negative evaluative reaction that reflects an adherence to folk epistemic norms.

Suppose the hypothetical person who needs the DSM to guide their delusion-attributions needs it because they simply don't have those negative evaluative reactions. What the DSM is doing is not defining the word "delusion" in the sense of picking out a natural kind, or a key theoretical concept. It is simply describing the kinds of things that typically arouse that negative evaluation in the linguistic community that uses the word "delusion". We can draw an analogy with the moral case. Suppose the psychopath, with no moral sensibilities, wants to behave in ways that are perceived by others as morally right. He might use a description (or even a list) of the sorts of things that are right or wrong. But the description doesn't guide normal people's judgements of what is right or wrong, and even less is such a description a *definition* of the word "wrong". And although the psychopath may use this description to guide his judgements, there are bound to be cases that don't fit the description.

Since theorists can recognise delusions when they encounter them (cf. "patients who everyone agrees are deluded"), when people criticise the DSM, it is not fear of misdiagnosis that is motivating them. It is that the DSM definition is failing to "capture the concept" of delusion, whether this is what it is trying to do or not. These attempts to capture the concept are driven by philosophical rather than clinical concerns. I think it is important to keep these separate. It is worth noting that there are a several terms in David's grievance that are epistemic ("unfounded", "irrational", "incorrigible"). So the following question seems to arise quite naturally.[8] Might it not be the role of the philosopher (and perhaps, more specifically, the epistemologist) both trained in conceptual precision and, most importantly, freed from the daily and pressing concerns of the clinic, to "define" delusion?[9]

1.2. Delusion, Rationality and Epistemology

Suppose that, instead of thinking of delusion as a marker of mental illness, (as Jaspers, David, and countless others have) we thought of the relationship between mental illness and delusion as follows. Although there remains a great deal of controversy surrounding the concept of mental illness (see Perring 2010 for a review), let's suppose illness, and mental illness as a subset thereof, is a practical notion (e.g. Reznek 1987), in the sense that a condition is pathological if it interferes with daily functioning or flourishing. Since it is often disruptive to be delusional, delusional subjects will *tend* to be mentally ill. However, even if it were always disruptive to be delusional, it would be fallacious to think

[8] Definitions sometimes blend epistemic and medical criteria. For example, we get (McKay et al. 2005) claiming that: "A person is deluded when they have come to hold a particular belief with a degree of firmness that is both utterly unwarranted by the evidence at hand, and that jeopardizes their day-to-day functioning" (McKay *et al.* 2005, p. 315).
(Freeman 2008 pp. 24–26, arrives at a similar analysis).

[9] I put "define" in scare quotes because, strictly speaking, you define *terms*, not phenomena themselves; you *characterise* phenomena. And furthermore, you don't define all terms. Some escape adequate definition, and, as we will see, delusion is one of these.

that this should make pathology enter into the very *concept* of delusion. Rather, it's perfectly possible that the concept of delusion can be analysed without reference to pathology, and yet the phenomena that fall under that concept resultantly have the kinds of properties that will always make them pathological. This conceptual independence is sometimes overlooked because pathological (or at least clinical) delusions are the paradigm cases, since (trivially and by definition) they come to the attention of clinicians.

However, there are sometimes cases of (what it seems natural to call) harmless delusion. For example, Leader (2011) writes about a patient of his who was being treated for anxiety. He later discovered that this man harboured an unrelated conviction that any man who shared his first name must also share a common quality with him. And yet he "took care not to broadcast his delusional thought, and it never caused him any problems " (Leader 2011, p. 13).[10] Whether or not one wants to call this a delusion (but I am intuitively drawn towards doing so) a non-pathological delusion does not seem to be a *contradiction*.

It is with this in mind that we look next at philosophically, rather than clinically, minded definitions of delusion.

2. The Standard Philosophical Approach

The aim of such a philosophical definition is not to help with clinical practice, but to provide a true definition of "DELUSION", an analysis of the concept, if you will. The aim of such a definition is typically thought to be that it should enable us to take any case, real or imaginary, and put it into the category of "delusional" or "non-delusional", and to do so for

[10] Other examples can be found in Hosty (1992), who speaks of "beneficial delusions", and Roberts (1991) who speaks of "preferred realities". A nice illustration of how clinicians get confused when pathology and need of treatment is built into the very concept of delusion, is found in Hosty (1992): "If the purpose of such a treatment […] is to relieve distress and to improve quality of life it could be argued that his psychosis is not a problem [i.e. doesn't need treatment] in that he enjoys his experiences and finds them life-enhancing" (p.373).

its own sake, not for the sake of clinical diagnosis. This tends to be attempted by proposing necessary and sufficient conditions. One can even draw a parallel with the concept of knowledge. Just as epistemologists have looked for the extra ingredient you need to add to true belief in order to get knowledge, so have recent theorists looked for the extra ingredient you need to add to false belief in order to get delusion.[11] As with analyses of knowledge, real or imaginary counterexample can be appealed to in criticism of analyses of delusion. We will look at potential attempts at doing this, before motivating an abandonment of the overall approach.

2.1. Delusions break clearly definable norms

Whatever we take delusions to be, one thing that seems fairly obvious is that they (and the subjects who have them) are breaking norms. In other words, delusion involves belief or belief-formation that is "bad" or has "gone wrong" in some way. When you say that something is a delusion, or that someone is delusional, you are performing a negative evaluation; you are claiming that they are somehow deviating from some standard or benchmark.[12] It has been standard philosophical (in particular epistemological) practice to isolate these norms or standards in objective, factual, theoretical terms. We will see that attempting to do this for delusion is grounded in a fundamental mistake.[13] However, in the interest of diagnostics, lets examine this approach. So, what kinds of norms might help us define delusion?

2.2. Deontic Norms and Norms of Biological Function

There are two kinds of norms that we need to put aside at the outset. These are *deontic norms* and *norms of biological function*. Deontic norms are norms that don't only set

[11] Except, as we've seen, delusions needn't be false.

[12] To put it another way, being non-delusional is normal – at least with regards to being delusional or not; there are obviously plenty of other ways of being abnormal.

[13] I suspect a similar mistake applies to analyses of knowledge, but that is a topic for a different thesis.

the standard for our evaluations of certain things: they entail some kind of obligation in the following sense. You don't just say that murder is wrong in the sense that it is merely defective (as you might for a belief that is false, or a heart that fails to pump blood); you expect that to have consequences for the way people behave. These kinds of norms, therefore, seem to apply only to actions, namely, to phenomena that an agent has voluntary control over. Regardless of whether one takes belief to be subject to voluntary control (we will touch on this issue in Chapter 4), it seems clear that delusions (or at least most of them) do not arise as the product of anything that is under the voluntary control of the subject. We should therefore put aside deontic norms altogether as being of little use for defining delusion. The delusional subject (at the very least in most paradigmatic cases) has not neglected any obligation of any sort. The delusional subject is more often than not a victim to be pitied, rather than a wrongdoer to be blamed.

Another norm to be put aside is that of proper *biological* function. Thus a heart "ought" to pump blood, and a heart that fails to pump blood, that strays from this norm, is functioning "badly" or malfunctioning (it is, in a sense, a bad heart). Biological function is usually cashed out in terms of that which contributes to the fitness of the organism (e.g. Millikan 1989). When a heart malfunctions, the organism as whole is somehow impaired (or potentially impaired). The same point can be made in terms of value: a well-functioning heart has biological value.

Some naturalistically minded philosophers (e.g. Papineau (2012)) have claimed that all norms can be reduced to biological norms. However, it seems intuitively implausible that the function of belief (construed as an epistemic state) is biological. In other words, it is hard to see how the norms that govern belief, what makes us call beliefs good or bad, are biological ones in any obvious sense. Indeed there are very well-known cases where a belief is epistemically defective but biologically valuable (e.g. cases where the cancer patient in denial goes on to live longer as a result of her self-deception (Dean and Surtees 1989)), and also cases where a belief is epistemically perfect, but biologically detrimental (e.g. having realistic estimations of self-worth may lead to depression (Taylor and Brown 1988)).Whether there is a way around this, it seems independently implausible that we can define delusion simply as *biologically malfunctioning belief*.[14] The notion of biological malfunction is slippery enough and in any case seems to operate at the wrong level. There are many cases we can think of where biologically well-functioning subjects are delusional. In cases of *folie a deux*, the secondary delusional subject often has nothing biologically wrong with them. Interestingly, similar problems arise if we try to tie illness to biological malfunction. Thousands of little biological malfunctions occur in my body all the time, but they don't interfere with my life, and therefore don't cause me to seek treatment or consider myself ill. And some illnesses, in particular mental illnesses, may be caused by a properly functioning response to a highly unusual environment (Graham 2013 holds this view for addiction).

Since delusion is a belief-like state, it seems appropriate to think of it as defined by norms that are proper to beliefs. These norms are called epistemic norms.

2.3. Epistemic Norms

We saw, for example, that Radden (2012) thinks that delusions are to be identified with "epistemic lapses". Can we characterise these epistemic lapses in terms of traditional epistemic norms?

An epistemic norm is a standard or benchmark that specifically concerns epistemic matters, matters concerning the phenomena of cognition and belief, and that will guide evaluations of these phenomena. These kinds of norms are commonly taken to be clearly definable by theoretical work. Indeed, one might even see the main task of epistemology as

[14] Perhaps we can be pluralistic about the norms of belief or the value that belief can have. We can say that there are ways in which beliefs can be *good* biologically (or, also, good psychologically), and independently, there are ways in which beliefs can be good *epistemically*. Whether we accept this pluralism or not, two things seems fairly clear. First, a belief is not anything like a heart. It is a very different kind of entity. Second, when we evaluate a belief *epistemically*, we are engaged in doing something rather different from what we are doing when we evaluate a heart biologically.

being one of mapping out the normative landscape for belief and cognition. They all ask some variety of the question: When and in what way is a given belief good?[15] We might call the norms that arise from answering this question in terms of clearly definable conditions, "traditional" epistemic norms.

2.3.1 Process-dependent vs. Process-independent evaluations

In epistemic evaluations we can first distinguish between an evaluation of a belief that is independent of the cognitive process that gave rise to it from an evaluation that explicitly takes that process into account. An example of the former kind of evaluation might be an evaluation for truth or falsity.[16] A belief is true or false in virtue of its content, not in virtue of the process by which it came about. Of course the process will have played a role in determining the content (if someone tells me that they'll be late for dinner, I am unlikely on *that* basis to believe that the moon is made of cheese), and, of course, some types of process are more truth-conducive than others (more likely to produce true beliefs), but even processes that are unlikely to produce true beliefs can do so by luck. So evaluating a belief for truth or falsity is what we might call a *process-independent evaluation*.

An example of a process-dependent evaluation might be rationality: a belief's being rational, one might plausibly think, is all about the process that gave rise to it. Thus the same belief (the same content, not the same state, obviously) can be rational if produced by one kind of process and irrational if produced by another kind of process. Evaluating a belief for rationality is what we might call a *process-dependent evaluation.* It should come as no surprise, therefore, that these two different kinds of evaluation can come apart: rational beliefs can be false, just as much as irrational beliefs can be true.[17]

[15] Indeed, in the narrow tradition of epistemology as theory of *knowledge*, the central question is typically: When is a belief *good enough to constitute knowledge*?

[16] Another example might be something like "plausibility."

[17] A less obvious question is whether the rational and the justified can come apart. This will depend on your epistemological leanings (e.g. externalism/internalism etc.).

A true belief is "better" than a false one, and there is an almost trivial sense in which what we believe "aims at the truth" (we will see much more on this in Chapter 4). It is in the nature of belief, as Velleman (2000, p.16) puts it, "of going right or wrong by being true or false." Although very much a traditional epistemic norm, it seems clear that truth will be of little use for characterising delusion, since there are clearly lots of false beliefs that aren't delusions, and, as we've mentioned, there may even be some delusions that happen to be true. This might suggest that part of what makes something a delusion involves the cognitive process that gives rise to it. In other words, it suggests that delusion attribution is a process-dependent evaluation. As Coltheart (2007) puts it, "couldn't a true belief be a delusion, as long as the believer had no good reason for holding the belief?"

In light of this, a better candidate norm than truth seems to be epistemic rationality (this is the kind of "epistemic lapse" that Radden has in mind). *Ceteris paribus*, a rational belief is better than an irrational belief. Might delusions simply be beliefs that are epistemically irrational or perhaps, more cautiously, irrational in a certain way, or to a certain extent? To answer this we must ask: what is it to be epistemically irrational?

2.3.2. Introducing Epistemic Rationality

Note, first, that rationality has a categorical and a normative sense (see de Sousa 2007, for nice expression of this). Thus we speak of humans being rational, in the categorical sense that they are capable of being rational, in the normative sense. As a first step towards understanding epistemic rationality, it is useful to compare it to practical rationality. The norms of *practical* rationality tell us when *action* is rational (or irrational), whereas the norms of *epistemic* rationality tell us when a *belief* is rational (or irrational). Perhaps the simplest way to think of rationality is in terms of a *telos* or *aim*. Practically rational action is action that maximizes our chances of fulfilling our own aims (these aims are often called "desires"). So what is epistemically rational belief? Whether *we* can aim at anything while believing is controversial (and we will examine this in Chapter 4), however, the idea that belief *itself* aims at truth is seen by many as highly plausible (see e.g. Williams 1973). I think therefore that a simple, but working, characterization of epistemically rational belief is belief-formation that has maximized its chances of achieving its goal, namely, *truth*.

Practical irrationality arises when, to put it in terms that echo William James, our nervous system, our urges and conditioned responses and evaluations, cease to be our allies and become our enemies.[18] Our short-term, short-sighted, brute motivational states conflict with our long-term deliberative judgement. Thus the life-long smoker smoking outside the cancer clinic is exhibiting practical irrationality: he doesn't want to be smoking because he knows that it is killing him and he doesn't want to die, but at a lower level, beyond his control, he can't help smoking since he can't over-ride the urge to do so.[19] The ability to be practically rational, to act in accordance with one's deliberative reasoning, certainly seems valuable. Practical irrationality can then be plausibly understood as a kind of agentive defect. Epistemic irrationality, by contrast, can be seen as a cognitive or epistemic defect. The goal of action is to fulfil (the right kind of) desire; the goal of belief is to be true. But a rational belief isn't simply a true belief, any more than rational action always fulfils desire. The point is that in forming a belief rationally, your belief-formation is doing the best it can. But sometimes the best doesn't attain truth, and sometimes poor belief-formation gets lucky.

How might we characterise epistemic rationality? We could start with the uncontroversial claim that a belief is epistemically irrational if it has no support, or, worse, is held in the face of contrary information. On the flipside of this we might say that a belief is rational if there is enough support in its favour. What kind of support might this be? (And, indeed, what counts as enough?) Common terms used to express epistemic support include "evidence", "reasons", "justification." These are terms of art, and are used to mean different things by different people. Now is not the time for a survey of contemporary epistemology,

[18] The particular quote I have in mind is, "The great thing, then, in all education, is to make our nervous system our ally instead of our enemy" (The Principle of Psychology).

[19] Although, Becker and Murphy (1988) have famously tried to accommodate the behaviour of addicts into their "Theory of Rational Addiction" which views addicts, within the micro-economic framework of rational choice theory, as rationally maximizing utility. This view has since been widely criticised.

but we can get some idea of how one might use these terms by looking at a well-known example from popular culture. The hope is that it will provide us with stable intuitions on which we can build.

In the film “The Matrix”, the experience of being an embodied agent living in, and interacting with, the world is perfectly replicated for individuals who are actually kept in motionless pods of fluid. There is an important sense in which people in The Matrix can be seen as perfectly rational. In fact, let us imagine a perfectly epistemically rational agent (call him Neo) who happens to be in the Matrix. Neo forms perfectly reasonable beliefs on the basis of what “people” in the Matrix tell him (testimony), on the basis of his perceptual experiences and so on. These beliefs are all (or mostly) false, but they are, in an important sense, rational.

Of course, a lot of epistemology deals with knowledge. Almost none of Neo’s beliefs count as knowledge because they are false (or formed through a deviant process). But once we put a concern for knowledge to one side, and are instead concerned with an evaluation of a cognitive system, e.g. a human central nervous system, there are many things that are commendable about Neo’s belief-formation process. Insofar as rationality is seen as your belief-formation processes doing the best they can to achieve truth, Neo is, by stipulation perfectly rational. He is just supremely epistemically unlucky.

This is one thing that epistemic rationality could plausibly be: it involves correctly using certain evidential inputs. Irrationality would then involve *incorrectly* using certain inputs (implementing biases), or forming beliefs on the basis of no evidential inputs at all. Can delusion be defined in terms of epistemic irrationality thus construed? This is far more promising than defining it in terms of falsity. It also fits nicely with our intuition that Neo is not delusional. However, it is crucial to see why this won’t work either.

In the next two sections we will see how delusion comes apart from irrationality. Not only is there non-delusional irrationality, but also there is rational delusion (or at least conceptual space for it). This suggests that being delusional is not conceptually tied to being irrational.

2.4. Non-delusional irrationality

Stating the obvious, there are instances of irrational beliefs that we would be unwilling to call delusions. As we mentioned, there are heuristic biases that stray from epistemic ideals, but that are adaptive short cuts most of the time. We may say that this is not ideal rationality, but rather "bounded rationality" (Gigerenzer 2002): there are temporal constraints, or conflicting norms (perhaps even various epistemic norms) that are weighed against one another (you might, for example, sacrifice accuracy for exhaustiveness or vice versa (cf. Proust 2012)). On a different note, as we have mentioned, there are various "healthy biases" in the normal population. These include, for example, the fact that it is considered statistically healthy to have an unrealistically positive self-image (i.e. "Positive Illusions", cf. Taylor 1989). People who are realistic, who approach what would be epistemic rationality in these domains, often suffer from clinical depression (cf. Alloy and Abramson 1979). Although we wouldn't, as I said earlier, want to tie the concept of delusion to that of pathology, our taxonomical intuitions would not want to call these "healthy biases" delusional, and they would certainly not want to call these *people* delusional. However, I don't think that it is because of "health" that we are reluctant to call these people delusional. The considerations that count when we attribute delusional status are not ones to do with health or pathology. We will later see what I think they are.

In any case, it should come as no surprise that there are cases of irrationality that aren't delusional. No one expects to be able to say that *any* degree of irrationality is a sufficient condition for delusion. But perhaps a certain degree of irrationality may suffice. Perhaps somebody needs to be irrational *enough* and then they will count as delusional.

However, I think we can quite easily show this to be false. Person A might be *more* epistemically *irrational* than Person B, but in fact turn out to be less *delusional*. Consider, for example (from Nozick 1993, cited in Murphy 2012), a mother whose son has been convicted of murder. We can understand that she will be *highly* resistant to evidence that suggests that he is guilty. We will not, however, call her delusional. People in these situations are believing in ways that are *epistemically* deeply irrational, but they are intuitively not delusional.[20] Why is this? I would suggest that it is because we can recognise their motivations, and we can recognise the influences that these can have on belief-formation and maintenance. This means that we find their epistemic irrationality unsurprising and *understandable*. We might even recognise (implicitly or explicitly) that in similar circumstances we would do similarly.

This marks a major change of approach. However, before moving on to the new approach, let's look at another reason why delusion cannot be defined in terms of epistemic rationality.

2.5. Rational delusion?

We saw that irrationality is not sufficient for delusional status, but it may not even be *necessary* for something to be a delusion either. In other words there may be cases of delusion that don't involve *any* irrationality in the sense we have just sketched. There are two very different kinds of grounds one might have for claiming this. One is on the basis of already existing (and in principle empirically testable) theories about how certain cases of delusion come about (these are the cases that Radden denies have true delusional status, because they are not "epistemic lapses"). The other is a conceptual argument that can be supported with thought experiments.

With scientific advancement, in particular in cognitive neuropsychology, we have moved beyond the observable behaviour of delusional subjects, especially in cases caused by

[20] Consider (from Bayne and Fernandez 2009) the active young man, who is diagnosed with terminal cancer, but who is in denial, refusing treatment and planning adventurous trips that go well beyond his life expectancy. This is more of a borderline case. An informal opinion poll I conducted through facebook, suggested that slightly more than half (32 out of 50) of the friends I asked do not think of this as delusional. In contrast, no one thought of the mother as delusional.

localised brain damage, to some understanding of what might underpin the formation of these delusions. What has constituted a major breakthrough is the increasing support for the view that these delusions are in fact formed on the basis of some kind of anomaly at the level of experiential input. To put it in more intuitive terms, if you or I were to experience what these patients experience, then we too would form the delusions that they form. As Brendan Maher, presciently put it, at a time before neuropsychological theories of delusions were available, "The delusional belief is not being held "in the face of evidence strong enough to destroy it," but is being held because evidence is strong enough to support it" (1974, p.99).[21] The point is that we can think of these delusions as arising from a misuse of normal input (what Maher calls "evidence") or we can think of them as arising from correct use of very bizarre input (misleading "evidence") (or, as we shall see in the next chapter, as a combination of the two).

As we will see in the next chapter, most philosophers and neuropsychologists in the field agree that many paradigm cases of delusion have, at least some, experiential grounds. In particular they take the nature of the experiential input to be instrumental in explaining why the delusion has a certain subject-matter: the feeling of unfamiliarity in the presence of loved one's might lead you to believe that they are an impostor, as in the Capgras delusion; a global lack of affective response to your environment may lead you to believe that you are dead, as in the Cotard delusion, and so on. The main source of contention is whether this experiential anomaly is strong enough (carries enough epistemic weight) to explain why the delusion is maintained for so long, or whether we need to postulate a bias of some kind. In the latter case, the delusional patient would be charged with epistemic irrationality.

However, whether or not there actually are biases at work is an empirical question, and as such is reserved for the next chapter where we look at competing ways of explaining

[21] This is perhaps a non-standard use of evidence in recent philosophy, which can construe evidence as being *de facto* truth or knowledge conducive. However, Maher's use of "evidence" is coherent, intuitive, and points to an important point. Namely, that when belief formation "goes wrong" it can be down either to the process or the input (or a combination of both).

delusion as an actually occurring phenomenon. In this chapter our aim is to ascertain, regardless of whether certain real-world delusional patients are epistemically irrational or not, whether, *if* there *were* people who believed these bizarre things on the basis of fully adequate private grounds, and hence are plausibly epistemically rational, we *would* still rightly consider them to be delusional. To put it another way, if Maher's theory happened to be correct, would these patients still count as delusional? Radden's answer is openly "No". Mine, however, is a resounding "Yes". Furthermore, it certainly wasn't Maher's intention to show that these patients that we previously had taken to be paradigm cases, were not, after all, really delusional. Rather, the question is: *granting* that they are delusional, how can we explain their delusional state? This is precisely the question we will address in the next chapter.

What if somebody who holds Radden's view, simply digs in her heels? What if she simply insists that we must tie delusion to irrationality? Since my response to this provides the motivation for my change of approach, I will now recap and then address this objection.

2.6. Recap and an Objection

To sum up, then, we have seen that the DSM definition is open to counterexamples, if it is seen as providing necessary and sufficient conditions. We have seen that psychopathologists tie the notion of delusion to mental illness. I have gestured towards why this is unattractive. In particular, we need to be more careful about the relationship between mental illness and delusion. A more epistemologically minded approach was then examined. Such an approach might claim that delusion is tied, in one way or another, to irrationality. There were problems for this view. We saw that there are cases of non-delusional irrationality. We then examined whether it could be a question of degree of irrationality. However, it seems that degree of irrationality does not always correlate with our propensity to call the subject delusional (Nozick's example of the mother whose son was accused of murder was a case in point). Furthermore, we also saw that cases where the delusion may be rational, for example, in the Capgras delusion, do not thereby cease to be delusional. Now suppose that someone insisted that, if one takes the Capgras delusion to be sufficiently supported by experiential evidence, as Maher claims (and as Reimer 2009 has recently defended), one *just is* thereby taking the delusion out of delusion. How does one adjudicate between these two positions? It is with this that I present a change of approach.

3. A Change of Approach

The change of approach I have in mind is as follows. We need to reflect on the kind of term that "delusion" is (or the kind of concept that DELUSION is), how it came about, the role that it plays, and, importantly, whether there is any good reason to think that it needs or deserves revision or regimentation, and if so, in light of what. I think that careful reflection on this yields the result that "delusion" is a folk evaluative term, and one that should not be revised.[22]

3.1. To revise or not to revise? Natural Kind, Theoretical and Folk concepts

One thing that is rarely discussed in the context of "defining delusion", is what we are aiming at with such a definition, and why. In other words, what are the desiderata of such a definition? What are the considerations that we take into account when we decide whether

[22] In two recent papers, Dominic Murphy (2012, 2013) has presented a view that closely resembles mine. However, it differs in two respects. First, he presents his view as an alternative approach based on looking at practices of belief-formation that are acceptable in other cultures (he draws on work done by the anthropologist Pascal Boyer), but doesn't argue for why this approach is preferable, and the current mainstream approach is misguided. I want to argue that the mainstream approach is misguided because there is no such natural kind or useful theoretical concept to do the regimenting. Second, he thinks that the attribution of delusion and mental illness are very closely linked ("Ed's [talking to trees is evidence that there is something mentally abnormal about him" 2012, p.22). I, on the other hand, think that it is important to see that our folk epistemic evaluations and our folk judgements of well-being take very different considerations into account.

we have got such a definition right? This can be presented in terms of “revisionism” versus “conservatism”. Revisionism takes the desideratum of a “definition” to be that it should pick out a concept of importance and thereby have the authority to revise and regiment current usage. Conservatism, on the other hand, takes the desideratum of a definition that it should accord with our current usage. Now, whether one should be conservative or revisionist will depend on the kind of concept or term that one is dealing with. What is interesting about “delusion” is that some theorists are revisionists in their definitions, and others are conservative.

Take two opposing “definitions” or “analyses” presented earlier, which will categorise phenomena very differently into “delusional” and “non-delusional”. Radden’s takes some phenomena that are paradigmatically delusional and claims that, once we understand their underlying nature, (namely, along the lines of Maher’s much earlier claim, that “the delusional belief is not being held “in the face of evidence strong enough to destroy it,” but is being held because evidence is strong enough to support it” (1974, p.99)) they aren’t really delusions. This is revisionist. It doesn’t aim to fit with our current usage of the word “delusion”. It aims to regiment and correct it, by identifying delusion with an epistemic lapse. As we’ve seen, Maher, on the other hand, seems to think that his theory about delusions leaves the delusional status as it is. Maher is, in this way, implicitly conservative. I am very explicitly conservative.

We should ask: For “delusion”, should we be revisionist of conservative? And does it matter? I’d say that it does matter, and that revisionism about delusion is mistaken. To answer this, it is useful to reflect on the kind of term that “delusion” is and to compare it to other, perhaps less elusive, terms. In particular, *revisionism* is motivated by what we might call *answerability*. We revise the usage of a term, if it is the kind of term that is answerable to something *beyond the usage*.

First, take *natural kind terms*. Their usage is answerable to how the world is, and revisable in the light of discoveries concerning this. Thus the discovery that water is H_2O

makes us revise our application conditions of the word “water” (see, most famously, Putnam 1975). If we came across something that looked, smelled, tasted etc. like water, but which we were reliably told did not have the chemical composition H_2O we would admit that it was not water (we might even call it “fool’s water”). We would be wrong to insist that by “water” we simply mean that which we deem to be water.

Now think of *theoretical terms* like “the equator.” Suppose we thought that the equator ran through the village of San Cristobal but then realized that we had miscalculated the location of the line that connects all the points that are equidistant from the north and south poles, so that it no longer ran through that village. We would revise our claim that San Cristobal is equatorial. Somebody who insisted that by “equatorial” he meant all of the places that we used to think the equator ran through would be exhibiting deviant usage. The application of fundamental theoretical terms, like natural kind terms, is answerable to how the world is, but in a slightly different way, namely, not to kinds, but to stable relationships between elements of the world, that are significant for various explanatory enterprises.

Now contrast these two kinds of terms with other terms that don’t have the same degree of answerability. These terms, I don’t want to say can be used in any way we like, but the application conditions are of a very different nature. Illustrative examples are the best way to proceed here. Consider terms like “indignation”. It is well known that there is major cross-cultural and linguistic variation in the application conditions of words that express concepts of this nature. Setting the application conditions for these terms isn’t answerable to anything in the world *beyond usage in the linguistic community.* Call these *folk terms*.[23] The application of the term is correct to the extent that it is aligned with how it is actually used by the linguistic community. Deviant usage involves failing to use the term as the linguistic community does. But unlike theoretical or natural kind terms, where this *does* happen, it makes no sense to say that the usage of the entire linguistic community is mistaken.

[23] There is some ambiguity here. On an understandable use of “folk term”, “water” is a folk term. In the sense I’m intending here, “folk” is not to be equated with “non-scientific”. “Water” is a natural-kind term, but merely using non-scientific vocabulary. It is not a folk term in my sense.

Here conservatism is warranted, but it is needn't be lax or unscientific. Indeed, how a linguistic community uses a term can be usefully treated as an empirical fact, worthy of careful scrutiny. Certain branches of linguistic (notably anthrolopological or descriptive linguistics) investigate precisely this. Vainik (2002), for example, does this for Estonian emotion terms. This is taken to "plot" the Estonian "folk model for emotion", which is strikingly different from English.

3.2. Describing (fact-stating) vs. Evaluating

Although the first two kinds of term (namely, natural kind and theoretical terms) are "answerable" in a way that the third kind (namely, folk terms) is not, and therefore are subject to revisionism in a way that the third is not, all three are "descriptive" in that they attempt to pick out properties or kinds in the world, and say fact-stating things about the world. "Indignation" is folk-*descriptive* in this sense. Consider the exchange:

- "You're indignant."
- "No, I'm not."

Both of these claims are fact-stating. Namely, calling someone indignant is taken to be determinately true or false.[24] What determines the meaning of the word "indignation" is how it is used in various contexts. The person who denies that she is indignant is taking herself to fail to exemplify such a context. Now think of some terms that are not descriptive in this sense but evaluative: terms like "good" and "beautiful". These, we might say are *folk-evaluative* concepts.[25] They express a certain kind of evaluation. The kind of evaluation that they express, their linguistic meaning, if you will, is determined by the linguistic community, but it is not taken to have the same kind of factual power. Obviously, theorists can describe norms factually (e.g. biological malfunction, epistemic warrant etc.), and these norms will

[24] We must not confuse vagueness and factual indeterminacy. Vagueness is factually determinate. "Child" is a vague concept, but is used in making factually determinate statements. An 8 year-old is a child. A 21 year-old is not a child. Is a 14 year-old, a 16 year-old?

[25] Note that there are also non-folk (viz. scientific or theoretically sophisticated) evaluative concepts like biological malfunction (or, perhaps, "warranted" in a reliabilist framework).

give rise to evaluations in a certain, more objective, sense. Thus, "traditional" epistemology will attempt to *describe* the factual conditions that need to obtain in order for a belief to be good. Philosophy of biology can do something similar for biological malfunction. This is not what I mean by evaluative terms. I mean the terms that ordinary language-users typically use when they are *in the psychological state of evaluating rather than describing something*.

Thus we can construct the kinds of terms outlined here in a table, as follows:

	Descriptive	Objectively Evaluative	Fully Evaluative
Answerable Terms	Natural kind and theoretical terms ("Water", "Equator")	Biological norms (proper function), design norms ("broken")	
Folk Terms	e.g. Emotion terms ("Indignation")		Value-judgment terms ("Good", "Beautiful")

3.3. What kind of term is "delusion"?

Now, where would we want to locate "delusion" in this table? A revisionist theory wants to put it in either the answerable/descriptive box, or the answerable/objectively evaluative box. "Delusion" picks out either a natural kind (Samuels (2009) explicitly thinks that this is the case) or something of fundamental theoretical importance (which can either simply describe or present an objective benchmark from which a deviation generates an objective negative evaluation). On such a view, it may turn out that many of the things that we think are delusions are not really delusions.

I want to propose the following: that delusion is a folk-evaluative term. Delusion-attribution is folk-determined and non-fact-stating. A delusional assertion or belief is an assertion or belief that is "bad" in the sense that it leads to a certain negative appraisal on the part of the "folk". Such an exchange does not seem out of place, in either a daily or clinical context:

- "You're delusional."
- "No, I'm not."

This seems to have the same form as the anger case. However, doesn't one get the intuition that the disagreement is not factual, but rather evaluative? The second speaker seems to be questioning the validity of an evaluation, rather than claiming that that first speaker has mis-described the world.

So far, I have merely stated that delusion is a folk-evaluative term. However, here are some explicit reasons for thinking that it is:

> - Natural kind terms refer to things with a unified underlying nature. There is no reason (*pace* Samuels) to think that delusions do (recall the Jones (1999) quote from earlier). Or that it would be a good thing if they did.
> - Does delusion play a specific theoretical role? Is it clearly definable in terms of other theoretical terms/relationships? No.
> - Does the fact of something's being delusional actually do any work? If I dispute that something is water, or is equatorial I am disputing a fact. What if I dispute whether something is delusional? In fact, theorists do dispute whether a given case is delusional. However, when they do, this is not disputing the facts. When theoretical dispute about these phenomena is substantive, it is put in more basic, objective, theoretically useful, terms. For example, it centres on what mechanisms are implicated, what is experienced by the subject, if there are biases at work or not. But the attribution of delusional status is never a firm base on which to found substantive, factual, disagreement.

Radden and Maher could (and perhaps do) both agree about the aetiology of the Capgras delusion, about what is going on inside the delusional subject. And yet they would disagree about whether the subject is delusional. Would either of them, on this basis, deny that the Capgras patient is in need of treatment? Clearly not.

There is lots of interesting work (both empirical and philosophical) to be done with specific delusional phenomena, e.g. with the Capgras delusion, but there is no reason to think that the concept of delusion itself is doing any important work (or that a carefully honed concept of delusion *would* do any important work). Calling someone delusional clearly has a rule-of-thumb clinical utility. It also works rather well as an insult ("You're delusional if you think the lunar landings were faked!"). So, what I think is worth doing, rather than providing a strict definition of delusion, is a *characterisation* of the kinds of things that we call delusional.[26] This will be picking out a kind of "family resemblance", rather then delineating necessary and sufficient conditions. I will put this in terms of a "folk epistemic norm" I call "understandability".

4. The "Understandability" Criterion

Recall the cases of severe epistemic irrationality, but that we would be reluctant to call delusional. One example we used was that of the mother whose son has been convicted of murder, and who refuses to believe in his guilt. These cases tell us something very important.

Regardless of how we cash out the details of a theoretically sophisticated notion of rationality, our intuitions about who is and who isn't delusional are not explained by appeal to irrationality, but rather by appeal to our folk-reactions towards something we might call "understandability". This understandability involves what I briefly mentioned earlier, namely, understanding the motivations of the subject, given the context, and understanding that the epistemic irrationality that this will give rise to is understandable. Recognising that many people, including oneself, would do similarly if they found themselves in relevantly similar circumstances. Henceforth, I will call this "understandability." Indeed, self-deception and wishful thinking, insofar as they are cases of motivationally influenced belief-formation,

[26] In a sense, I think that this is what the DSM is trying to do, or ought to be trying to do. However, what I propose is less rigid, and hence not subject to counter-examples in the same way.

are "understandable" in the relevant sense.[27] One interesting upshot of this "understandability" criterion is as follows.

Many theorists find it plausible that self-deceptive phenomena overlap with delusional phenomena (see, for example, Bayne and Fernandez 2009, Radden 2012 etc.), perhaps with only serious or tenacious cases of self-deception counting as delusional. Thus, the denial of a cancer diagnosis may be a delusional *state* that arises out of motivationally driven *process* of self-deception. This overlap, or gradation, is a clear consequence of tying delusion to epistemic irrationality: if self-deception strays *far enough*, the belief it gives rise to will count as delusional.

This does not seem *prima facie* implausible. However, contrast this with the position that ties delusion to understandability. This view yields the interesting consequence that factors making a state a delusion and factors that make a process self-deceptive don't pull in the same, but in *opposing* directions. The motivational influence on the belief, which makes something self-deceptive, can also serve to make it understandable. This is not to say that there are *no* cases that are both self-deceptive and delusional. But rather, that such cases aren't delusional because of the degree of epistemic irrationality. They are delusional because the motivational influences fail to render the epistemic irrationality understandable. This could be either because the motivational influence is perplexing (as we will see, this applies to Joseph Capgras's original account of the delusion that bears his name) or because the degree of irrationality is too extreme to be rendered understandable by the motivational strength.

However, since these understandability intuitions do not have carefully thought-out theoretical criteria, we sometimes operate double standards when attributing motivations to the subject, namely, when attributing what is "desirable" for the subject. Thus, the most common cases of self-deception that are labelled as delusional, are cases where someone is

[27] This is assuming they do not occur in the context of brain damage or psychosis (like some cases of "motivated delusions" (cf. Bortolotti and Mameli 2012)).

in denial about something harmful to them, e.g., a cancer diagnosis (e.g. Bayne and Fernandez 2009). It is not the denial in itself, or its strength or degree of epistemic irrationality, that makes it be seen as delusional. The strong cases of denial we get in cases of sexual abuse (by both victims and perpetrators) (cf. Leander 2010) are not called delusional. The reason why denial of cancer is called delusional is because, when it persists the subject refuses treatment, and dies much sooner as a result. That makes us call them delusional because we see it as "desirable" for the subject, to seek treatment. We therefore readily switch perspective from what the subject actually desires (as in the case of the murderer's mother), to what is desirable for the subject (or what the subject ought to desire), namely, medical treatment. But this is not to do with epistemic irrationality. The mother whose son is guilty of murder can keep up *extreme* resistance to the evidence, and we may still not count her as delusional. We understand the motivation and psychological benefit of her believing in her son's innocence. Indeed, we may even see something wrong in a willingness to accept a son's guilt that matches the evidence in a way that might be describe as epistemically rational. We might say, "Gosh, that's not very mother-like of her!"

5. Some illustrative cases

One thing that should be obvious is that delusions come in degrees, and as a result the boundaries between the delusional and the non-delusional will be vague. But we can use non-borderline cases to see the sorts of things that infringe our understandability intuitions.

The Capgras Delusion

I will start by looking at the best documented of the monothematic delusions. It is also the one that we will be dealing with the most in this thesis.

What makes the Capgras delusion a case of *delusional* misidentification rather than straightforward misidentification? If Jennifer Radden is correct, then the Capgras delusion merely looks like delusional misidentification. I don't think that this is an attractive option, in part, as we saw, because it is revisionist, but also in part because there is a clear distinction between delusional misidentification and a misidentification that arises from a straightforward mistake. Neither, on Maher's aetiology of the Capgras delusion, are epistemic lapses, but we can separate them easily by other means. Our folk epistemology, our everyday evaluations of beliefs and their formation, i) have certain expectations about how people identify individuals and ii) make allowances for mistakes. Regarding i), you normally judge that people are a certain person, either because you recognise their appearance (or voice etc.), or because you have tracked them spatiotemporally (you might know that an old friend is working the morning shift in a local bar, but has changed beyond recognition). The Capgras delusion clearly defies this expectation. The Capgras patient may have a private epistemic reason (we will see more on this in the next two chapter), but doesn't have a publicly accessible reason, a reason that can be given to a third party. We know that the Capgras patient can recognise people, can even recognise that, for example, the misidentified person *looks exactly* like the person in question, but still judges that it's not them. Our expectations, our folk norms, are grievously upset. Whether the Capgras patient is "really" rational (as Maher would have it) is not the point: we brand them delusional.

It is also illustrative to compare the Capgras delusion with another delusional disorder: intermetamorphosis. Both the Capgras patient and the patient suffering from intermetamorphosis misidentify people, and both are classified by the psychiatric profession as delusional misidentification disorders. However, whereas the Capgras patient misidentifies people in spite of being aware that the person looks like the person they are misidentifying, the patient with intermetamorphosis experiences a change in appearance. In fact, the patient with intermetamorphosis forms her judgement on the basis of the experienced appearance, so, unlike the Capgras patient there is a match between who somebody *looks like* to the patient and who they *actually are*. This means that what is most perplexing (and hence most "delusional") element in these two disorders is located at a very different part of epistemic space. In spite of the fact that the content of the explicit delusional utterance can be the same ("This person [gesturing to her mother] is not my mother"), the differences in their route to making this claim betray the implicit epistemic commitments of the subjects. The most perplexing thing about the Capgras patient is the implicit acceptance that people can look exactly like a given person and yet somehow fail to be them (the patient with intermetamorphosis clearly isn't committed to this). On the other hand, the most perplexing thing about the patient with intermetamorphosis is the implicit acceptance that people can spontaneously change both appearance and identity. Both are delusional disorders because we don't expect people to form beliefs that imply these possibilities.

To sum up, regardless of our aetiological leanings, and, in particular, whether we take the Capgras delusion to be rationally grounded in bizarre experience or not, it is still a paradigmatic case of delusion, namely, a belief that affronts our folk epistemic norms, and therefore lack understandability.

Folie à deux

Having started with a case of delusion that is paradigmatically a delusion on my view, I will now examine a more problematic case, a case that is problematic on most definitions of delusion.

In folie à deux, a "primary" patient transmits a delusional belief to one or more "secondary" patients (so you also get "folie à trois", and "folie à famille") who then adopts the belief (Langdon 2013). What is difficult for standard theories of delusion is the status of the secondary. The primary tends to suffer from a psychopathology, like schizophrenia, whereas the secondary, although personally very close to the primary, and often in a position of subordination (e.g. a son or daughter, or younger sibling), and sometimes of below average IQ, is very often perfectly "healthy" by medical standards.

This is especially difficult for those who tie the notion of delusion too closely to pathology. Furthermore, since the secondary often abandons the belief very easily after separation from the primary (Langdon 2013) this may seem problematic for any view that ties delusion too closely to "tenacity". One option for these sharply definitional views is simply to bite the bullet and say that the secondary is simply supremely gullible, but not delusional.

However, if we look at some of the extreme cases, this becomes less and less palatable as a response. The challenge is: how do you separate gullibility from delusion? The understandability criterion responds to this in an obvious way. If the gullibility is sufficiently extreme to engender a high degree of perplexity, to heavily disrupt our folk expectations (as these extreme cases do), then they are delusional. But of course, this is a matter of degree. And this is highly unsurprising, since "delusional" is an evaluative term like "bad" or "ugly", and evaluations come in degrees ("better", "worse", "more delusional", "less delusional").

Religious belief and religious delusions

Contrary to Richard Dawkins's high-profile view, religious belief is not in and of itself, delusional. We accept that people are capable of believing things that we think are false or abhorrent, and our folk epistemology accommodates that. Dawkins's labelling of religious belief as delusional is more like the non-literal use we saw at the start of the Introduction. It is an expression of *disapproval* of religious belief. However, whether we are friends or foes of religion, we have expectations that people will have religious beliefs of a certain nature, in certain contexts (N.B. Although we have then now, we may not have these expectations forever). However, there are some beliefs, with religious content, that do infringe these expectations. These are typically called religious delusions, and actually have a high prevalence among schizophrenics (Siddle et al. 2002, found that 45 out of 193 examined cases of delusional schizophrenic patients, had delusions with religious content.) What distinguishes them from standard religious beliefs is that the subject takes himself to be special, namely, to be a deity, or to have a special relationship with the deity.

Similarly to folie à deux, there is an interesting question about what we are to say of the cases where the person with the delusional disorder, convinces others and forms a cult. I think Reznek (2010) is right to say that it depends on *how* the followers are acquired. This can have varying degrees of understandability. However, I disagree with him that it depends on whether "the thinking of the leader is delusional" (Reznek 2010, p.112). He cites Jesus as an example of a leader who was probably not delusional. He may be right, but I don't see how that distant historical fact has an impact on whether today's Christians are delusional or not. It seems to me clear that Christianity is something that does not (yet) infringe our folk epistemic norms.

Delirium: Duration and Tenacity

It is widely claimed that delusions are strongly held in the face of counter-evidence. What are we to make of short-lived cases of delusion-like states? These cases are often called "delirium", and occur in various contexts, from high temperatures to dementia. The boundary between delirium and delusion is not clearly defined, and some theorists (e.g. Broome 2004, Silva et al. 1997), talk of delusional misidentification occurring in the context of delirium, or a delirious episode. What are we to make of this?

First, I think it is important to distinguish the duration of a belief from its "tenacity", by which I mean its sensitivity to counter-evidence. I could hold a preposterous belief for a very long time, but only because no one ever corrected me. Suppose that if someone *had* corrected me, I would have been open and grateful to such correction. It seems that my epistemic state, so described, in an important sense does not infringe folk epistemic norms to the extent that we would want to call this belief delusional. This belief happens to be long lasting, but it is not tenacious. In delirium what we have is the opposite: a confused state where a false belief (say, a delusional misidentification, or a delusional intermetamorphosis) is tenaciously held, but only as long as the delirious episode lasts. We get cases like these in the context of extremely high fevers (Silva et al. 1997), or induced by drugs (such as ketamine, e.g. Corlett et al. 2010), or even during a migraine (Bhatia 1990). I would suggest that our folk epistemological norms also accept that people momentarily under the influence of drugs, or fevers, will believe very strange things. We may call them momentarily delusional, or simply delirious. Whether we call these cases delusional or not, it is clear that the subject, first, does not infringe our epistemic norms as much as one whose belief has both tenacity and duration, and second, does not require the same kind of attention as such a subject (e.g. a schizophrenic, or brain-damaged Capgras patient). In the next two chapters we will see potential explanations of this tenacity.

Conclusion

The concept of delusion is a folk concept that should not be regimented by a strict definition. Such regimentation is unmotivated, since the concept itself does no important scientific work, except insofar as it is a useful clinical diagnostic tool.

Why, you might ask, if I think this about the concept of delusion, am I writing this thesis? And why should you go on reading it? It is because several of the things that we call delusions are very interesting in their own right. For example, some of these may tell us, as exemplars of defective human cognition, how healthy human cognition works. Schizophrenia research has already shown promising signs of illuminating the role of neurotransmitters (e.g. Kapur 2004, Corlett et al. 2010), and the Capgras delusion, as we will see in Chapter 3, may tell us some interesting and surprising things about how we track individuals (namely, judge that a given individual is that individual). Most importantly for the overarching aims of this thesis, some of the phenomena that we call delusions (most certainly not all), cause problems for mainstream philosophical views about the nature of belief.

In the next chapter, we look at how we might go about explaining some of the phenomena that arouse in us the negative reaction that leads to them being called delusions.

CHAPTER 2

Explaining Delusions

Introduction

In the previous chapter we looked at the concept of delusion, independently of the kinds of delusions that actually exist in the world.[1] In this chapter we will be more closely concerned with actual cases of delusion, and, in particular, with how to explain them. We will start by looking at the constraints and desiderata that might apply to explanations of delusions, and then look at different attempts that have been made for explaining delusions caused by brain damage, and, in particular, the Capgras delusion. This fits in with the rest of thesis by showing, first, how we can make sense of something physical (namely, brain damage) affecting (more or less directly) something paradigmatically mental. And second, by showing the kind of explanation that belief both takes part in, and is itself subject to.

I will start by looking at the nature of explanation generally, and in psychology in particular. I will then draw attention to two different constraints. One constraint concerns "levels" of explanation, whereas the other concerns "kinds" of explanation. Personal (or "personal-level") explanations are importantly different in kind from subpersonal explanations, and cannot be provided in subpersonal terms. However, this is not to say that subpersonal psychology cannot inform personal explanations, and we will see the ways in which it can do so. Then I will look at different approaches that have been taken towards explaining delusions generally, and look at competing aetiologies for the Capgras delusion. I will end by showing how different aetiologies have different consequences for the availability of personal explanation.

[1] In other words, we could imagine or describe a hypothetical case (or a hypothetical subject), which doesn't, or perhaps even *couldn't*, exist and yet we'd have a good idea as to whether it would count as delusional. In particular, we saw that it would count as delusional if it infringed folk epistemic norms to the point of lacking "understandability".

1. Explanation: Levels and Kinds

To what extent can the cognitive sciences (neuroscience, neurobiology, cognitive psychology etc.) help to provide us with an explanation of delusions? It is clear that they can make major contributions, but what is the nature of, and the constraints governing, such contributions? My aim in this first section is to draw attention to two distinct constraints that are present in answering this question. The first constraint concerns *levels* of explanation. How do we connect explanations, usually mechanistic explanations, at different levels, namely the different levels provided by the disciplines of the cognitive sciences? The second constraint concerns *kinds* of explanation. How do we tie two very different kinds of explanation? These two kinds of explanation are sometimes called "personal level" and "subpersonal level" explanations (e.g. Dennett 1969, Davies 2000).[2] However, the presence of the word "level" here is unfortunate, since it obscures the need to keep the two constraints separate. In my view it is vitally important to not view the "personal level" as just another "level" of explanation. It provides an altogether different kind of explanation. Indeed, many theorists (exemplified and unfortunately disseminated to beginners in philosophy of psychology in the go-to manual on the topic (Bermudez 2005)) take the two challenges to be roughly the same.[3] They take the problem of connecting, say, a neurobiological story to a cognitive story to be the same kind of problem as that of connecting a subpersonal cognitive story to a personal story. It is not. And the aim of this opening section is to show why not. I will therefore drop the word "level" and use simply "personal explanation" and "subpersonal explanation".

[2] Other terminology used is "rational" vs. "mechanistic" explanation (Davidson 1963, McDowell 1985) and "teleological" vs. "causal" (Smith 1987) explanation.

[3] See, in particular Bermudez's insistence (2005, p.32) that personal explanation (what he calls "folk psychological" explanation) is simply one kind of "horizontal explanation", namely, explanation of an event in terms of causal antecedents, rather than constituent parts (which is "vertical explanation").

An important preliminary to introducing this difference, is to reflect briefly on what we mean by explanation generally.

1.1. Formal-logical, Ontological and Pragmatic Views of Explanation

As Achinstein (1983), among others, has noted, explanation is subject to a "process-product" ambiguity, which is to say that it can refer either to an act of explaining, or the product of such an act. What we are interested in here, like most theories of explanation, is the product. However, as we will see, different theorists view the importance of the act of explanation to our understanding of the product differently.

Differing answers to the following two questions yield different views about the nature of explanation (*qua* product). These two questions are:

1) What kinds of things are the relata in explanations? (viz. When we say that x explains y, what kinds of things are the values for x and y?)

2) What is it for x to successfully explain y?

Following Faye (2007), I think it is useful to distinguish between three kinds of views of explanation, namely: Formal-logical, Ontological and Pragmatic views of explanation. My aim is not to adjudicate between these in any general sense, but rather to show that one of these provides an especially helpful way of approaching the issues in this chapter.

1.1.1. The Formal-logical View

On the formal-logical view, first and famously put forward by Hempel (1948), an explanation is an abstract entity; in particular, it is a logically valid argument with propositional structure. Indeed, an *explanandum*, according to Hempel, is a proposition that follows *deductively* from an *explanans*. Thus, in answer to the two questions above, one *proposition* explains another, and does so by standing in a certain logical relation to it. A number of things should be noted about this approach.

i) Scientific and ordinary (everyday) explanations are profoundly different in nature. The things we call "explanations" in daily life never, or at best rarely, pick out logically related propositions.

ii) This characterisation is prescriptive rather than descriptive. It is neither interested in capturing how we use the word "explain", nor in capturing what scientists are actually engaged in doing when they explain things.[4] It aims to tell us what something *ought* to be if it is to count as an explanation in this refined, ideal, sense. One might alternatively put this in evaluative rather than constitutive terms and say that explanations are good explanations to the extent that they approximate this ideal.

iii) Explanations are objectively "out there" to be discovered.

The formal-logical view of explanation includes a number of views of explanation besides Hempel's original covering-law version. For example, it includes Salmon's statistical-relevance model as well as the unificationist theory of scientific explanation as elaborated by Friedman (1974) and Kitcher (1989).

The formal-logical view is not very helpful for our purposes. Practically speaking, it is too demanding to usefully apply to psychology, and it doesn't reflect what psychologists actually do or ought to do.

1.1.2. The Ontological View

On the ontological view, explanations aren't made up of logically related propositions. They are made of concrete entities like, for example, objects, states of affairs or events. For example, you might think that events explain other events. In particular, it is common within this approach, to think of causes explaining their effects. An instance of fire explains an instance of smoke.

[4] As Hempel (1965, p.413, quoted in Faye 2007) himself said: "Explicating the concept of scientific explanation is not the same thing as writing an entry on the word 'explain' for the *Oxford English Dictionary*."

So, in answer to questions 1 and 2 above, we get:

Events (or states of affairs) explain other events (or states of affairs), and they do so by standing in predicable law-like causal relations.

Like the formal-logical view, a number of things should be noted about this view:

i) Again, scientific and ordinary (everyday) explanations are different in nature.

ii) Again, explanations are out there to be unearthed. You discover them. You find a particular event, and you unearth the explanation of that event, namely, its cause or causes.

One recent theorist, who buys into this account in philosophy of science generally, is James Woodward (2003). Another, who applies a related view specifically to psychological explanation, is Donald Davidson. To simplify somewhat, Davidson (1970) takes events to cause other events. He also takes explanation to require the picking out of a cause (which is an event) to explain an effect (which is also an event). However, events are only explanatory "under a certain description". He is sensitive to the fact that picking out events that are causally related is not sufficient to be explanatory: you have the pick them out in a *causally relevant* way. For example, to explain why the scales go down when weighing some plums, you appeal to the weight of the plums, not their colour, even though those are two aspects of one and the same event (namely, the putting of plums on the scales). This has some affinities with the pragmatic view. However we will see that, crucially, the pragmatic view opens up the possibility of non-causal explanation.

1.1.3. The Pragmatic View

According to a pragmatic view of explanation, an explanation is a good answer (and, we shall see, a variety of factors, both psychological and objective may contribute to this "goodness") to an explanation-demanding question. An explanation is the product of a communicative act, and the relata of explanations are not events, nor are they propositions; they are speech acts that are heavily dependent on a number of contextual factors. These contextual factors include a number of issues (for example, conversational context, the epistemic state of the demander of the explanation etc.) but the most important for our purposes are the explanatory concerns of the demander of the explanation (which I will henceforth call "the consumer"). An explanation has to address the explanatory concerns of the consumer, and has to be (at least a candidate for being) considered satisfying.[5] This potential subjective satisfaction is a necessary but not a sufficient condition of something being a good explanation. Obviously, there are many objectively bad explanations that we may wrongly consider satisfying ("just-so stories" and "old wives' tales"). So they have to be satisfying in a non-illusory way. But conversely, an explanation that is objectively good by the standards of either the ontological or formal-logical view, but which leaves the consumer completely in the dark, is not considered a good explanation. Explanations are relative to a particular instance of a question being asked, and have to cater to the consumer's epistemic state. The consumer, it must be noted, isn't necessarily an individual, but could be a collective. The "question" could be asked implicitly by the scientific community as a whole (or a subset of that community), or explicitly by an individual.[6]

There are a number of variations on the pragmatic view. The *locus classicus* is Van Fraassen (1980). Achinstein (1983) has an attractive version that relies heavily on the tenets of ordinary language philosophy, and Faye (2007) puts forward his own refinements. Although it is not worth going into the details of these variations, it is worth enumerating what all versions of the pragmatic view have in common, in particular, in contrast to the formal-logical and the ontological views characterised above.

[5] The parenthetical "at least a candidate for being" is to account for cases where an appropriate explanation is provided, but is rejected on grounds of implausibility. I'm assuming, for theoretical purposes, that the demander of the explanation is maximally receptive and has no competing information of her own.

[6] The pragmatic view can easily explain away the following apparent intuitive support for the ontological view of explanation. We often use the word "explain" to talk about events "explaining" other events. We say, for example, that the presence of fire explains the presence of smoke. But for the pragmatic view, this is just a manner of speaking. In actual fact, a *person* explains the presence of smoke by *telling you* about the presence of fire. Or alternatively, the fire explains the presence of smoke, if you see it for yourself, by "telling" you of its own presence. Implicit in this being explanatory for you is that you have an understanding of how fire relates to smoke.

i) Scientific and ordinary explanation are essentially the same. The former simply has a more regimented context (viz. the explanatory concerns are regimented and shared across a community, namely the scientific community).

ii) Explanations, being the products of communicative acts, aren't unearthed, but carefully expressed. They are answerable to how things stand in the world, but they need to be selective and comprehensible to the consumer of the explanation. Explanations, *in the relevant sense*, simply do not exist in a possible world devoid of inquiring beings that demand and give explanations. Furthermore, these explanations are demanded within a wider pragmatic context, whether it is everyday life, the court of law, the lab, or the clinic.

A pragmatic view draws attention to the fact that explanations cannot be requested from a background state of total ignorance. The epistemic or informational state of the demander of the explanation will in part determine her explanatory concerns, and her explanatory concerns will dictate the kind of explanation that would be satisfying.

A broadly pragmatic view of explanation is what I'll be assuming for the remainder of this chapter. On a pragmatic view, one event can arouse different explanatory concerns, each of which demand different explanations. Suppose there is a plane crash. If one is conducting a legal investigation, one will, for example, ask for an explanation of the plane crash in terms of poor decision-making. If one is an aeronautical engineer one may ask for an explanations in terms of the weather conditions, and how they impacted on the flight of the aircraft. Depending on certain facts about the crash, either of these explanations may be unavailable. However, this does not involve forcing somebody who is asking the former question to accept an answer that is suitable to the latter question. Rather, it involves getting the person to see that they need to ask a *different question* because their particular question, in the circumstances, has no answer. For example, suppose the weather conditions were so severe and unpredictable that nobody is to be held accountable for the plane crash. Somebody looking to hold someone accountable for the plane crash is, given the situation, asking the wrong question. Things like this happen in countless fields of enquiry. We will see that this will turn out to be extremely relevant to delusions.

1.1.4. The Pragmatic View and Explanation in Psychology

A lot of issues concerning explanation in psychology become simpler when we adopt a pragmatic view of explanation. In psychology, broadly construed, there are so many different questions to be asked about the same subject matter (roughly, to use Jane Roland Martin's (2000) term, the same "chunk of reality"). Questions can differ for different reasons. It could be because we are looking at the phenomena at a different fineness of grain. Or we can ask different *kinds* of questions because we have a different *kind* of explanatory concern, for example, one that we will see crops up, in particular, when we are dealing with persons or agents and we want to know why someone has acted the way they have, or believes the things they do.

Since on a pragmatic view an explanation is a good answer to an explanation-demanding question (it is satisfying in a non-illusory manner to the demander of the explanation), this opens up at least as many explanations of a given phenomenon as there are explanatory concerns, and as many *kinds* of explanations as there are *kinds* of concerns. Some concerns in psychology are about understanding a causal system. Other concerns are about understanding people; why they do things, why they believe things, and sometimes with a view to evaluating them or attributing blame to them. Different sorts of answers will address these sorts of concerns. As I said before, you can't force someone who wants an explanation of a plane crash in terms of poor decisions to accept an explanation in terms of the weather. What you can do, however, is try to redress their concerns. This will be especially important when the type of explanation being asked for is not available.

Within the context of this chapter, what a pragmatic view of explanation makes room for is the claim that "levels" of explanation in cognitive science are about *causal* explanatory concerns. There are many different such concerns, but they are, at heart, the same kind of concern. In addition to this, however, there are important explanatory concerns that aren't causal in the same sense. Addressing these concerns produces not another level of explanation, but a different *kind* of explanation.

With the pragmatic view of explanation in the back of our minds, let's characterise "levels of explanation" and see how it differs from the "kinds of explanation" I think it is so important not to overlook.

1.2. Levels of Explanation: Causes and Mechanisms

Within explanations of a certain kind, which have causal and predictive explanatory concerns, one can distinguish between causal explanation and mechanistic explanation. Roughly, in asking for a causal explanation one is asking for a cause, without inquiring into the mechanism whereby that cause functions. This provides some degree of understanding, and a sufficient degree of understanding for some situations, for example, if one wants to avoid a particular effect. Thus we might establish that smoking causes cancer. We might not know exactly how it does so, but knowing that, at least, is enough to suggest that (*ceteris paribus*), if we don't want cancer, we shouldn't smoke. We can think of causal explanations as answering a certain kind of question, namely, a "How come?" question. "How come he got cancer?" This is often expressed with a causal use of "Why", as in "Why did he get cancer?" This causal use of "why" is very different from a justificatory use that we are about to encounter.

A mechanistic explanation answers instead a "How?" question. It provides not just the cause, but the mechanism whereby a certain causal process operates. Using the cancer example, it isn't enough to know that smoking causes cancer: what is required is a description of the mechanism, for example, in terms of carcinogenic disruption of genetic material through radiation. Mechanistic explanation provides a greater degree of understanding than merely citing causes, and it is of central importance in various areas of science, e.g. in biology, cognitive psychology, and neuroscience. I take mechanistic explanations to be a sophisticated brand of causal explanations, since the kinds of concerns addressed are of the same kind; namely, of understanding how a brute causal system will behave in relevant counterfactual circumstances.

The quality of an explanation in this sense increases as a function of the degree of causal understanding it conveys. We can think of this understanding in different ways, but one useful way is in terms of ruling out the relevant possibilities. We can see this by imagining a subject with a "full" understanding of a particular causal system. This requires ruling out all of the relevant possibilities down to one, namely, knowing what will happen, how it will behave under all conceivable variations.[7] However, for an explanation to convey this degree of understanding requires that the consumer of the explanation have no cognitive or conceptual limits. Human being have limits and these impose constraints, constraints on what constitutes relevant and comprehensible information, and these give rise to the selectivity of the demander's explanatory concerns. These different explanatory concerns give rise to what is called "levels of explanation". Different theorists have different views about what the levels are and how to demarcate them. However, what they all have in common is that explanatory concerns operate at one level, and should not be answered in terms that address the concerns of another level.

The *locus classicus* of "levels of explanation" is to be found in David Marr's book *Vision* (1982). He distinguishes "the three levels at which any machine carrying out an information-processing task must be understood":

[7] Such a being is not to be confused with a Laplacean demon, a being whose knowledge of the *actual* universe is so complete that it has complete predictive power. Wesley Salmon famously noted that such a being would lack scientific understanding. Scientific understanding is not just knowledge about the future, but an understanding of key scientific concepts like causation (which Salmon wouldn't want to give a counterfactual account of). The pragmatic view deals with this in a different, and I think neater, way. The Laplacean demon lacks scientific understanding, but that is not to be understood in terms of lacking an objective understanding of the concepts of causation (however construed). Rather, the Laplacean demon lacks the human epistemic concerns that underpins the entire scientific enterprise, and which renders concepts like that of causation useful. Hard though it may be to imagine, such an omniscient being wouldn't have the concept of causation because it would and could have no use for it.

Our subject with perfect understanding of a causal system cannot predict the future, but can understand how the (perfectly understood) system *would* behave in any circumstance. The future is *prepared for*, but it is not *predicted*. And if it were predicted, it wouldn't need to be prepared for.

Computational theory: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?
Representation and algorithm: How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?
Hardware implementation: How can the representation and algorithm be realized physically?
(Marr (1982), p. 25)

Marr's three levels make use of two main points that apply to any talk of levels of explanation. First, there is a *functionalist* point of there being multiple realizability of a high-level or functional property in lower-level properties. Second, there is the point that if one has explanatory concerns that operate at a certain level, addressing them at a different level is at best, sub-optimal, and at worst, completely irrelevant or opaque.

Some theorists see what is called the "personal level" as just another level in this sense: as a particular functional level where the causally relevant interventions are coarse-grained, and, in particular, coarse-grained to the extent where we are talking about whole persons, what they believe, desire, feel etc. Dennett's doctrine of the "intentional stance" seems to view things in this way. He presents us with the following thought experiment. Suppose:

> "some beings of vastly superior intelligence – from Mars, let us say – were to descend upon us […] suppose, that is, that they did not need the intentional stance – or even the design stance – to predict our behavior in all its detail" (Dennett, 1981, p.68).

The question then is: do these Martians miss out on anything in failing to use the intentional stance, the personal-level vocabulary of beliefs, desires etc.? According to Dennett, although they might be able to predict the exact motions of the fingers and the vibrations of vocal cords during an instance of a stockbroker buying shares in General Motors, if they fail to see

> "that indefinitely many *different* patterns of finger motions and vocal cord vibrations – even the motions of indefinitely many different individuals – could have been substituted for the actual particulars without perturbing the market, then they would have failed to see a real pattern in the world they are observing" (1981, p.69).

Note that even here, with its non-reductive take-home message, Dennett calls this "a predictive strategy". The plan is to predict how a causal system will behave at the relevant fineness of grain. The finger motions are not a relevant fineness of grain for gaining a predictive understanding of the stock market. The intentional stance is the relevant fineness of grain for gaining a predictive understanding of persons. Elsewhere (Wilkinson (in press)) I argue that this constitutes an important change from the views of the early Dennett of *Content and Consciousness* (1969), where the personal/subpersonal distinction was first coined, to the later Dennett of "the intentional stance" (e.g. 1981). I am here, as I do there, favouring the earlier Dennett. Personal explanation is explanation of a different *kind*. This is what I aim to show now.

1.3. Kinds of Explanation: Personal and Subpersonal

As I said, the pragmatic view of explanation allows us to have concerns that aren't causal or mechanistic (viz. predictive). And we do have such concerns. We are deeply social beings, who don't merely predict each other's behaviour: we evaluate it, we hold each other accountable, we ascribe obligations to each other, and so on. Being able to do this requires us to plug an epistemic gap in our understanding of a situation, but it is not an epistemic gap of a mechanistic and purely predictive nature. "Why did this person do that?" "Why does this person believe that?" We trade on the fact that there are correct and incorrect answers to these questions, and these answers inform us. But they do not do so by giving us causes or mechanisms.

Although the "personal/subpersonal" terminology was introduced by Dennett (1969), as Dennett himself acknowledges, it has analogues in Ryle and Wittgenstein (and also, arguably, in Jaspers).[8] Roughly, whereas subpersonal explanation is mechanistic or causal, answering "How" and "How come" questions (respectively), personal explanation answers a "why" question, where "why" is understood in a certain way. We do sometimes use "why", when our explanatory concerns are causal or mechanistic. For example, when we ask "Why is there a hole is the ozone layer?" We mean, "By what cause or process is there a hole in the ozone layer?" We know it isn't there for a *reason*. However, when we ask, "Why is there a STOP sign at the end of that road?" we are asking for a reason, a justification, a *rationale*, for its being there. Along with the distinction between justificatory and causal uses of "why" in the *question* ("Why is there a STOP sign there?/Why did you raise your hand?" vs. "Why is there a hole in the ozone layer?") we have the distinction between justificatory and causal uses of "because" in the *answer* ("Because there tends to be fast-moving traffic in the main road/Because I wanted to ask a question" vs. "Because he smoked too heavily"). Answering such a "why" question involves citing a person's reasons or grounds for believing certain things and acting in certain ways (or the general, publicly agreed, reasons, not attributable to a specific person, as in the case with the STOP sign).

With beliefs and actions, we often ask questions of one another: "Why do you believe that?" "Why did you do that?" In doing this, we are asking a very particular kind of question, and one that requires a very particular kind of answer. This answer is commonly called a *rational* explanation (Davidson 1963, McDowell 1985). As we saw in the previous chapter, "rational" has a categorical and evaluative sense. The way "rational" is used here is not evaluative: you can have "rational explanations" of *irrational* phenomena. It is in part

[8] As Dennett writes: "the lesson to be learned from Ryle's attacks on 'para-mechanical hypotheses' and Wittgenstein's often startling insistence that explanations come to an end rather earlier than we had thought is that the personal and sub-personal levels must not be confused." (1969, p.95)

because of this, and in part because the word “rational” is used in so many different ways, and in part because I see my view as sharing much with Dennett’s early view, that my terminological decision is to use “personal” rather than “rational”.

If you ask me, “Why did you raise your hand?” and I answer, “Because I wanted to ask a question,” that’s normally a satisfying explanation. If I tell you a full physiological story about what happened up until the point when my hand went up, that may be interesting, but it’s not an answer to *that* question. For starters, it doesn’t even describe me, the person, performing the *action* of raising their hand; it describes in very accurate detail a bunch of movements. Not only this, but a description of a causally related sequence of events is not what you asked of me. You were after a *reason*. The same applies when you ask the question “Why do you believe this?” You are after *reasons* for my belief, not any mechanistic story.

Now, why is this not a mechanistic explanation? It is not a mechanistic explanation because when I cite my reasons for raising my hand, it is not thanks to understanding a mechanism that you come to understand my action. It is in virtue of “meaning not mechanism” that the explanation I provide you with is satisfying. Now, you may object that we understand how meaning causally operates on people, so, in a sense, we do understand the mechanisms of meaning, the mechanisms of reasons and rationality. Two things can be said in response to this. First, this may be correct, but it is an unusual use of the word “mechanism”. What the objection describes more closely resembles causal explanation. We don’t know the mechanism by which certain reasons operate, but we know the sorts of things that they tend to give rise to. We know that certain beliefs and desires in certain contexts will give rise to certain actions. We certainly don’t know by what *mechanism* they do so. Second, I would say that, although it looks superficially similar, personal explanation is not causal explanation either, since the explanatory concerns of the consumer are subtly different. Many theorists will disagree with me here, but I will attempt to explain this now.

It is worth noting that there are those who are fully accepting of the explanatory autonomy of reason-giving explanation, who claim that reasons are causes. For example, unlike, say, Anscombe (and the early Dennett) Davidson sees reasons as causes. Now, although I side with Anscombe on this, I am willing to accept for the sake of argument, that reasons are causes, especially if one accepts a counterfactual theory of causation. If I hadn't wanted to ask a question, I would not have raised my hand. In that very simple sense, my *desire* to ask a question was a cause of my *action*. However, it does not follow from something being a cause, that something is explanatory *in virtue of being a cause*. I think that, since our explanatory concerns when we ask the special variety of why-question are not causal concerns, the explanation given in terms of reasons is not a *causal explanation*, even though reasons may well be causes. Why do I think that our concerns aren't causal? First, consider that causal explanatory concerns are closely tied to prediction. A causal understanding of a system allows us to predict how that system will behave in relevant counterfactual circumstances. Now consider a very common, retrospective, case of personal explanation. We want to understand why somebody did something (e.g. killed someone), so that we can hold him or her accountable. Finding out about this does not confer any future predictive power. (In fact, we already have the predictive power that we need: we already know how beliefs and desires give rise to and, sometimes justify, actions). However, it enables us to understand the individual. It renders them intelligible. Alternatively, consider a forward-looking case, but where the subject behaves in a deeply surprising way. When this happens, are we just bemoaning a failure to predict, to causally understand, the subject? I would suggest that there is more to it than this: we are perplexed by this *person qua agent*, by the fact that we find them unintelligible. To take an example that will crop up in Chapter 5, suppose someone, holds up a poisonous mushroom, and announces: "This mushroom is deadly poisonous and I have no intention of killing myself" but then pops the mushroom in his mouth. This behaviour is *surprising*, but it is not just surprise that you feel, but perplexity and confusion. In particular, you don't understand why this person ate the mushroom. You can hypothesize that they were lying, either about the poisonous nature of the mushroom, or their intentions to stay alive. Or they were demonstrating an antidote. However, on face value, this action is perplexing.

1.4. Connecting Subpersonal and Personal Explanations

We can use both personal and subpersonal explanations about the same case. To use an example that is highly relevant to us, we can ask about a patient with the Capgras delusion the subpersonal question, "How has their brain damage disrupted normal cognitive functioning?" A really good answer to this will make it altogether unmysterious why (how come) this particular damage has disrupted functioning in this particular way, and not in any other way. This will require causal and mechanistic understandings at different levels. However one can also ask, "Why (on what grounds) do they believe what they do?" In answering this, you can't use the same vocabulary as when answering the first, subpersonal, question. Dopamine dyregulation, modular damage, etc. none of these are even the right kind of thing to provide grounds for the subject.

Although it is correct that subpersonal vocabulary cannot feature in personal explanation, this is not to say that the cognitive sciences can't make very important contributions to personal explanations. They certainly can. The nature of the contribution is that it can give us an idea of the nature of the grounds that a subject might have (e.g. what experiences or emotions they might be undergoing) and how it is that they have them, or rather, why (how come) they have these experiences "and not others". For example, as we are about to see, subpersonal psychology can suggest that the Capgras patient is experiencing a feeling of unfamiliarity toward someone who appears to the subject like someone that is familiar. It can also suggest (for example by understanding the function of the damaged region) why it is that the Capgras patient experiences lack of familiarity to faces and not, say, to places or objects. Once we understand what the subject may be experiencing, there is scope for their claims to be rendered *intelligible*, namely, to be subject to personal explanation.

In other cases, the contribution of subpersonal psychology may be very different. It may be able to warn us when personal explanation is *unavailable*. That is to say, it may warn us when any attempts at understanding the subject in terms of subjective grounds would be a waste of time. We saw with our plane crash analogy that an understanding of the situation may lead us to conclude that no blame is to be attributed. Similarly, an understanding of some cases may lead us to realise that there are no reasons for certain behaviour or, perhaps even beliefs: they are simply caused. Later, and in the next chapter, we will consider something along these lines for the Capgras delusion.

For addictive behaviour, which are cases that *look* like goal-directed actions, something like this view is held by some theorists. Ross (2010) seems to take something like this view with regard to addictive gambling: "[we may] wonder what it is that keeps attracting [the subject] to gamble more. Is it thrills or money or relief from boredom or what? This is the wrong question to ask if we want to understand gambling *addiction*" (Ross 2010, pp. 140-141). I happen to have some reservations about Ross's view of addiction, but they are not central to our concerns.[9]

1.5. A Recap

So the general picture of explanation that we have looks like this. There are all sorts of different questions that we can ask, and they are satisfied by different answers. There are many different questions that can be asked about a complicated causal system, and they need

[9] Roughly, my reservations about Ross's view are that I don't see why we shouldn't try to understand practically irrational actions in personal terms. Can't we understand the smoker smoking outside the cancer clinic in terms of insurmountable urges? Sure, both we and the smoker, can *evaluate* the action of smoking as a bad one, an akratic one, one that is not a good way of fulfilling important goals in life (such as *not dying*). However, don't the urges provide motivation for smoking? Ross's view is, however, founded on the correct claim that "what's rewarding to the reward system and what's rewarding to the person are two different kinds of thing." But from this it simply doesn't follow that we can't understand gambling behaviour in terms of personal explanation, e.g, in terms of insurmountable urges, which are the product of a messed up reward system. One might say that it is through the subjectively felt urges that practical reasoning is hijacked, not, as Ross claims, that the reward system directly hijacks practical reasoning. Gambling addicts are sentient human beings, not zombies.

to be answered in a way that addresses the explanatory concerns of the person asking the question. The understanding that we are after in such a case is one that will confer predictive power. This can be achieved in a number of different ways, and there is a large and important literature on this, that we don't need to get into here. For example, as we briefly saw, there is the distinction between causal explanation and mechanistic explanation (see, e.g. Bechtel and Wright 2007). Roughly in causal explanation you draw law-like generalisations between cause and effect, so you would be able to predict in the present or absence of the cause whether there would be the effect. In mechanistic explanation, you try to understand how a mechanism works. You might then be able to predict how the mechanism will behave if interfered with in a certain way. Both are ways of understanding a causal system. In horizontal explanation, it enables predictions of temporally sequential events, in vertical explanation, it explains how a larger, usually functionally individuated, component is realised or constituted by the properties of smaller components.

However, there is an altogether different kind of question, which asks for the grounds and reasons for belief and action. These don't primarily demand answers that confer causal predictability (although there is predictability that is granted as a *side-effect* of assuming that people are rational in the relevant sense), but rather that confer intelligibility. With this picture of explanation firmly in mind, let's look, first, at the two main contrasting methodological approaches that have been used in attempts at explaining delusions, and then turn to different aetiologies of a specific delusion, the Capgras delusion. We will see that see that different aetiologies have concrete consequences for the availability of personal explanation. Their disagreement over how we are to understand the Capgras patient is, in a sense, analogous to disagreeing over the severity of the weather conditions during a plane crash.

2. Methodological approaches to explaining delusions

Let us look at the two main approaches that theorists have used in order to explain delusions. These are psychodynamic and neuropsychological approaches.

2.1. Psychodynamic approaches

Psychodynamic approaches attempt to explain delusions in terms of motivations, and, in particular, in terms of interactions between "conscious" and "unconscious" motivations (cf. e.g. Ahles 2004). They therefore, typically produce accounts that are totally compatible with there being no physiological abnormality to the delusion whatsoever. Although in principle you might claim that some of these motivations arise as the result of brain damage (see McKay and Anderson (2007) for a hypothesised "neuropsychodynamic" account of anosognosia, or denial of impairment), a purely psychodynamic account doesn't start with postulations of damaged brain regions, or chemical imbalances, or neurotransmitter deficiency.[10]

These accounts can provide seemingly satisfying personal explanations, in spite of the fact that, at least for the Capgras delusion, they lack plausibility as accurate accounts. People have a much better intuitive grasp of what it is to believe or desire something (granted, without putting it in terms of "belief" and "desire"), than they have of neurons or modules. They can also comprehend that having certain beliefs and desires can be painful or considered generally unacceptable by society and, as such, are themselves undesirable. There are things that one may want not to believe, and there are things that one may want not to desire. It is comprehensible that psychological processes will react in an epistemically unusual or deviant way in order to re-establish psychological order or bearableness. Thus someone who is told that they are dying of cancer may treat that evidence with bias, and may even end up (perhaps only indirectly), fooling themselves, against all the available evidence,

[10] For the early psychodynamic accounts we are about to see, this is unsurprising for historical reasons: the techniques that nowadays provide evidence for the postulation of such causes were unavailable.

into thinking that they are not dying of cancer. How might this be used to account, not for extreme cases of self-deception, but for our more exotic cases of delusion? How, for example, might it account for the Capgras delusion?[11]

Interestingly, Joseph Capgras himself presented a psychodynamic account of the delusion that now bears his name. He claimed that the Capgras delusion is adopted to resolve conflict between desires. He tied his explanation (cf. Capgras and Carette 1924) to the (then very popular) Oedipus complex as follows. The subject has a subconscious or repressed sexual desire for his mother. At some level he is aware of this and it disgusts him, and/or is worried about the social disapproval that such a desire would arouse. Therefore, he has a desire not to be sexually attracted to his mother, namely, a second-order desire not to have a desire.

How is this resolved in the subject's mind? Logically speaking, there are two options that would fulfil his second-order desire. On the one hand, he can either try to alter the fact that he is attracted to this very woman, perhaps by focusing on her negative qualities.[12] However, there is another option: he can continue to be attracted to that very woman, but simply tell himself that she isn't his mother. After all, it is the fact that this woman is "his

[11] Recall in the previous chapter that (some case of) self-deception, although epistemically deviant, is "understandably" so (and the epistemic deviance is also *intelligible,* namely, subject to personal explanation) so. Furthermore, one can argue that this comprehensibility often serves to excuse cases of self-deception from delusional status. To revisit the example from Murphy (2012), the woman who refuses to believe, against all the available evidence, that her son is guilty of murder, is not someone we think of as delusional. We understand that many people would do similarly in her stead. As a result, she doesn't conflict with our folk expectation and, as a result, the negative evaluation "that's delusional" isn't generated. In order for a psychodynamic account of a genuinely delusional phenomenon to work, the postulated beliefs and desires appealed to in the account would have to be unavailable to the folk (in the same way that the anomalous experience of the Capgras patient, on e.g. Maher's view, is unavailable to the folk, and so cannot confer intelligibility). As we saw in the last chapter, recourse to any kind of theoretical sophistication (whether neuropsychological or psychodynamic) cannot be used to exempt someone of delusional status (since this goes beyond "folk theory"). Psychodynamic accounts do precisely this: they postulate beliefs and desires that cannot be obviously postulated by the folk. (Note also that, just like Maher did several decades later, these accounts take themselves to be explaining cases of delusion, not "taking the delusion out of delusion".)

[12] This may be a common self-deceptive strategy: the person who has been left by the love of his or her life may, as a coping mechanism, build up an unrealistically negative image of the ex-partner in question (as a way of indirectly lessening attraction to that person, which is not something can be directly controlled).

mother" that is so troubling. Therefore he forms the delusional belief that this woman isn't his mother, and can continue to harbour desires for her unimpeded. These desires are mysterious subconscious ones that are never expressed. The delusional misidentification that was formed in the service of these desires, however, is expressed, giving rise to the delusional assertion.

Problems for this account are abundant. First of all, the majority of cases of the Capgras delusion are very obviously not going to be explained in this way. For example, not only are there cases of heterosexual men misidentifying their fathers, but there are even cases of Capgras delusion for pets (Ramachandran 1998, Ehr 1999), and delusional misidentification for inanimate objects (Abed and Fewtrell 1990, Castillo and Berman 1994, Ghaffari-Nejad and Toofani 2005). Are we to say that overtly heterosexual men can be attracted to their fathers, or that one can be attracted to pets or to a set of keys? Clearly this is preposterous. Even if these cases didn't exist, and we only had optimal cases for being explained in this way (e.g. a heterosexual male misidentifying his mother) there are still basic building blocks of this explanation that remain utterly mysterious? What are these subconscious desires? Why are they there? How do we know that they are there?

Over and above these problems, there is the fact that theoretical accounts have an obligation to make use of available relevant scientific evidence. With the arrival of certain kinds of evidence, in particular evidence suggesting underlying physical causes, comes the obligation to form certain kinds of explanation, namely, explanations that are in keeping with the advancements. In other words, explaining a phenomenon that we know is correlated with certain patterns of brain-damage without making *any* reference to the brain-damage in question (or its cognitive effects) is *prima facie* suspect since it is ignoring evidence that ought to be taken into account. Holding onto purely psychodynamic accounts would render this information and these exciting new methods and results redundant. As such, entirely psychodynamic accounts of the Capgras delusion have been all but abandoned. However, accounts that make partial appeal to motivational factors, and certainly that invoke personal explanations, still have explanatory use.

2.2. Neuropsychological approaches

Neuropsychological approaches essentially ask the following question: given what we know about the nature and location of the brain damage that a certain patient (or group of patients) has, and given what we know about the function of the damaged regions, how can we provide a plausible account about how these people end up exhibiting delusional behaviour? Neuropsychology, by its nature, is prone to two-way self-correction. Theories about the contribution that certain brain regions make to cognition will not only help us understand pathological cases, but will also be subject to revision if they end up failing to accommodate the relevant data from pathology (see Halligan and Marshall 1996, pp.5-6; Langdon and Coltheart 2000, for nice explanations of the neuropsychological project). Roughly speaking, when the ultimate aim is to diagnose and treat the brain-damaged patient, this goes under that label *clinical* neuropsychology, whereas when the ultimate aim is the theoretical one of understanding cognition generally, this is called *cognitive* neuropsychology.

One nice example of cognitive neuropsychology, as applied to the Capgras delusion, is to be found in the Ellis and Lewis (2001) review paper entitled, "Capgras delusion: a window on face recognition". In this they explain how the Capgras delusion sheds light on the neuroscience and psychology of face recognition.[13] Their account is built on pre-existing hypotheses about the role and function of certain regions, but gaps are filled in to form

[13] Since the Capgras delusion is considered a mental illness rather than a cognitive impairment, this is often called cognitive neuropsychiatry, rather than neuropsychology. I think there is a substantive distinction to be drawn here, but it is natural to see neuropsychiatry, especially methodologically, as a particular branch of neuropsychology. Roughly we think of prosopagnosia and ataxia as neurological impairments that an otherwise intact person has to cope with. Prosopagnosics are not *mentally* ill, they are face-blind. Capgras patients, in contrast, *believe* something unusual and are considered mentally ill.

hypotheses that are compatible with what is exhibited by Capgras patients, and other relevantly related deficits. This then feeds back into our hypotheses about face processing generally.

The advantages of neuropsychological approaches are abundant, especially when contrasted with purely psychodynamic ones. Firstly, they take into account all of the empirical data, and yet remain open to the possibility that the body of neuropsychological knowledge is open to revision if it encounters difficulties in accounting for something. Secondly, it has the scope not only for rendering the patient's condition causally tractable, but also for rendering the subject intelligible. This second point, in particular, is crucial for our purposes. It is with this that we examine existing neuropsychological accounts.

3. Existing Neuropsychological Accounts

In this section we will examine in more detail the original model of the Capgras delusion that Ellis and Lewis 2001 draw most heavily upon. This "inverse prosopagnosia" model is the classic "bottom-up" model.[14] We will them present a sharply contrasting kind of model, a "top-down" model. We will then end the chapter by reflecting on what these tell us about the availability of personal explanation for these delusions.

3.1. The Classic Bottom-up Account of Capgras delusion (Ellis and Young 1990)

Borrowing Bauer's (1984) model, whereby there are two streams for processing facial information – one covert, affective and anatomically dorsal, the other overt, semantic

[14] "Bottom-up" models are sometimes called "empiricist" models, because they hold that the subject's belief is grounded in experience. I dislike this appellation since it faintly suggests that the human subject is a perpetual hypothesis-making theorist. Furthermore, "empiricism" already has a perfectly good application, namely, to a certain family of philosophical views about the best way of getting knowledge.

and anatomically ventral – Ellis and Young (1990) put forward the influential proposal that Capgras delusion can be understood as a sort of "inverse prosopagnosia".

In cases of prosopagnosia, patients have difficulty in the overt recognition of faces, even though their ability to recognise other classes of stimuli can remain intact. Show them a picture of a familiar face (i.e. a friend, family member or famous person) and they will not be able to tell you whose face it is. However, fascinatingly, they appear to have differential autonomic responses to these faces, as measured by heightened skin conductance response (SCR). In other words, although they themselves cannot tell you whose face they are looking at, their affective system seems at the very least to be able to "tell" that it is someone familiar.

Ellis and Young hypothesized that Bauer's two streams can be selectively impaired, leading to (that Holy Grail of neuropsychology) double dissociation. According to them, whereas with prosopagnosia the affective stream for "covert recognition" is intact and the semantic stream for "overt recognition" is impaired, with the Capgras delusion it is the other way around. The Capgras patient, on their view, is presented with someone who, thanks to intact semantic processing, looks to them exactly like a loved one, but there is a lack of affective response. The patient therefore concludes that this person cannot be the loved one in question. This model was then given experimental support when it was discovered that, in contra-distinction to prosopagnosia, Capgras patients show diminished SCR when presented with familiar faces (Ellis et al. 1997; Hirstein and Ramachandran 1997).[15]

This model is highly attractive for a number of reasons. Firstly, it builds on and is consistent with, past research, in particular on prosopagnosia and face-processing (e.g. Bauer 1984, 1986; Landis et al. 1986). Secondly, it presents a hypothesis and shows how it might be falsified. If Capgras patients did not show diminished affective response then their theory

[15] However, it must be noted that this is inconclusive evidence since the lack of SCR could be a consequence rather than a cause of the misidentification, i.e. it could be that there is a lack of affective response because you judge them not to be a loved one, rather than vice versa. A normal SCR, however, would have presented seemingly falsifying evidence.

would be *prima facie* falsified. Thirdly, it presents an account for understanding the formation of the delusion that not only renders the phenomenon causally tractable, it also gives us an idea of what it might be like for the subject. It therefore makes the delusion at least in principle intelligible: you can put yourself in the patient's shoes by asking yourself, "What would it be like to see your mother, but to not feel that affective warmth, or to experience her as deeply unfamiliar?" In a related manner, it enables us to answer the why-question: "Why does the subject believe that this person isn't her mother?" We can answer something like: "Because she feels unfamiliar to him". This provides a personal explanation, in the sense sketched earlier. Finally, it explains the statistical and behavioural data. It explains why Capgras patients don't come across as deeply psychotic, or globally irrational. It also has the tools for explaining the fact the Capgras patients only misidentify loved ones (however, we will later see that this is not so with all cases). As Ellis and Young (1990) make explicit, it is only with people who look like loved ones to the subject (namely, the people we onlookers know *are* their loved ones) that one *expects* a strong affective response. When this expectation is frustrated, misidentification ensues. That explains why those who are not loved ones are not misidentified.

With all of this going for it, it is hardly surprising that the Ellis and Young model, at least in rough outline, has been widely accepted and constitutes an orthodoxy, or at least a starting point for other more recent theories. Accounts that rely on the basic idea of the Ellis and Young model include, but are not restricted to, Stone and Young (1997), Davies et al. (2001), Bayne and Pacherie (2004), Coltheart (2007). In the next chapter, we shall see that I have some suggestions of my own to make in improving this model.

3.2. Explanationist versus Endorsement Models

It is worth noting that Brendan Maher foreshadowed the Ellis and Young model. As we saw in Chapter 1, as far back as 1974 he claimed that delusions are unusual beliefs that are held on the basis of unusual experiences. However, if we accept this, there is an important issue to resolve concerning the precise nature and role of the anomalous experience. What content is carried or encoded in the experience? In other words, what does the experience "tell" the subject?

Following Bayne and Pacherie (2004), we can isolate two options, which correspond to what they call "explanationist" and "endorsement" models. For the explanationist, the experience only encodes a sparse content (something like "unfamiliar") and the delusional judgement is a hypothesis that has the role of *explaining* the anomalous experience. Roughly speaking, the subject reasons: "This man, who looks like my father, feels unfamiliar (or fails to invoke a feeling of warmth or familiarity in me). What could explain this? It must not *be* my father". In other words, the delusion is formed through an abductive inference on the basis of sparse content. Maher (1974) seems to hold something like this view, as do Ellis and Young (1990) (who refer to a "rationalization strategy") and Stone and Young (1997).

Bayne and Pacherie (2004), on the other hand, want to replace the explanationist model with their own "endorsement" model. Endorsement models claim that the content of experience is rich: the content of the delusional judgement ("that's not my father") is encoded directly in the unusual experience, and all that suffices, namely what the delusional judgement consists in, is *endorsement* of that content. To put it another way, what the subject's experience is "telling them" is something rich like "This person is not your father" rather than something sparse like "unfamiliar".

How are we to adjudicate between the two? As Pacherie (2009) neatly puts it:

> Experience-based accounts of delusions involve (at least) two components: (a) an explanation of the delusional patient's experiential state, and (b) an explanation of the delusional patient's doxastic state. Endorsement and explanationist models face distinct challenges in providing these explanations. Explanationist models appear to have an easier job of (a) than endorsement models: the less one packs into the content of the perceptual experience, the easier it is to explain how the experiential state

> acquires its content. […] But what explanationist models gain with respect to (a) they lose with respect to (b). The explanationist holds that delusional beliefs are adopted in an attempt to explain unusual experiences. The problem with this suggestion is that delusional beliefs are typically very poor explanations of the events that they are supposedly intended to explain. (2009, p.107)

In other words, the downside of the explanationist model is that it fails to explain why the subject adopts what we might call the "doppelganger hypothesis" and not some other, more plausible, hypothesis, whether this be self-constructed ("maybe I'm in a bad mood", "maybe I don't like dad anymore") or actively suggested by their doctor, like the "brain-damage hypothesis" ("the reason why this person feels unfamiliar is because I've damaged my brain in such a way that they will feel unfamiliar").[16] As Fine et al. (2005) ingeniously point out, if, as the patients sometimes admit, the doppelganger is an *exact* replica of the original loved one, then the patient ought not to expect them to feel odd. If a perceived individual is perceptually indistinguishable from one's father, one would expect them to feel like one's father too, namely, one would expect the deception to be complete. After all, what, other than perceivable properties, can the disguise, indeed any disguise, or indeed *any* conscious judgement of identification, be based on? (In the next chapter, we will see what my response to this question is.) Therefore the doppelganger hypothesis is not only a bad explanation compared to competing hypotheses: it is a bad explanation in that it almost seems to defeat itself.

Conversely, the main selling point of the endorsement model is that it can explain this. There is no hypothesizing: only endorsement. It seems to me that P, therefore P. However, as Pacherie points out, endorsement models "pack more into the content of the experience", and

[16] Reimer (2009), calls the delusional hypothesis "the impostor hypothesis" but that is a less neutral term than "the doppelganger hypothesis" and implies a malevolence on the part of the double that is not always perceived by the subject (e.g. patient DS in Hirstein and Ramachandran (1997) who says: "He's a nice guy doctor. He's just not my father.").

therefore have more explaining to do on that front. In the next chapter we will see some options for how to achieve this.

At this point, it is interesting to note, given what we established at the beginning of this chapter, that the "explaining" that is done at a) and b) correspond to subpersonal and personal explanation respectively. Explaining how the experience has the content that it has requires us to understand what has been disrupted, and how that will impact on experience. Experience itself is a starting point for personal explanation, an input for inference or belief formation. It is not *itself* explainable by personal explanation. You can't ask someone: "Why (on what grounds) did you have that experience?" We will see in the next chapter what I have to suggest in this regard. Explaining how one gets from the experience to the belief is paradigmatically personal. On the endorsement model, you get:

Q: Why did the subject judge that P?

A: Because it perceptually seemed to her that P.

Alternatively, on the explanationist model, you get:

Q: Why did the subject judge that P?

A: Because it seemed to her that Q, and the best explanation for Q was P.

It is extremely important to note, as Bayne and Pacherie emphasise, that both the explanationist and endorsement models count as bottom-up accounts in the sense that the delusional judgement is grounded in experience. They merely differ in their views about what the content of this experience is, and therefore, what kind of reasoning process the subject has to go through in order to form the delusion.

3.3. One-Factor versus Two-Factor accounts

Note, however, that delusions are not only formed, but also tenaciously held in the face of contrary evidence. Following Coltheart (2007), we might call these two features "formation" and "maintenance".[17] Although these bottom-up accounts explain the formation, neither the explanationist, nor the endorsement theorist provides an account of why it is that the delusion is held so tenaciously in the face of contrary evidence. In particular, regardless of whether the experience carries sparse or rich content, it still constitutes evidence on the basis of which the delusional judgement is formed, and presumably the subject can refrain from making this judgement, or can reject it later. On the endorsement theory, the experience tells the subject something like "this man is not your father", but the subject could in principle reject the testimony of his senses. After all, we can do this when we are presented with illusions that we know are illusions, but which have conceptually rich content: we may see the stick as bent, but we don't *believe* that it is. Bayne and Pacherie themselves are aware that they need to say more to account for the tenacity of the delusion. The richness of the experiential content cannot account for this alone. What else might be appealed to is what we will look at now.

As the name implies, a one-factor theory is one that claims that the patient has one deficit, and that that is sufficient to explain both the formation and the maintenance of the delusion. The deficit is usually taken to produce an anomalous or bizarre experience on the basis of which the delusion is formed. A two-factor theory claims that the bizarre experience, though very much necessary, is not sufficient to explain the formation and maintenance of the delusion. It needs to be supplemented by some abnormality, either a deficit or a bias, at the level of reasoning or belief-formation.

[17] Coltheart goes on to make the rather strong claim that "if we can find the answers to just [these] two questions [about formation and maintenance] in any case of delusional belief, we will have arrived at a possible explanation of that delusion." In fact, I take these two questions to potentially subdivide into two further questions, namely ones asking for personal and subpersonal explanations. Both the formation and the maintenance can, at least in principle, be addressed in mechanistic terms and in terms that ask for grounds. The Ellis and Young model potentially addresses both for formation (viz. damage to affective processing and feelings of unfamiliarity, respectively) but as it stands, is silent on the issue of maintenance, both in terms of the mechanisms that cause it, and in terms of the subjects grounds for holding the delusion for so long.

The first theorist to clearly express a one-factor view (which was also the first clearly formulated bottom-up view) was Maher (1974), who, as we have already mentioned, foreshadowed the experimental work that, over a decade later, came to support his bottom-up theory. Maher claimed that delusions are hypotheses that the subject forms and adopts in order to explain an unusual experience. What makes this a one-factor theory is that the delusions are developed as a result of "the operation of normal cognitive processes" (Maher 1974, p. 103). There is no reasoning abnormality: only an experiential one. The disagreement between one and two factor accounts is usually seen as a disagreement about whether the delusional hypothesis is a rational response to an anomalous experience. Assuming rough agreement on what the delusional hypothesis is, they disagree, first, about the nature of the anomalous experience, in particular its evidential weight, and secondly, about what constitutes a rational response. This disagreement is very important since it is a dispute about whether deluded patients have a more global pathology of thought and reasoning. One-factor theorists claim that they haven't; two-factor theorists claim they have.

In direct opposition to the official DSM definition's claim that delusions are held in the face of contrary evidence, Maher states:

> The delusional belief is not being held "in the face of evidence normally sufficient to destroy it," but is being held because of evidence powerful enough to support it. Where the patient may differ from the normal observer is not in the manner of drawing inference from evidence but in the kinds of perceptual experience that provide the evidence from which the evidence is being drawn. (Maher 1974, p. 99)

Two-factor theorists agree that there is an unusual experience (cf. Davies and Coltheart 2000), but think it is implausible that the experience alone can account for the relevant phenomena. They think that, without postulating some post-experiential abnormality, the one-factor theorist cannot explain, firstly, why such an implausible hypothesis is formed and adopted, and, secondly, why the patient doesn't reject the hypothesis once she is presented with better explanations. In other words, assuming that the delusion involves an inference to the best explanation of the anomalous experience, one would expect the delusional hypothesis pretty swiftly to cease to be the best explanation. For example, one would expect the following assurance on the part of the doctor to provide defeating evidence and hence cure the Capgras patient: "The reason why this person doesn't feel like your mother is not because she isn't your mother, but because you have had an accident that has damaged your brain and will make her feel unfamiliar." But this doesn't cure the patient; hence the need to postulate a second-factor, according to the two-factor theorist. The two-factor theorist is not only in a strong position to criticise the one-factor theories. She also has independent reasons for putting forward her positive view of what the second factor might look like. Since we will go into this is quite some depth in Chapter 3, I won't go into it here.

What can the one-factor theorist say in response to this claim of implausibility? Seeing as the criticisms the two-factor theorists launch rest on the idea that the delusional hypothesis is a bad explanation of the anomalous experience, there are two clear and complementary strategies open to the one-factor theorist. First, she can claim that the delusional hypothesis is not as poor an explanation as it appears to be. Second, she can claim that the experience is stranger, and hence in need of a stranger explanation, than the two-factor theorist realises. I say these are complementary because in combining the two strategies, they help each other out. The less implausible you take the delusional hypothesis to be, the less strange the experience needs to be in order for it to be a good explanation of it. On the other hand, the stranger the experience is, the stranger the explanation is allowed to be that is constructed to explain it. Reimer (2009), for example, protests that the nature of the anomalous experience makes all the difference. We just cannot underestimate what these patients are going through, experientially speaking. I think there is a great deal of truth to this, but it will require us do give up on some fairly ingrained presuppositions about experience and judgement and how they relate to one another. We will see more along these lines in the next chapter.

However, for the purposes of this chapter, it is important to see that there are views that don't fit the bottom-up orthodoxy at all. In a sense, they avoid these issues, altogether.

3.4. An alternative approach: Top-down theories

Within the camp of neuropsychological approaches it is not universally accepted that delusion formation is the "bottom-up" process that the Ellis and Young model suggests. In other words, it is not universally accepted that the delusional belief arises as result of an anomalous experience, an experience that (i) precedes the delusional judgement and (ii) serves as evidential input or grounds for it. According to alternative "top-down" theories, the delusion is certainly the result of brain damage (obviously, since we know that these patients weren't delusional before their lesions and became delusional subsequently), and there is an anomalous experience, but it is not the experience that is caused first by brain damage. Rather, the way things appear to the subject, experientially speaking, relies on their expectations and beliefs, which can be prior to the experience and colour it (or alternatively are a crucial component of the experience). As Campbell 2001 puts it, delusion "is a matter of top-down disturbance in some fundamental beliefs of the subject, which may consequently affect experiences and actions" (Campbell 2001, p. 89). Therefore, applied to the Capgras delusion, the subject believes that this person is not their mother, and as a result this person will not be experienced as such, or will be experienced as unfamiliar. It is not the experience that drives the judgement ("bottom-up"), but the judgement that alters the experience ("top-down").

This has been less popular than the bottom-up theories that have been built more directly on the Ellis and Young model. More specifically, top-down theories seem to lack many of the explanatory advantages mentioned earlier. Firstly, they cannot explain as easily why it is only loved ones that are misidentified. Secondly, they do not enable one to understand the subject's point of view: how can you simply judge on the basis of nothing, that this person, who looks just like your mother, is not your mother? Finally, and in related fashion, there is something perplexing about the idea that a knock on the head can directly induce a change to the subject's beliefs. It is far more understandable to think of the damage causing some deficit to the input of a belief-formation mechanism.

Implausible as top-down theories may seem, there are valuable insights to be extracted from them, especially concerning that way in which belief (or belief-like states) need not always be held for epistemic reasons. Furthermore, they have the *prima facie* advantage of accounting for the tenacity of the delusion, without having to appeal to a second-factor reasoning deficit. For the top-down theorist, the delusion is held for so long because it is not rationally held on the basis of evidence, experiential or otherwise, which the subject could reject. Rather than being the product of a rational process, the delusion is caused by brain damage, by something pre-evidential, pre-rational. Some have even claimed that it has the epistemic role of a premise or a Wittgensteinian hinge proposition (Campbell 2001; Eilan 2001).[18] Perhaps, for these reasons, namely, that is doesn't have evidential grounds, it doesn't even deserve to be thought of as a belief at all. We will examine this issue in Chapter 4.

4. Aetiology and Explanation

We looked at two importantly different kinds of explanation, and looked at different theories about how the Capgras delusion comes about. What impact have these aetiologies

[18] Wittgenstein describes hinge propositions as follows: "[...] the *questions* that we raise and our *doubts* depend upon the fact that some propositions are exempt from doubt, are as it were like hinges on which those turn. That is to say, it belongs to the logic of our scientific investigations that certain things are *in deed* not doubted. But it isn't that the situation is like this: We just *can't* investigate everything, and for that reason we are forced to rest content with assumption. If I want the door to turn, the hinges must stay put." (Wittgenstein 1969 §§341-3)

got on explanation? We will now see that top-down and bottom-up aetiologies have rather different consequences for the availability of personal explanation.

4.1. Explaining delusion within the bottom-up framework

As we have already mentioned, bottom-up theories open the possibility of rendering the Capgras delusion intelligible, since it can be seen as something that is inferred on the basis of experiential evidence.[19] A complete bottom-up account will contain a mix of personal and subpersonal explanation. The presence of the anomalous experience is explained in terms of mechanism (in the Capgras case, on the Ellis and Young model, this could involve explaining how lesions disrupt affective processing of familiar faces). However, the judgment, which is grounded in, inferred from, the experience is open to personal explanation. The *person* infers from their experience. To understand this properly, one has to ask the relevant kind of question, which is "*Why* does the *person* believe that this woman is not his mother?" And the relevant answer is "Because this woman feels deeply unfamiliar to him" (Or, as the endorsement model would have it: "Because the subject experiences her as not being his mother"). This is not a causal or mechanistic explanation, but a personal one. It renders the belief intelligible and, if correct, tells us all we need to know *within the scope of personal explanations*. We may shift our concerns to wondering, for example, what causes that particular bizarre experience. That is a subpersonal issue, to be sure. We may even look into the neural processes that underpin the inferential processes. None of these can answer the personal question. To re-iterate: mechanisms, brain regions, neurotransmitters; none of these are even the right kinds of things to provide grounds for the subject.

[19] Something similar was proposed by Chris Frith for delusions of control in schizophrenia. On what grounds could someone possibly deny authorship of his or her actions (which they are nonetheless performing successfully and in accordance with their intentions)? Well, Frith thought, if there is a monitoring deficit, then the normal experience of authorship will be removed or altered, and this will provide experiential, evidential grounds for delusions of control.

4.2. Explaining delusion within the top-down framework

However, as we have seen, not everyone subscribes to bottom-up theories of delusions. Some theorists claim that the delusion is not inferred, nor grounded in evidence, but caused. On such a view, any report (or even experimental evidence from SCR), for example, that the mother feels unfamiliar, is a consequence of the delusional judgment, but not grounds for it. She feels unfamiliar because she is judged to not be the subject's mother, and not the other way around. An upshot of this is that the presence of the belief (or delusional state, if you don't want to call it a belief) cannot be explained personally. In answer to the same question, "*Why* does the *person* believe that this woman is not her mother?" one cannot appeal to grounds or justification, since there are none. One can only answer: "Because (in the justificatory sense of because) she just does." This is obviously not a very satisfying response, but, then again, if these top-down aetiological theories are correct, it is vital for us to see that a satisfying answer to *this question* is not available. Another question must be asked if progress is to be made. An analogy with this might be the judge who is looking to see if there is somebody to blame in the wake of a plane crash. The discovery that the plane crash was not the result of poor decision-making, but was in fact caused by extreme weather conditions, results in a change of question. There simply is not answer to the question: "Who's to blame?" Similarly, if the top-down models are correct, there is no answer to the question: "Why (on what grounds) does the subject believe that this woman is not her mother?"

An upshot of this is that, unlike with bottom-up theories, the delusion must be explained subpersonally. The only question with an illuminating answer is not the personal one, but rather: "What has *caused* this person to behave the way she does?" (or, if it even makes sense, and we will see in Chapter 4 that that is debatable, "What has (directly) *caused* the subject to believe what they do?").[20]

However, note that, although, on these top-down theories, the delusional belief

[20] This accords with Jaspers' well-known claim that delusional subjects are "un-understandable" (by which he means "unintelligible" in my sense).

may not be amenable to such an explanation, any *action* (arguably, as we shall see in Chapter 4, by definition, if it really is an action) performed on the basis of the belief will be amenable to such an explanation, and this explanation will appeal to the belief. Personal explanations explain not just beliefs, but actions. We ask the relevant "why" questions about both beliefs and actions. In such a situation we get the following series of questions and answers.

> Q: Why did the patient stab her father (even though they seemed to have a good relationship prior to the event)?
>
> A: Because she believed that he was not her father, but an identical-looking impostor.
>
> Q: And why (on what grounds) did she believe this?
>
> A: We can't say. She just did. There were no grounds for her belief. It was just caused.

At this point we would need to delve into the subpersonal mechanisms to understand what is underpinning the (groundless) belief mechanistically.[21]

4.3. A Clarification: The relationship between "understandability" and "intelligibility"

In Chapter 1, I said that we call something delusional if it breaks a folk epistemic norm that I called "understandability". And in this chapter, I claim that personal explanation confers intelligibility. If Maher is right, then the judgement of delusional misidentification in Capgras delusion is intelligible in the sense that there is an answer to the question "Why does the subject believe that this person is not their father/mother/etc?" However, it is worth noting, as we did in Chapter 1, that Maher's view, if correct, doesn't "take the delusion out

[21] Note also that, although there may not be action that is explainable (given a plausible understanding of the concept of action, which we will examine in Chapter 4) in subpersonal terms, there is brute behaviour that requires this. We cannot ask of somebody with a nervous tick *why*, for what reasons, there are doing what they are doing. They are not doing it for any reasons. (Indeed one might even say that *they* are not *doing* anything). But there is presumably a perfectly good physiological explanation of why (how come) they have that tick.

of delusion". Delusions can be rendered intelligible, but they cannot be "understandable". So what, exactly, is the difference between "understandability" and "intelligibility"? In the rather technical sense I introduced in Chapter 1, "understandability" is a folk norm, formed by the expectations we have of how people form beliefs. As a folk norm, it is not informed by scientific advancement, at least not *directly*. It is still not OK, in the sense that it still grievously infringes our folk-epistemic expectations, to claim that someone who looks exactly like your mother, and who you accept looks just like your mother, is not your mother. That you have subjective grounds for doing so doesn't change that. However, "intelligibility" is not a folk concept. It is theoretically sophisticated, and the intelligibility or otherwise of a belief (or indeed an action) will turn on facts that can be brought to light by facts that are closed to the daily interactions of the folk. Indeed the top-down theorists are precisely denying that the delusion is intelligible in this sense. Bottom-up theories (in principle) confer intelligibility, but neither kind of theory can confer "understandability".

Conclusion

What we look for in an explanation of delusions (or indeed of any psychiatric phenomenon) is an understanding of how it comes about, *qua* causal process, but also an understanding of the subject and their grounds for believing what they do. Sometimes, the latter kind of explanation may not be available. However, if it is available and we don't provide it, we are missing something really important. Not only are we missing out on some important factual information in its own right (facts concerning, e.g., a subject's grounds for believing something), but also information that has crucial consequences in other settings, such as medical or legal settings. Looking at the brain of an addict won't, at least not by *itself*, tell me whether their drug-seeking behaviour is an autonomous action, deserving of a negative moral appraisal, or something that the subject cannot control and which requires pity and treatment.

Although the vocabulary of cognitive science cannot feature in the kind of explanation that we have called "personal" explanation, (neurons and modules aren't, and cannot be, grounds for the subject), cognitive science can in principle tell us, i) when personal explanations are available (for example one hopes that cognitive science will eventually adjudicate between bottom-up and top-down aetiologies), and ii) where their starting point might be (namely if it is grounded in experience, it could help us understand what that experience might be like). For example, Ellis and Young's work (1990) suggested (inconclusively, of course) that the Capgras delusion is i) amenable to an explanation in terms of grounds, and ii) that these grounds have something to do with unusual affect in the presence of loved ones. We then switch to personal vocabulary (or even the phenomenological vocabulary of "what it's like" for the subject) and look at how it renders the delusion intelligible.

In the next chapter, I will present my own aetiology, which is based on various problems I see with existing accounts. My aetiology resolves these problems, but brings up another issue. The Capgras delusion, after all of the hope of rendering it intelligible, may not be intelligible after all. In other words, it may not be the product of an inference (in the personal sense of inference, which is the intelligibility-conferring sense), and therefore it may not be amenable to a personal explanation. Then in Chapters 4 and 5 we will look at what that might tell us about its status as a belief.

CHAPTER 3

A Proposed Aetiology of Delusional Misidentification

Introduction

The purpose of this chapter is to provide a plausible aetiology of the Capgras delusion, with the hope that this aetiology can be extended to some other cases of delusional misidentification. Although the arguments in favour of the aetiology I propose are all contained within this chapter, the chapter fits into to rest of the thesis as follows.

In the previous chapter, we saw that there are two different kinds of explanation: personal and subpersonal explanations. We also saw that, within a bottom-up (experience-based) view of the Capgras delusion, one has to (i) "explain" how the experience has the content that it has, and also (ii) "explain" the passage from experience to belief. The kind of explaining in (i) is not going to be personal, but rather subpersonal. One can't ask for what *reasons* the experience is the way it is. The experience just is the way it is and one has to understand, rather, what has caused it to be that way. Roughly, one has to look at the patient's nervous system to get some further clues (beyond the subjective reports of patients) as to what the experience might be like. Conversely, the sense of explain in (ii) is personal. We want to know on what grounds the subject comes to form the delusional belief. How much work we have to do at (ii) will depend a great deal on our answer to (i), namely, to paraphrase Pacherie (2009), on "how much we pack into the experience". In the last chapter, we saw that this corresponds roughly to the *explanationist* and *endorsement* models, where the former packs less into the experience, and the latter packs rather more. In this chapter, I will attempt to justify packing rather a lot into the experience. As Fine et al. (2005) rightly point out, endorsement accounts (what they call "expression accounts") "require an explanation of the deficit causing the experience that a familiar person (or object) has been replaced" (p.148). In a sense, this is what I hope to start providing here. However, what I propose goes beyond the endorsement model as initially presented (e.g. in Bayne and Pacherie 2004, and Pacherie 2009).

How this chapter fits into *subsequent* chapters is as follows. The aetiology I propose explains, among other things, why (how come) the delusion is so resistant to counter-evidence. Indeed it does so by suggesting that the delusional misidentification is actually *caused* rather than *inferred.* In the next chapter, we see that a combination of this aetiology and a certain mainstream view of belief results in the Capgras delusion not being a belief. In short, a mental state has to be sufficiently sensitive to evidence in order to count as a belief, and, as I characterise the Capgras delusion, it is not sufficiently sensitive. I will argue in Chapter 4 against such a view of belief, and build on that in Chapter 5, where I will claim that the delusional misidentification is genuinely doxastic.

As we have seen, the Capgras delusion is the delusion that one or more known persons have been replaced by identical-looking strangers. Along with a number of other delusions, (e.g. the Frégoli delusion, mirrored self-misidentification, delusional intermetamorphosis) it is classified as a delusion of misidentification, which is to say that the subject misidentifies someone or something that one would expect them not to misidentify.[1] What is unexpected in the Capgras case is that the subject's perceptual system seems to be working just fine. She doesn't have any form of blindness, visual impairment, so no apparent problem with visual "input", so to speak. Nor does there seem to be any problem with categorising or processing this input: she has no form of agnosia. She recognises colours, shapes, objects and faces. However, in spite of this, she mis*identifies* people, and often (although not always) a strikingly small set of people, namely loved ones. In the presence of, say, her father, the Capgras patient will say, "That man is not my father; he looks exactly like my father, but isn't him." As we've seen, utterances with similar content can occur in the context of more global pathologies, in particular schizophrenia and dementia, but in

[1] Recall that in Chapter 1, I endorsed the view that delusion is attributed when folk epistemic expectations are infringed.

cases where they are caused by brain damage, the subjects have been known to lead surprisingly ordinary lives, both cognitively and emotionally, when kept away from the so-called impostor (see e.g. Hirstein and Ramachandran 1997).

In the previous chapter we saw that, over twenty years ago, Ellis and Young (1990) proposed a bottom-up aetiological model for explaining the Capgras delusion. It was so well received that it generated an approach to thinking about monothematic delusions that still constitutes an orthodoxy. I will label this orthodoxy the *inferential-evidential approach*, (IEA for short) since it takes the delusional misidentification to arise as an *inference* on the basis of some kind of *evidence* that is present in the subject's experience. Prominent proponents of IEA include: Maher (1974, 1988, 1999 etc.), Stone and Young (1997), Davies et al. (2001), and, to a much lesser extent, Bayne and Pacherie (2004), among several others. The aim of this chapter is (i) to characterise this orthodoxy, (ii) to criticise it, and (iii) to present an alternative.

I will proceed as follows. First, I will explain what I mean by "inference" and "evidence". Then I present Ellis and Young's position as well as some options that have recently been taken by others in elaborating this into different forms of IEA. I look at three problematic issues for IEA. The first two (the "logical" and "epistemological" issues) are philosophical in nature. They consist in explaining how the tracking of individuals (*identification*) differs more profoundly from the recognition of qualitative similarity (*recognition*), both logically and epistemically, than IEA seems to suggest. Furthermore, it is shown to be beneficial that a cognitive system should be capable of considering something's identity entirely independently of its appearance, and that this should involve some kind of tracking mechanism. The third issue (the "psychological" issue) is empirical in nature, and is elaborated throughout section 3. It amounts to claiming not only that the two tasks of identification and recognition are different, and that it would be beneficial to have a dedicated mechanism for the former, but that humans can and do perform them using two dissociable mechanisms. I present evidence for this claim (and in particular the claim that identification can be non-inferential in the relevant sense) from empirical work on dreams and the Frégoli delusion. In section 4, I tentatively present a theory, which introduces the notion of identity files, for the underpinnings of identification, and by extension of *mis*identification. I then examine how the theory compares with its competitors. Finally, I will conclude by suggesting some consequences that this aetiological model may have for an issue that forms the subject of the next two chapters.

1. What is the Inferential-Evidential Approach (IEA)?

This section is dedicated to presenting and clarifying IEA. I will begin by clarifying what I mean by "inference" and "evidence". I will then present the Ellis and Young model and clarify its epistemological and psychological consequences.

1.1. Clarifying "Inference" and "Evidence"

Both the terms "inference" and "evidence" get used in different ways. As I am using the term, an "inference" is a process that has personal, though not necessarily conscious, steps as part of a process of reasoning (in this case abductive reasoning). I mean "personal" in the sense that was sketched in the previous chapter. It is attributable to the person and not to a subsystem: it is the person, and not some part of them, that performs the inference. It is also the person who, through performing an inference, has grounds for belief, and citing these grounds provides intelligibility. This sense of "inference" is certainly more universally used in philosophy of mind and epistemology than in psychology. In psychology there is a use of "inference" on the basis of which parts of your brain can be described as performing "inferences". This is the sense in which people talk of your visual system performing inferences (e.g. on the basis of ambiguous input). Some recent paradigms (viz. the "Predictive Coding" paradigm, see Clark 2013 for a review) even think of the brain as constantly performing statistical "inferences". When I say that I am interested in whether delusional misidentification is inferential or not, I am clearly referring to personal inference. It would be trivial, or very nearly so, to say that the delusion arises as the result of an inference in the low-level sense, since almost everything in human cognition can be thought of in this way.

An inference, in the sense that interests us, can be personal but achieved automatically (it can be, to use Pryor's (2000) term, *psychologically immediate*) but the epistemic steps, if it is indeed an inference, should always be ascertainable (it is not *epistemically immediate*). How can we, therefore, tell between an automatic, personal inference, and something that is non-inferential in the personal sense? To use Pryor's own example, I might look at the fuel gauge in my car and, in a psychologically immediate way, come to believe that my car needs filling up. However, this is not epistemically immediate in the sense that if somebody were to tell me that the gauge is not functioning properly, I would thereby cease to take the gauge to be an indication that my car needs filling up. The following is not an entirely uncontroversial claim, especially in certain areas of epistemology, but I take many perceptual judgements to be epistemically immediate, namely, non-inferential. You don't *infer* from your visual experience of a glass on the table that there is a glass on the table.

How we think of "evidence" further clarifies what is meant by inference. "Evidence" is simply that which justifies a belief, drives an inference, and makes it epistemically rational (or irrational). So evidence is also a personal notion. Just as the low-level "inferences" that psychology sometimes appeals to are not inferences in this sense, so low-level information within a subsystem is not evidence.[2]

Evidential steps mediate a judgement that is based on inference, and an inference can be rationally defeated by competing evidence. Crucially, for the orthodox view of

[2] In psychology, and 'inference' can be a very low-level phenomenon indeed, such as the Bayesian 'inference' postulated in early visual processing.

delusions, evidence thus construed, firstly, is the sort of thing that you can ignore or misuse, namely, it is that on which reasoning bias operates, and secondly (and in a way that arguably strays from philosophical orthodoxy) allows for private experiential evidence.[3] It is, for example, not "evidence" in the everyday legal sense of publicly accessible entities such as fingerprints.

With these clarifications out of the way, we turn to Ellis and Young's model, and look at the sense in which it is a form of IEA.

1.2. Capgras Delusion as "Inverse Prosopagnosia"

You may recall from the previous chapter that Ellis and Young (1990), borrowing Bauer's (1984) model, whereby there are two streams for processing facial information – one covert, affective and anatomically dorsal, the other overt, semantic and anatomically ventral – put forward the influential proposal that the Capgras delusion can be understood as a sort of "inverse prosopagnosia." Prosopagnosics have difficulty in the overt recognition of faces. Show them a picture of a familiar face and they will not be able to tell you whose face it is. And yet, surprisingly, some of them appear to have differential autonomic responses to these faces, as measured by heightened skin conductance response (SCR). In other words, although they themselves cannot tell you whose face they are looking at, their affective system seems at the very least to be able to "tell" that it is someone familiar. Ellis and Young hypothesized that Bauer's two streams can be selectively impaired, leading to double dissociation. According to them, whereas with prosopagnosia the affective stream for "covert recognition" is intact and the semantic stream for "overt recognition" is impaired, with the Capgras delusion it is the other way around. This means that the Capgras patient is

[3] In philosophy, there are two very different readings of 'evidence'. On one, more externalist, reading, x is evidence for p if it actually increases the likelihood of p being true. On another, more internalist reading, x is evidence for p if it makes the subject more likely to believe p. To illustrate, on the first reading, a brain in a vat has no evidence for its "perceptual" beliefs, even though it thinks it has. On the latter reading it could have good evidence. This subjective kind of evidence is what Maher has in mind when he says that "delusional belief is not being held "in the face of evidence strong enough to destroy it," but is being held because evidence is strong enough to support it." (1974, p.99)

presented with someone who (thanks to intact subpersonal semantic processing) looks to them exactly like a loved one, but the experience is somehow strange (which is the result of deficient subpersonal affective processing). The perceived person feels unfamiliar and the patient therefore concludes that this person cannot be the loved one in question. This model was given experimental support (Ellis et al. 1997, Hirstein and Ramachandran 1997) when it was discovered that, in contra-distinction to prosopagnosia, Capgras patients show diminished SCR when presented with familiar faces.[4]

With this in mind, let us examine the notions of overt and covert recognition.

1.3. "Overt Recognition", "Covert Recognition" and the Delusional Inference

Let's start with the term "recognition". In philosophical terms, recognition involves a *judgment*, in particular, a judgment that you are in the presence of something that you have encountered before (or perhaps that you have been familiarized with, prior to encountering, through descriptions, pictures or diagrams). The categorical nature of this recognized thing can differ: it can, for example, be a kind or an individual. Thus an ornithologist can recognize the kind "rook" (a judgment he would express as, "This is a rook"), but can also "recognise" his mother (a judgment he would express as, "This is my mother"). One can also recognise an appearance: a property or collection of properties. Someone can recognise the appearance of a rook ("This bird has the appearance of a rook") and the appearance of one's

[4] It is worth noting, however, that Bauer's dual stream model is far from being universally accepted. Breen, Caine and Coltheart (2000), for example, deny that the dorsal stream is capable of recognition. Rather, with prosopagnosics the ventral route is damaged enough to prevent conscious recognition, but, in some cases (in particular in associative rather than apperceptive prosopagnosia), not so damaged as to prevent referral to the limbic system. This critique undermines the claim that prosopagnosia and Capgras delusion exhibit a double dissociation, but it can leave the epistemic gist of the Ellis and Young model intact, namely that a central *reason* (i.e. evidential subjective grounds) for the delusional misidentification is a lack of affective response. Breen, Caine and Coltheart (2000) themselves endorse this much, and attribute it (as Hirstein and Ramachandran 1997 do) to a disconnection between the FRU ("face recognition unit") in the temporal cortex and the amygdala. This explains the emotional response in associative prosopagnosia, which can still be thought of as "covert recognition" in the absence of overt recognition, without claiming that the dorsal stream is capable of recognition.
It is also worth noting that the claim that some prosopagnosics exhibit covert recognition remains untouched. Rather the *anatomical* claim that this is due to intact *dorsal* processes is to be rejected.

mother ("This woman has the appearance of my mother"). What all of these uses of "recognise" have in common is that they denote cases of *re*-cognition; there is some familiarity or expertise which is brought to bear on a particular perceptual case, and it is common to think of this in terms of stored information pertaining to the recognized thing in question (in these instances pertaining to rooks, mothers and appearances of rooks and mothers). What sort of evidence this judgment is based on will vary from case to case, but one thing is clear; certain classes of very important stimuli, e.g. faces, voices etc., receive enhanced processing. So, when healthy subjects see a face, a lot of very clever and complicated subpersonal processing ("inference") is going on, which we take for granted. Indeed prosopagnosia is a perfect illustration of what happens when this goes awry. Furthermore, our pre-theoretical wonder at the phenomenon of prosopagnosia is an illustration of the extent to which we take this for granted. As we will see, I think something very similar can be said about certain cases of the Capgras delusion: a certain tracking mechanism that we take for granted, and have no pre-theoretical understanding of, goes awry. As we saw in Chapter 1, this will give rise to a folk epistemic surprise and perplexity that will lead to delusion-status attribution.[5]

So much for recognition: what are we to say about "overt" and "covert"? As we saw, Ellis and Young think that the Capgras patient, when presented with her father, judges, "This man has the appearance of my father". This constitutes overt recognition. "Overt" information (on the basis of which one makes an overt judgment) can be understood as that which is accessible to consciousness and provides citable justification for the subject. If asked, "Why does this man have the appearance of your father?" she might reply, "Because he has the same nose, same eyes, same hair etc." And these are demonstrable features that (assuming her perceptual apparatus is working normally and she's not hallucinating) are available for others to see, and in principle could convince them of the correctness or

[5] We may express wonder at prosopagnosia, but not *epistemic* surprise since it doesn't give rise to belief.

plausibility of her claim. Granted, as we've just mentioned, our facial recognition is modularized and holistic (see Palermo and Rhodes 2007 for a review), so we have a great facility in visually processing faces. We don't *form* the judgment by consciously matching up facial features in the way just described, but the crucial point is that *post hoc* justification (and support) is available to the subject (and to anyone who might ask the subject).

What is meant by "covert"? One might think that "covert" means subpersonal, but we need to be careful here. The processes underpinning overt recognition of faces are subpersonal too (viz. to be explained subpersonally). The judgment, however, is personal. Similarly, the damage to affective processing is subpersonal, but insofar as the covert recognition is a judgement, it is personal, and just as personal as the judgment of overt recognition. It is the Capgras patient, the person, who judges "This man does not feel familiar/like my father." So this is covert, not in the sense that it is subpersonal, but rather in the sense that it is not overt in the sense just described. The judgment "This man does not feel familiar" is personal-level all right. Rather, if someone were to ask her, "Why doesn't this man feel familiar?" there is no justification that can be adequately put into words and, relatedly, there is no inter-subjectively available property that can be appealed to. The subject is thus robbed of the ability to justify her judgment to others (although it may be supported by private, experiential evidence) and would have to reply something like, "He just doesn't." An inability to (successfully) justify judgments to others is (and even more clearly so in the light of our discussion in Chapter 1) an integral part of any case that is likely to be called delusional.[6] As we have seen, delusions (whatever else they may be) are beliefs that lack justification or reasonableness by other people's lights. Although this covert judgment isn't yet delusional (she *could* claim: "Although this person doesn't feel familiar, like my father ought to normally feel, I can plainly see, and reason my way to the conclusion that it must be my father"), it is an important precursor.

[6] Some delusional patients try to offer (probably confabulated) justification for their delusions, whereas others do not. Sometimes they confabulate slight differences in appearance, like "his moustache looks different" (cf. e.g. Hirstein and Ramachandran 1997).

However, the delusional subject isn't merely claiming that this person doesn't feel familiar, as her father should feel. She is claiming that he actually *isn't* him. So where does the full-blown delusional judgment "This man is not my father" come in? Well, according to IEA, it is the product of a personal inference, more specifically, an abductive inference to the best explanation (Ellis and Young themselves call this a "rationalization strategy"). It takes something like this form:

1. This man looks like my father.
2. This man feels unfamiliar (hence doesn't feel like my father).[7]

[How do I explain that, although this man looks like my father (and hence should feel like my father), he fails to feel like my father?]

3. He must not *be* my father.

1.4. Accounting for Tenacity: Two-factor Theories

If the inference is to be an epistemically rational one, where "epistemic rationality" roughly means that a judgment goes only as far as evidence permits, then 2 (and what it suggests) would have to be strong enough evidence to defeat 1. If this is plausible, (and, as we saw in the previous chapter, it's far from obvious that it is) it explains the initial formation of the delusion. But the delusion is also highly tenacious: it is not just formed briefly and then rejected, but maintained over time and in the face of much contrary evidence. For example, no matter how much one tries to convince the subject, she still thinks that the man we know is her father, and whom she agrees looks exactly like her father, is

[7] To avoid possible confusion, IEA does not claim that there is a specific feel for the individual and that it is this that is lacking. Rather there is a general lack of familiarity (or general presence of unfamiliarity) and that, for the subject, excludes the relevant familiar people. In this case the relevant familiar person is the father, since the person the subject sees looks just like his father. There is not a specific fatherly feel that is lacking. Note also that 'this man doesn't feel familiar/like my father' could be replaced by "this man doesn't arouse a sense of warmth or familiarity in me". Whether it is a property attributed directly to the perceived individual, or indirectly as a property that ought to produce a response in the subject, the inference still goes through.

(seemingly on the basis of no grounds at all!) not her father.[8] Therefore 2, the judgment grounded in the affectively unusual experience, has to defeat not just 1 but also a great deal else: the testimony of others, including people of authority such as doctors. Not only this, there is, of course, the obvious implausibility of the delusional scenario: statistically people *just don't* look exactly like a certain individual and yet somehow fail to be them.[9]

As we saw, Maher insists that the affectively anomalous experience can play such an epistemic role: "delusional belief is not being held "in the face of evidence strong enough to destroy it," but is being held because evidence is strong enough to support it." (1974, p.99; as we saw, Reimer 2009 has recently supported this view, claiming that we simply can't underestimate the strength or strangeness of these patients' experiences). However, most have found this implausible. Instead, they claim, delusional subjects aren't rational and must have, in addition to their unusual experience, a reasoning bias or deficit (Stone and Young 1997; Davies et al. 2001).[10] In epistemic terms, delusional subjects attach more credence (i.e. more weight) to 2 than they ought to. The initial version of these "two-factor theories" builds naturally on IEA and we will address them later. They amount to claiming that the patient's judgment is based on an irrational inference.[11]

1.5. Explanation, Endorsement and the Role of Experience

Whether one is a two-factor theorist or not, if one adopts IEA, the affectively anomalous experience still has to constitute strong subjective evidence in favour of the

[8] There are sometimes confabulated grounds.

[9] The patients are sometimes sensitive to this implausibility. Patient DS (Hirstein and Ramachandran 1997) when asked why this man was pretending to be his father, he replied "That is what is so surprising, doctor; why should anyone want to pretend to be my father? Maybe my father employed him to take care of me … paid him some money so that he could pay my bills." Note, however, that this assumes that the judgment of misidentification is correct: the implausibility is located at the level of the impostor's motive

[10] Stone and Young 1997, however, unlike Davies et al. 2001, claim that the biases involved in delusion formation are not necessarily different from those that are present in ordinary belief formation.

[11] Some theorists have since pushed everything down to something that should be explained subpersonally (Coltheart 2005). On this view, the nervous system performs all of the relevant inferences.

delusional hypothesis. As we saw, there are, broadly speaking, two different views on the role of the anomalous experience in the formation of the delusion. So-called explanationist models are closer to Ellis and Young's initial proposal. They (e.g. Maher 1999) say that the content of the affectively anomalous experience is something sparse like, "This experience is strange" or perhaps "This man feels strange", and that the delusional judgment *explains* the bizarre experience (roughly, the subject reasons: "This man, in spite of looking like my father, doesn't *feel* like my father would feel, therefore he cannot *be* my father").[12]

Endorsement models (e.g. Bayne and Pacherie 2004) say that the delusional content is encoded directly in the unusual experience, and all that suffices is endorsement of that content. As will become clear, I am far more sympathetic to endorsement theories that to explanationist ones. In a sense, as I said in the introduction, a large part of what I am doing in this chapter is giving an argument in favour of "packing a lot into the experience" (and beginning to suggest what we might pack into it). However, there is a sense in which endorsement models hold onto an important aspect of IEA. Here is why.

According to endorsement theorists, there are not two separate judgments (one about how the person looks, and another about how the person feels) that, through a process of inference, lead to the delusional judgment. Rather there is one experience that "tells" the subject something like: "This man may look like your father, but it's not him". This is so far, so good. However, according to them, one in principle could have exactly the same subjective experience as the Capgras patient and not (even *should* not) go on to form the delusion. In other words, the Capgras patient in principle could (and *ought* to) disregard what his experience is "telling him", namely the *evidence* provided by experience, as one might, for example, with a well-known visual illusion (both Bayne and Pacherie 2004, and Pacherie 2009 explicitly cite the Muller-Lyer illusion in support of this). My proposal denies this

[12] One *prima facie* problem with such a view is that, if the experience is sparse and non-specific, why is there not a wider array of potential hypotheses used to explain it? ("Maybe I don't like dad anymore", "Maybe I'm tired" etc.).

possibility. First, there is an empirical point. The evidence that claims that there are patients who have "the same experience", but who don't go on to form the delusion (Tranel, Damasio and Damasio 1995) is, as we shall see, not very convincing. For starters, "sameness of experience" is asserted on the grounds of diminished skin conductance response. But this physiological response can be interfered with in any number of ways and does not demonstrate "sameness of experience". The second point is more philosophical in nature. On an attractive, but overlooked, view of judgment, experience and the relation between the two, the difference between judging and not judging *is itself and experiential difference* (see Pitt 2004). The sense of trusting or not trusting your senses is an experiential (or phenomenological, if you prefer) difference (see Ratcliffe 2004 and 2008 for an approach running along similar lines).

There is, therefore, even for the endorsement theorist, *inference* going on, namely, the weighing up of *evidence* from what the current perceptual experience is telling the subject, against the subject's background beliefs (including the general implausibility that people can look exactly like others and yet somehow fail to be them). This means that endorsement theorists need to rely on something like a reasoning deficit to explain the tenacity of the delusion just as much as the explanationists. Indeed endorsement theorists (Bayne and Pacherie 2004) express this explicitly.

In spite of this difference, my target in this chapter is primarily the explanationist model, and some of my arguments for "packing a lot into the experience" might be supportive of the endorsement model. However, if you're going to pack in the whole content of the delusion into the experience, why not consider whether you can pack in the judgment too? I think there is some support for thinking in this way.

1.6. Distinguishing Bottom-up and IEA

This brings me on to a potential query that I want to address, which concerns the relationship between the cluster of aetiological models called "bottom-up" models and IEA.

It is easy to think that these are the same, but I think that IEA does not exhaust “bottom-up” models, and is rather a sub-category. “Bottom-up” and “Top-down” refer to the relationship between experience and belief. Bottom-up aetiologies claim that the experience comes “first” whereas top-down models claim that the belief or judgement comes “first”. Now, I have but “first” in scare quotes because it is not clear that this means *temporal* priority or *explanatory* priority. There is a view that could give experience priority of either kind without claiming that the judgement is *inferred* from the experience, or, which comes to the same thing, that the experience provides *evidence* for the judgement. In other words, there is room for a non-inferential bottom-up theory. In other words, it would be somehow *based* in experience, but not *inferentially based*. Such a view would pack the judgement into the experience: it is the experience of judging that p, the experience of truly being in the presence of a certain individual. Such a view is certainly not top-down, (since the judgement is neither temporally nor explanatorily prior to the experience) and is arguably bottom-up (since the experience is at least explanatorily prior to the judgement: the experience explains the judgement and not vice versa), but it is certainly not IEA, since the judgement is not formed from an inference on the basis of evidence present in the experience. But this view does not enable personal explanation of the belief any more than a causal, top-down, view does, since, if the judgment is just part of the experience, and the presence of the experience is *just there*, it too can only be explained through a subpersonal story.

2. Problems with IEA: Queries from the Armchair

We have seen that several theories fall under IEA (one-factor, two-factor, explanationist and, to a lesser extent, endorsement theories). What they all have in common is that the delusion is formed on the basis of an inference based on evidence in the subject’s experience. On the explanationist view, the subject recognizes that this person *looks like* A, and this would usually be evidence to support the judgment that this person *is* A, but there is an experiential anomaly that epistemically defeats this.

There are three issues that I'd like to raise for IEA. The first two are non-empirical: they constitute a logical issue and a related epistemological issue. The third, empirical issue, the "psychological" issue, is the most important and will constitute section 4.

2.1 The Logical Issue

We noted that ordinary language allows that you can "recognize" an individual as well as a property or kind: you can "recognize" a rook, but you can also "recognize" your mother. Here, superficial linguistic similarity is obscuring a deep logical difference. The former ("This [bird] is a rook") is of the form Fa whereas the latter ("This [woman] is my mother") is of the form a = b. Namely, one predicates a property of an individual (the referent of a singular term), whereas the other makes a connection of identity between two individuals (or rather implicitly claims that two singular terms have the same referent). This is why I prefer to use "identification" rather than "recognition" for judgments of the form a = b.

However, you might protest, doesn't IEA take both the overt and covert judgment to be of the form Fa? Namely, something like "This woman looks like my mother", and "This woman feels like [or fails to feel like] my mother." On this view, what I call identification happens later and is inferred from this. That is IEA's position, but is precisely what I think is wrong with it. We shall later see that I take what, I suppose, is the explanatory equivalent of the affective processing to be *directly* involved in making a judgment of the form a = b. Namely, in my terms, it underpins identification, not recognition of any kind. This terminological distinction between identification and recognition is not arbitrary, but serves to mark what I take to be a deep difference in the epistemology and psychology of recognition and identification, namely, the judgment that something looks a certain way (or, less relevantly, belongs to a certain kind) and the judgment that something is a certain individual.

2.2 The Epistemological Issue

Before making my central epistemological claim it is important to see precisely what IEA epistemologically amounts to.

An analogy might be as follows. An ornithologist, who owns many acres of woodland, has a favourite rook, call it Bob. However, he often gets confused between Bob and the vast number of very similar looking rooks (although he can tell after a while from the rook's behaviour when he has got his initial identification wrong). To put an end to this confusion, he puts a ring on Bob's leg. Now when he sees a certain rook he might think, "This has the appearance of Bob (and therefore, so far, could *be* Bob)". Upon closer inspection, the ring is absent; therefore the possibility of it being Bob is ruled out. For IEA, the affectively underpinned covert recognition (or lack thereof) is epistemically analogous to the ornithologist seeing this ring (or seeing its absence). Both are about a supplementary and identifying piece of evidence.

However, this supplementary evidence can be defeated, and, more specifically, undercut.[13] If the ornithologist had reason to think that the ring had fallen off (e.g. he had found it on the ground), he would think twice about judging on the basis of the ring's absence that the rook in question isn't Bob.[14]

This is analogous to the Capgras case according to IEA. And yet, if we look at actual cases, the patients do have reason to believe that that evidence would be missing: they are made aware by their doctor of their brain damage and its effects. However, unlike our

[13] Following Pollock (1986) we can distinguish, within defeating evidence, between undercutting and rebutting evidence. Undercutting evidence defeats evidence e for p by putting the reliability of e in question. Rebutting evidence, on the other hand, defeats e by directly presenting stronger evidence for not-p.

[14] Conversely, if he had reason to think that a prankster had attached a similar ring on other identical looking rooks (e.g. if he saw two rooks with the same kind of ring), he would think twice about judging, on the basis of a rook's having the ring, that it is Bob.

ornithologist, Capgras patients do not refrain from judging on the basis of this lack of identifying information that the individual they encounter is not the individual in question. One way that IEA has of explaining this, as we saw, is by saying that the delusional subject doesn't give the undercutting evidence as much weight as she ought to (and gives her experiential evidence too much weight). However, another way, which I am proposing in this paper, is to say that the delusional judgment is non-inferential.

Recognition, one might think, is needed to set up the possibility of identification, and hence misidentification. Someone's not feeling like your father is not evidence that defeats the judgment that this man is your father if this man looks nothing like your father, and you were not tempted to judge that they were your father anyway. But identification should not always be thought of like this. Intuitive as this may seem, we shall see in the next section, especially when we examine the Frégoli delusion, that identification demonstrably isn't always parasitic on recognition in this way, but can operate entirely independently.

The fact that it can operate independently, though perhaps counterintuitive (and at times making it hard to verbally justify judgments of identification), is no bad thing when we consider the central epistemic point I want to make, which is as follows. Judging that two things are the same individual (*identification*) and judging that two things look the same (*recognition*) are very different kinds of judgment. You can judge, and indeed be justified in judging, that x at t1 and y at t2 are the same individual, in spite of them looking nothing like each other. Indeed (and this is the way around that is relevant to the Capgras delusion) you can judge that x at t1 and y at t2 are distinct individuals in spite of looking the same.

Some theorists, (e.g. Evans 1982) are sensitive to the idea that you might make identification judgments not on the basis of recognitional evidence, but on background knowledge (i.e. you might be able to locate your old friend James, since you've been told that he's working the morning shift at the sandwich shop, but you might also be warned that he has changed beyond all recognition). I want to take things even further. Sometimes identificational judgments are made on the basis of no evidence at all, not even spatiotemporal location. They are not the product of inference. These “intuitive” judgments can be based on a reliable mechanism. Sometimes this reliability breaks down, and you get misidentification (and if this breakdown is stable and persistent enough, you get *delusional* misidentification). This constitutes the psychological issue I have to raise against IEA.

2.3. The Psychological Issue

The Epistemological Issue was a necessary claim, stemming from the difference between the metaphysics of unique individuals and that of multiply instantiable appearances. In any possible world, for any intelligent subject, a judgment of identification is such that it can in principle be correct (and reliable/justified/non-accidental) without any support from recognition.

The Psychological Issue is very different. It is a contingent claim about how we humans happen to make such judgments of identification. The whole of the next section is dedicated to this. It makes steps towards providing the explanation that the endorsement theorist needs, of how the experience gets its content. However, it may even go beyond that to explaining how the experience gets its *force*.

3. On the Independence of Tracking Individuals (Identification)

Here I will just state the claim plainly: human judgments about numerical identity (identification) can and do bypass judgments about qualitative similarity (and spatiotemporal trajectory) altogether. In such a situation the judgment of identification (which could, of course, be wrong, as in delusional cases) is not something that is inferred. In other words, it is not the product of personal inference, biased or otherwise. It is with this that I introduce and elucidate the notion of tracking individuals, namely, judgments of identification.

3.1 Bypassing both Similarity of Appearance and Spatiotemporal Considerations

I don't want to deny that we often make judgments of identification on the basis of judgments about appearance, with the addition of some kind of abductive inference. We say to ourselves, "This looks exactly like my tennis racket; it is unlikely that there should be a tennis racket that looks exactly like mine, and that isn't mine; therefore it is mine". And of course, there are also certain principles about spatiotemporal location. If the racket is where I left it, then that further supports my judgment. Or if I see it in one place, and then an identical-looking racket somewhere where it couldn't possibly be, then I ought not to judge that this racket, on the second encounter, is mine. The error of IEA is to think that this generalizes across the board for all judgments of identification (including delusional misidentification). There are many routes that lead to judgments of identification. There are certainly different *inferential* routes, different kinds of *evidence* that can lead to a judgment of identification (e.g. perceiving similarity or keeping spatiotemporal track of an individual). However, my empirical claim is that sometimes there are routes to identification that don't make use of any personal evidence, viz. that aren't inferential at all. Sometimes, the judgment of numerical identity relies entirely on a non-inferential mechanism, which involves neither the judgment of qualitative similarity (e.g. "this woman looks like my mother") nor the application of spatiotemporal principles.[15]

Consider what happens when you encounter someone and you (implicitly) ask yourself "Have I met this person before?" In the vast majority of cases, there is no inference. You don't *work out* whether you've met them before. The question is just answered. To look at it from a different angle, when you see someone you haven't met before, how do you know you haven't met them before? The answer is often: you just *know*.

This appeal to intuitions about everyday cases may not convince, so I turn to more robust support for the possibility of non-inferential judgments of identification.

[15] A rare example of where (purported) identification not only doesn't use, but also actively flouts, spatiotemporal principles is in reduplicative paramnesia: the delusion that a place or individual is existing in two places at once.

3.2 Support for Bypassing: Identification and Recognition come apart in Dreams and Delusions

3.2.1. Dreams

Schwartz and Maquet (2002) compared the content of dream reports with functional imaging during REM sleep (i.e. the sleep stage during which vivid dreams are reported).[16] The central idea is that, in principle, dream features can be mapped onto specific distributions of brain activity.[17]

Dreaming uses similar resources to waking cognition, and depends on a large-scale neural network.[18] As Schwartz and Maquet point out, the content of certain bizarre elements of dream experiences, as well as the underlying brain activity, may resemble the experiential content and brain activity associated with the delusions of misidentification that we are concerned with here. I want to suggest that the mechanisms underpinning recognition and identification regularly come apart, not only in delusional cases, but also in the dreams of healthy subjects. Consider this dream report: "I had a talk with your colleague, but she looked differently, much younger, like someone I went to school with, perhaps a 13-year-old girl." Or this one: "I recognize A's sister […] I am surprised by her beard, she looks much more like a man than a woman, with a big nose" (both quoted in Schwartz and Maquet 2002). This dissociation between how someone looks and who they are taken to be, coupled with imaging results, shows, as Schwartz and Maquet neatly put it,

[16] Recently, Gerrans (2012) has made use of dreams to explain aspects of delusion. However, he uses them to make a very different point, namely, that the delusional misidentification should be seen as a response to experience.

[17] What is assumed by this approach is that the anatomical segregation of brain functions remains similar when one is dreaming to when one is awake. It has been shown that this assumption holds for audition: the presentation of auditory stimuli during REM sleep elicits responses that are similar to those elicited during wakefulness (so with no contrary evidence presenting itself the assumption is, not without justification, extrapolated across the board).

[18] However, during sleep there is highly diminished activity in the frontal lobe (which is normally taken to perform supervisory control functions). This is perhaps why dream experiences can be extremely strange, and yet go completely unquestioned by the dreamer. We will also see that frontal lobe lesions or atrophy may contribute to delusions.

> that neuronal processes during sleep can *simultaneously and independently* engage (1) unimodal visual areas underlying the internal generation of a perceptual representation of an individual's face [i.e. "recognition"][…] and (2) distinct multimodal associative areas in the *temporal lobe* responsible for triggering the retrieval of a familiar individual's identity [i.e. "identification"]. (2002, emphasis added)

Firstly, although simplistic, it is worth mentioning that the brain damage in Capgras patients is to the right temporal lobe, to which Schwartz and Maquet ascribe the role of "identity retrieval" (we will, in particular, see the relevance of this in section 4). Secondly, such a postulation of two distinct mechanisms for the tracking of numerical identity and qualitative similarity would suggest that, in principle, the two processing tasks can come apart.[19] In the Capgras delusion there is recognition without identification (someone looks like a certain person but is taken not to be that person) so there ought in principle to be cases where there is identification, or at least *purported* identification, without recognition.

3.2.2. The Frégoli Delusion

We have seen that this can happen in dreams (with bearded man-sisters), but it can also happen, like with Capgras, following brain damage. The Frégoli delusion (that a known individual is taking on the different guises of surrounding strangers) involves precisely this: purported identification without recognition.[20] Like in dreams, people are (wrongly)

[19] There are two readings of "identification": one is factive, so on this reading "identification" would mean successful identification. The other is non-factive, so identification could simply mean the activation of the identification mechanism, i.e. *purported* identification (which may or may not be correct). I use the factive reading and explicitly specify "purported" or "wrong", when I am referring to misidentification.

[20] Interestingly, there have been some cases of delusional misidentification for inanimate objects (e.g. Alexander *et al.* 1979) (and there have been reports of Capgras-style misidentification for animals and places too). This seems to illustrate just the kind of distinction I'm getting at. These patients will say

identified without the slightest overt justification. Feinberg et al. (1999), for example, report a brain-damaged patient, BJ, who:

> …approached a severely disabled, wheelchair-bound patient in his early twenties whom he had never met before, and claimed that the patient was his younger son. [...] He maintained this misidentification despite clear differences in physical appearance between the falsely identified patient and his son. Even when these distinctions were pointed out by staff, BJ maintained his original assertion that the patient was his son. (1999, p.378)

Both Capgras and Frégoli patients have no difficulty matching perceivable properties, namely, recognizing the *appearance* of thing. They just don't identify the people they misidentify on the basis of this matching. Both involve disorders of identification, not of recognition.

However, it raises the following question. If these patients were to base their judgments on appearances then they would get them right. The Capgras patient would judge that these people, who look just like her parents are her parents; the Frégoli patient would judge that this man who looks nothing like his younger son is not his younger son. So why aren't we designed to base our judgments of identification on appearances all the time? To answer this, we must consider the utility of that very same identification mechanism that wreaks such havoc in pathological cases. One nice upshot of this (which is a common kind of result in neuropsychology) is that we pre-theoretically think that we judge who people are on the basis of their appearances. If the view just presented is correct, we often, in fact, don't. It only seems to us that we do, because out identification and recognition mechanisms are functioning correctly and tend to "pull in the same direction."

things like, "that's not my set of keys: it *looks* just like it, but it isn't it". They fail to re-identify previously encountered objects, just as Capgras patients fail to re-identify previously encountered people.

3.3 Why the Bypassing?

It is very important that we should be capable of considering something's identity independently of its appearance, or indeed its voice. The properties of an object that can be perceived (usually visually, of course) don't reliably singularize, as billiard balls, identical twins, or identical-looking animals abundantly prove. In other words, you can be completely right about how something looks (or sounds), but very wrong about which individual it is. And this could be dangerous. Suppose I own two seemingly identical dogs, one is docile and the other is vicious: it is vitally important that I can *somehow* tell which is which. I may have developed the ability to do this, but am totally incapable of justifying (of having access to) how I do it.

What is usually most relevant when you encounter another person is not whether this person looks a certain way, but whether she is the same individual as the one you encountered at a previous juncture. This will let you predict her personality, how she will behave (so you will know whether you should be relaxed or on edge) and will let you know what she knows and what experiences she has shared with you (with loved ones, this can be rather a lot). The way an individual looks or sounds is only indirectly relevant insofar as it is often a pretty reliable indication of identity. Most people, if their spouse went missing, wouldn't settle for someone who merely looked like (or even behaved like) their spouse. They would want their spouse: a certain individual with a certain personality and past.

However, it is *prima facie* puzzling that the property of being an individual cannot physically trigger a tracking mechanism. It is not because your mother is a certain individual that she impacts on your nervous system in the way in which she does. It is because of her physical (e.g. visible, tactile, auditory, behavioural) properties. If someone were to recreate a duplicate of your mother, who looked, sounded, smelled, behaved etc, just like your mother, this duplicate would impact on your nervous system in the same way, in spite of being a different individual with a very different history.[21]

As is so often the case, one must not confuse the causally efficacious properties that the tracking mechanism exploits, with the property tracked, which in this case isn't really a property, but an individual.[22] When tracking identity, our cognitive system exploits all sorts of information (voice, mannerisms, appearance, smell etc.). But it can do so very automatically. When there is a match, this will provide a more conscious phenomenon that triggers or underpins the judgment of identification. We will judge that this person is a certain someone, or someone we've met, or someone we haven't met, without having consciously worked it out from their perceivable properties, and at times without having access to why we have judged this.

Since taking an individual to be in your presence is (to pick up on the turn of phrase that Schwartz and Maquet use) a "multimodal associative" phenomenon, the modality that triggers the identification (or misidentification) needn't only be vision. It could be the dominant modality, or a modality that is sufficiently used such that the affective response to that individual is regularly elicited and indexed through that modality. And indeed we needn't take "presence" too literally either. Consider what underpins your taking yourself to be talking to your mother on the phone. Or it may not even involve recognizing something uniquely produced by her, like the sound of her voice. Consider the warm feeling you get from receiving a text message from a loved one. Although a text message could easily have been written (forged) by somebody else, you trade on it being from the loved one, and it imbues even the simplest message with significance.

[21] Indeed, as we've mentioned, as Fine et al. 2005 point out, this is another perplexing thing about the Capgras delusion. Some patients admit that the impostor looks just like the replaced loved one (others confabulate slight differences). Surely then the patient would expect to be fooled by the impostor's perfect disguise!

[22] Unless you take "the property of being an individual" to be a property, but as 2.1 aimed to show, this is merely a mirage set up by natural language.

This view therefore has scope for explaining reported cases of the Capgras delusion in blind patients (see Rojo et al, 1999; Dalgalarrondo, Fujisawa and Banzato 2002 for a brief review), where the misidentification occurs on the basis of audition. In contradistinction, fully sighted patient DS (Hirstein and Ramachandran 1997) doesn't misidentify his father on the phone, but only visually (i.e. his dominant modality). Note also that the existence of one single blind Capgras patient is highly damaging to any theory that takes the delusion to necessarily arise from damage to neural pathways that are dedicated to the *visual* processing of facial information (e.g. Ellis and Young 1990).

3.4. Recap

Let's recap. We started by characterising IEA. We then looked at three problematic issues for IEA: two are philosophical in nature, and one is empirical. Of the philosophical issues, one concerns logical differences between judgments of identification (which are of the form a=b) and recognition (which are of the form Fa). The second philosophical issue, the Epistemological Issue, concerns the epistemology of judgments of identification, namely, the sorts of things that can support them, given the kinds of judgments that they are. Then there was an empirical issue, concerning, how human beings actually appear to make judgments of identification. Two important points have arisen, which have been supported by evidence from dreams and from reflection on the Frégoli delusion, namely, (i) the independence of the mechanisms underpinning these tasks and (ii) the non-inferential (i.e. non evidence-based) nature of some judgments of identification. Now I'd like to go back to something perhaps a bit more philosophical to ask: what does this judgment actually amount to? What is going on when a subject, any subject, makes a judgement of identification? It is with this that I introduce the notion of identity files.

4. Introducing Identity Files

In order to identify an individual, one needs to have an adequate conception of that individual *qua* individual. This means that the individual to be identified needs to have some *salience* for the subject *as an individual* (rather than, say, as a kind or as a bearer of certain properties). Think of people who are not salient in the relevant sense, that you encounter once, casually and in passing. These people you might qualitatively categorize on the fly in terms of kinds and properties (girl, red hair etc.). But if I asked you to *identify* such a person, if I asked you "*Who* is this person?" you would rightly protest that the question makes no sense. Answering that question requires one to make the connection between the encountered person and somebody one has encountered previously, whether in person, on television, in the newspapers etc. When I ask you to identify someone who has no salience for you *qua* individual, I am asking you to perform an impossible task: I am asking you to make a connection between this person (under a perceptual, or perhaps short-term recollective, mode of presentation) and a person at a previous encounter, but where there is no such previous encounter. On the other hand, walking past someone who looks like, say, John Cleese, it makes perfect sense for me to ask, "Is that John Cleese?" I'm asking, "Is the man we just walked past the very same man we are so familiar with from Monty Python?" Now what does this connection involve?

Some philosophers of language, when theorizing about reference to individuals use the somewhat metaphorical notion of a "mental file".[23] Mental files are created for individuals (upon first encounter) and retrieved (upon subsequent encounters) and are filled with information (viz. properties, predicates) which is taken to apply to the individual in question, but it is the file, or rather the acquaintance relation (which can be perceptual or through various informational chains, such as the media) on which the opening of the file

[23] Strawson (1959), Evans (1982), Bach (1994), Perry (2001), Recanati (1993) all use something like mental files.

was grounded, which fixes the reference. The point is that a mental file can be virtually empty, or filled with largely inaccurate information, and still achieve reference.

Suppose I meet someone, John (call him John-1), and I am subsequently told information that I wrongly take to pertain to him (e.g. that he has climbed Everest), by somebody who unwittingly thinks that I met a different John (John-2). No amount of false information, even though that information accurately pertains to John-2, will mean that my beliefs (and other thoughts) cease to refer to John-1 and start referring to John-2. My belief about a given individual, that he has climbed Everest, is obviously a false belief about John-1, not a true belief about John-2. Namely, to whom I refer is determined by who I actually encountered and not who happens to fit the description I take to apply to an individual.[24]

The content of a file for an individual constitutes our conception of that individual. This conception will involve a variety of different kinds of information that is taken to pertain to that individual; what they have done, when you have encountered them in the past, character traits etc., as well as what they look like. This conception, we have seen, can be false in many ways, and yet still be about that individual, since it is the initial encounter that caused the opening of the file that determines the referent. The conception, the contents of the file, can also be updated. Suppose my school friend, Jez, unbeknownst to me, went on to become a bank robber, underwent radical plastic surgery and fled to Australia. If I think to myself, "I wonder what happened to Jez?" I am thinking about that individual who is now in Australia with a radically different appearance (but, because I don't know this, I am not thinking about him under that mode of presentation). When I discover this information from a friend, I update my file for him. Rather, what it is to discover this *just is* to update my conception of him. If I go and see him in person and recognize his voice or his eyes or his manner, I update my conception of him in a more fine-grained and perceptual way. I momentarily open a file for "this man here present" but that gets very swiftly merged with

[24] Mental files resemble the PINs (Personal Identity Nodes) postulated in several models of face recognition. But I do not intend them to be entities in cognitive modelling; rather they are abstract, philosophical entities that function at the level of *thought* (or, rather, "informational content").

the file for “My old friend Jez”. The perceptual information I then get goes into that file and updates it. But in order for this to happen, the file needs to be correctly retrieved.

Put generally, judging that someone in your presence is a salient previously encountered individual involves the triggered retrieval of the correct file; correct in the sense that it is in fact the file that you had initially opened or created for *that very individual*. And it will be incorrect if you retrieve a file that wasn’t opened for that individual (or fail to retrieve the file and erroneously open a new one) and that will be *mis*identification. The file contains all kinds of information and physical appearance is, firstly, only one kind of property among many, and secondly it is neither an essential nor a singularizing property (viz. A can change her properties, and others can look just like A). This means that a file can be retrieved in spite of the person looking superficially very unlike the person for whom the file was opened. All that needs to happen is that this person is seen as someone previously encountered, either on the basis of perceptual evidence (either voice, appearance, mannerisms), or indirect evidence (spatiotemporal trajectory) or, as suggested by our reflections on dreams and the Frégoli delusion, pre-evidentially. The file can also, as in the Capgras case, fail to be retrieved even though the person in question *does* look like the person for whom the file was opened. On this view, the delusional misidentification in the Capgras case consists in an inability to associate biographical information about the loved one with the person now present (who is *de facto* the loved one). This is made nicely explicit in Lucchelli and Spinnler (2007, p.189), who tell of the case of Fred and Wilma (both fictional names, of course). When Fred denies that Wilma is his wife, he cites that he “knew [her] very well as his sons’ mother.” Obviously, “mother of my sons” is not a perceivable property, but it is in the “Wilma file”, which is failing to be correctly retrieved, and which he cannot associate with the woman here present.

Whereas the Capgras delusion illustrates a failure to retrieve a file (and, one might say, the erroneous opening of a new file) in the presence of someone who looks like the person for whom the file was opened (because they are that person!), the Frégoli delusion involves the erroneous retrieval of a file upon contact with someone whose appearance is – and, strikingly, is recognized by the subject to be – nothing like the person for whom the file was opened. It is worth noting that IEA, although *prima facie* plausible for the Capgras delusion, is less plausible for the Frégoli delusion. Presumably, since it appeals to diminished affect in the presence of loved ones to explain Capgras, IEA would need to appeal to heightened affect in presence of strangers to explain Frégoli. But why should heightened affect lead to the judgment that one is in the presence of a *specific* person (as is the case in Frégoli delusion). One would instead expect a certain "haven't I met you before?" response (a response we are all familiar with). On the other hand, the file-retrieval model explains this easily. The patient judges that it is that specific person because their cognitive system has retrieved or activated the file for that person.

5. Consequences of the Non-inferential File Retrieval View

What does the view that the delusional misidentification is non-inferential rule out? As I explained at the start of this chapter, when I say that delusional misidentification is non-inferential, I am clearly referring to the sort of inference that a person, and not a part of the person, does. To avoid a merely terminological dispute, let me outline what I take to be the concrete consequences of this view, and how it differs from existing views.

(i) It differs most sharply from the explanationist model. The content of the experience, on the file-retrieval view, is most certainly not "This person feels unfamiliar." It is rather something like "This is not my father."

(ii) Since the delusion isn't an inference on the basis of experiential evidence, it makes no sense to speak of subjects who have "the same experience" but who don't go on to form the delusion. (Recall that this is where I part company with Bayne and Pacherie)

(iii) The delusion will be something that the patient will be certain of when the perceived impostor/stranger is present. The subject will have no actual grounds for their belief, but may confabulate some *post hoc*.

5.1 File retrieval and different interpretations of the second factor

IEA claims that the patient adopts a delusional hypothesis in order to explain an affectively anomalous experience. We have seen that such views, as they stand, plausibly fail (*pace* Maher) to account for the tenacity of the delusion. Assuming that delusions involve an inference to the best explanation of the anomalous experience, one would expect the delusional hypothesis pretty swiftly to cease to be the best explanation (viz. a *rational* person *ought* to abandon such a hypothesis). For example, one would expect the following assurance on the part of the doctor to provide defeating evidence and hence cure the Capgras patient: "The reason why this person doesn't feel like your mother is not because she isn't your mother, but because you have had an accident that has damaged your brain and will make her feel unfamiliar."[25] But it doesn't. As we saw, this has led some (Stone and Young 1997; Davies et al 2001) to postulate a "second factor" in the form of a reasoning deficit.

I think that two-factor theories are broadly correct, in particular, in the claim that these delusions require two loci of damage (one temporal and one frontal). I am obviously not going to blindly contradict anatomical data. However, the second factor, according to the view I am proposing here, should not be seen as a reasoning bias, at least not in the personal, epistemological, sense. For this reason I am sympathetic to two-factor theories, but urge that we need to tread carefully when making claims about the second factor. The view I am proposing is explicitly opposed to early versions of the two-factor theory (Davies et a. 2001), which seem to imply IEA, or take it for granted. I will call such views "two-factor IEA". It must be noted that recently, two factor theorists have pushed the second-factor to a "lower level", and I have no issue with that. However, I think it is important to see that, if this is the

[25] As mentioned, this would be undercutting, rather than rebutting evidence.

case, the frontal damage, the second factor, will have a direct influence on the conscious experience itself, and it makes no sense to speak of the subject inferring *differently*, on the basis of the *same* experience, as another subject who is not delusional.

Proponents of two-factor IEA (e.g. Stone and Young 1997, Davies et al. 2001) tend to take two things to support their view. The first is that there are patients who appear to have a lack of affective response to familiar faces, hence that are *prima facie* similar to Capgras patients, but who do not go on to form the delusion (Tranel, Damasio and Damasio 1995). What is hypothesized is that these patients have the first factor, but lack the second-factor reasoning deficit. In spite of this, these cases fail to offer good support for this because their lesions are very different from those found in Capgras delusion (which tend to be associated with a combination of right lateral temporal lesions and dorsolateral prefrontal damage). These cases have ventromedial prefrontal damage, which is absent in Capgras patients.[26] In related fashion, such patients, although not delusional, are in many respects more globally disabled than Capgras patients. Unlike Capgras patients, they treat loved ones and strangers identically, and have severely deficient decision-making capabilities.

A second kind of support is more direct in claiming that delusional patients demonstrably exhibit reasoning biases. However, the cited experiments (Garety 1991, Huq et al. 1988, John and Dodgson 1994) all involve schizophrenia rather than brain-damaged patients. In patients with such a global psychiatric pathology as schizophrenia, one would expect general reasoning biases. As far as I know, there is no evidence that such deficits exist in Capgras cases that occur in the context of brain damage only.[27] Brain-damaged

[26] One might, for example, hypothesize that although these patients have no difficulty making the correct identification, they lack the ability to associate the appropriate emotional salience to the episodic information that is telling them that they are in the presence of a loved one.

[27] Granted, McKay and Cipolotti (2007) document a case of an internalizing attributional style in purely organic cases of Cotard delusion. This supports Young and Leafhead's (1996) hypothesis that Capgras and Cotard delusions may have the same 'first factor' (viz. an affective disruption) but a different second factor (viz. attributional style). Whereas Capgras involves an externalizing attributional style ("she feels affectively flat because *she*'s not my mother"), Cotard involves an internalizing attributional style ("she feels affectively flat because *I* am dead"). As the vocabulary suggests, these are not *biases* in the sense of being misuses of evidence. Rather they are attributional *styles*, default (possibly mood-based) stances on the world.

Capgras patients (i.e. ones that aren't schizophrenic) appear to be surprisingly reasonable and level-headed. You can see it for yourself in interviews with them. That is part of what is so striking about the delusion.

On my proposal, since the identity file is retrieved (or fails to be retrieved) as the result of a non-inferential mechanism, the delusion is not the product of any inference, biased or otherwise. This certainly fits with the accounts of Capgras patients. Every time they lay their eyes of their loved one, they cannot help judging that the perceived person is not the loved one in question.[28] Now, as I mentioned, what the two-factor theorist may then be right about is that frontal damage (in particular to DLPFC) is necessary for the formation of the delusion. What this may cause is some kind of "reality-testing deficit" (Hohwy and Rosenberg 2005). This deficit, however, needs to be understood subpersonally, and would then explain in part why (how come) the delusional experience, carries the force that it does (another contributing factor is that the tracking mechanism I describe, which underpins identification, is so often used, relied upon and hence taken for granted). It is illuminating to note that DLPFC is hypoactive during dreams, which may explain why we are so uncritical of our dream experiences. But note that our uncritical stance towards our dream experiences is not something that should be explained in terms of biased inferences operating upon experience. It is simply part of the phenomenology, the experience, of dreaming.

5.2 Misidentification of Loved Ones Only?

It has often been noted that IEA has the attraction of explaining why Capgras patients only misidentify loved ones. Namely, the patient sees someone who looks like a loved one and expects that person to feel like a loved one (or to provoke an appropriate affective response). It is this frustrated expectation, this mismatch, that drives the delusional inference.

[28] Although in many cases there are confabulated reasons ("His moustache looks different), Ramachandran's patient, DS, at times insists (on a video available online) "I just *know* that's not my father".

This expectation is not present in the case of those who are not loved ones, and hence they are not misidentified. How does the file retrieval view account for this pattern?

The first thing to note is that several Capgras patients go on to misidentify far more than their loved ones. Ellis et al. (1997) mention a patient who was under the impression that almost everybody in a town had been replaced. Furthermore, they need to explain the cases where the duplicates themselves are duplicated (i.e. the impostors are replaced, as in the original case described by Capgras and Reboul-Lachaux (1923)). This cannot be the result of expected emotional response, since the subject would not expect warmth from the impostor. Nevertheless, the fact that some patients seem to only misidentify their loved ones needs to be explained by the file-retrieval view.

Capgras cases are messy and varied. But to think of it in terms of files and tracking mechanisms gives us flexibility to explain a wider variety of cases. IEA cannot explain Frégoli delusion, and among Capgras cases can only explain ones where loved ones are misidentified, and only misidentified once. This is all speculative, but the file retrieval view can at least potentially explain the variety of cases in terms of different damage to mechanisms for retrieving or triggering the retrieval of files (and also the creation of new files). Patients who misidentify differing classes of people (viz. either just loved ones or all acquaintances) in either stable or reduplicative ways (viz. the doubles are stable or reduplicated) will be understood in terms of differing damage to these mechanisms. Suppose one common trigger for mental file retrieval is affective response. With the category of known people who are not loved ones, a strong affective response is not going to underpin identification, so, either a total absence of affective response is not enough to disrupt or over-ride the retrieval on the basis of cues such as appearance and mannerisms, or there is enough affective processing intact to allow for file retrieval. Non-loved ones are, in these cases, not misidentified because they are not salient enough. When a small class of people is misidentified, it is usually people who play a substantial role in the daily life of the subject, such as parents (often that the subject still lives with) or a spouse. In the cases where more than just loved ones are misidentified, this may be due to an even greater affective deficit and would indicate that even identifying people who aren't loved ones requires an affective component. But even in cases where only loved ones are misidentified, it has nothing to do with frustration of expectations set up by intact recognition (as Ellis and Young suggest). Indeed the very existence of the Frégoli delusion demonstrates this, since surely strangers would not be expected to arouse any emotional response.

Furthermore, the Ellis and Young model cannot plausibly account for the following case. Patient DS, who we mentioned before (Hirstein and Ramachandran 1997), is a 30-year-old Brazilian man who lived with his parents, and developed the Capgras delusion towards them after being in a car accident. Although DS misidentifies his parents and correctly identifies mere acquaintances, he occasionally duplicates *newly encountered* people. How could the expectation of positive affect drive the required inference, since, presumably, with new people there wouldn't be one?

The file retrieval view, on the other hand, can in principle explain this. When DS meets someone for the first time, a file is created in the normal way for them and this gets filled with any associated information. But if this person leaves the room for 30 minutes and returns, instead of the old file being retrieved and added to, an entirely new file is created: DS behaves as though he is meeting somebody new.[29] This is not the lack of recognition that you get with amnesiacs. On the contrary, he'll have no difficulty remembering that he met someone, and he'll remember what he or she looked like. He'll, for example, say: "I've just met your twin, and he was dressed just like you." But it seems that the timing has to be right for this unusual duplication to occur. This is obviously speculative, but we might postulate the following.

[29] Within the identity file framework, I take it that when we fail to retrieve a file upon encountering someone, we automatically create a new file for "This person", however short-lived that file may be. Unless, of course, we are no longer experiencing that person as an individual at all. Perhaps this happens in the very advanced stages of dementia.

When you meet someone briefly and for the first time, your brain has a fast-learning mechanism that makes an affective profile for the next encounter with that person, a very weak one that constitutes the identification in the absence of other higher-level non-affective information. When the affective response fails to occur on the second encounter, in DS's case, due to his brain damage, the file is not retrieved: a new file is erroneously created.

But, given time, there will be enough knowledge of surface cues acquired for DS to be able to make the identification on the basis of inference from such cues. That is why although DS duplicates both new people and his parents, he doesn't duplicate the intermediate category, i.e. when he sees people a) who are not newly encountered but b) who are not loved ones. And that is also why he doesn't duplicate, over and over again (i.e. the doubles themselves have doubles), everyone that he has met since his accident.[30] To illustrate, he doesn't duplicate those he has seen repeatedly or for a longer period of time, such as his neuropsychologist, whereas he might sometimes duplicate a cleaner or a nurse. DS's cognitive peculiarities will obviously reflect the nature of his brain damage, and patients with different damage will exhibit different cognitive profiles. The patients, for example, who, unlike DS, duplicate over and over again may, for whatever reason, have problems with forming stable, and correctly retrievable, files for individuals. Patients who misidentify only their parents may differ from patients who misidentify all acquaintances in the problems they have with file retrieval cues.

Conclusion

Let's take stock. We have characterised IEA, noted epistemological differences between judgements of identification and judgements of recognition. We have noted that a non-inferential route to identification, one that bypasses recognition (and spatiotemporal

[30] Interestingly, the first reported Capgras patient, Mme M. (Capgras and Reboul-Lachaux 1923) duplicated people over and over (i.e. the doubles themselves have doubles) – her husband as much as eighty times!

considerations), would be useful. We have seen from intuitions about meeting strangers, and from evidence from dreams and the Frégoli delusion, that such a route would not only be useful, but actually seems to exist and be put to use in human cognition. Undoubtedly, more needs to be said about the mechanisms underpinning all this, and we need a better understanding of what role affect really plays in the misidentification (and what we mean by affect in the first place). My hope is that I have shown that IEA has serious problems and that the non-inferential file-retrieval view, although in need of some scientific fleshing out, deserves serious consideration. Aside from its plausibility, it deserves to be considered because it is extremely relevant to current philosophical debates that concern delusions. One is to do with our previous chapter; the other is to do with our next two.

The first is the issue of whether delusions are "intelligible". This is the question (which goes back at least as far as Jaspers (1963)) of whether we can understand *why*, in the sense of "why" that asks for epistemic grounds, subjects believe what they do. According to IEA, we can in principle understand why they believe what they do, even if it requires a large, perhaps impossibly large, imaginative leap ("What would it be like to see your mother, but to have her feel very unfamiliar?"). The affectively anomalous experience provides grounds for the belief. But if my non-inferential view is right, that question cannot be answered, since it is the wrong question. The *why* question should be replaced by a *how* question. The relevant question becomes: "What causes the misidentification (which I suggest may be understood as a deficit in file retrieval)?" The belief itself, in the terminology introduced in the previous chapter, is not amenable to personal explanation.

The second debate, which we will examine over the next two chapters, is whether delusions are to be thought of as fully-fledged beliefs, or beliefs at all (e.g. Bayne and Pacherie (2005) say 'yes', Currie (2000) says 'no'). Now, if you think that there is a constitutive link between beliefs and evidence (or reasons), and if, as I am suggesting, these delusions of misidentification are based on no such evidence, then you may wonder whether delusions of misidentification count as beliefs at all (might they not be rather "experiences" (cf. Hohwy 2004) or “imaginings” (cf. Currie and Jureidini 2001)). Of course, the relevant question is then whether you want to think of beliefs in this way, namely, as constitutively tied to evidence or epistemic reasons. That is the central topic of the next chapter: How should we think of beliefs?

CHAPTER 4

On Failing to Believe: A "Downstream Only" View

Introduction

As we have seen, the Diagnostic and Statistical Manual of Mental Disorders begins its official characterization of delusion as follows:

> A false *belief* based on incorrect inference about external reality that is firmly sustained despite what almost everyone else believes and despite what constitutes incontrovertible and obvious proof or evidence to the contrary (DSM-IVTR, 2000: 821, emphasis added).

As we saw in Chapter 1, this characterization can be contested on a number of grounds. However, the contention that I will be focusing on in the next two chapters is as follows: do delusions deserve to be counted as beliefs at all, let alone defective ones? Following Bayne and Pacherie 2005, call the debate concerning whether delusions are beliefs, the doxasticity debate (henceforth DD). Call those who claim that they are beliefs "doxasticists", and those who deny that they are beliefs "anti-doxasticists".

Now, one's answer to the question, "Are delusions beliefs?" will obviously depend, in part, on what one takes belief and believing to be. It will also depend on what one takes delusion to be. Furthermore, unless one takes the features that make a phenomenon delusional to be the very same features that prevent it from being a belief, one's answer may also depend on what cases or kinds of delusion one focuses on. We can't assume out-of-hand that the belief-status will be constant across the wide array of phenomena that we call delusional. It will then also depend on the nature and aetiology of the delusion that one has chosen to focus on.

Bearing this in mind, the next two chapters are as follows. In this chapter, I put aside considerations about specific cases of delusion, and predominantly focus on different views about the norms of belief. In particular, I ask what any subject must do (or, rather, exhibit) in order to count, or fail to count, as a believer. Then, in the next chapter, I focus on the Capgras delusion, and examine whether, and in what sense, it deserves to be counted as a belief. As for which aetiological model of the Capgras delusion I use, it should come as no surprise that it will be the one expounded in Chapter 3, namely, the non-inferential file retrieval view. One way of locating this issue within the thesis as a whole is to note that, on certain views about the norms of belief, which we will examine in this chapter, if the Capgras delusion really is formed in the way that I claim it is formed, namely, non-inferentially, it thereby fails to be a belief state (it is, for example, something more like a perceptual state, or a state of imagining). Whether this is an unacceptable conclusion or not, the view of belief that I will put forward in this chapter will not have this as a consequence. The next chapter will address whether *other considerations* may rob the Capgras delusion of belief status.

As just mentioned, in this chapter, I examine what a subject must do in order to count as believing. I proceed as follows. I begin by examining why anyone might be tempted to deny that delusions are beliefs. This serves to give the reader an idea of what we mean when we say that someone is failing to believe what he or she seem to believe. It also serves to introduce some important issues and notions, and gives the reader an idea of how the work in this chapter relates to this debate concerning delusions. Then, I clarify some methodological issues concerning the question "What is belief?" I then put forward one answer to the relevant version of that question; namely, I put forward a view about the phenomenon of belief. The view in question is called the "downstream only" view, for reasons that will become apparent as we progress. I will end by defending the "downstream only" view against two highly influential objections. One objection is the objection from "strong normativism." The other is from the claim that non-belief states can explain actions in the same way that beliefs do.

1. Why Might (Some) Delusions not be Beliefs?

On what grounds have theorists denied that delusions are beliefs, and what kinds of responses have typically been made in response to such views? We will quickly look at this as a way of giving an idea of the debate that the discussion in this chapter is contributing towards.

1.1. Evaluative and Constitutive Norms

It is useful to see the DD as raising a particular instance of a general problem, namely, the problem of distinguishing between evaluative and constitutive norms; in other words, distinguishing between doing ϕ badly or well (evaluative) versus doing ϕ and not doing ϕ at all (constitutive).[1] In this instance, ϕ-ing is believing and the anti-doxasticist thinks that the delusional subject is breaking constitutive norms of belief. Unlike the doxasticist, who thinks that delusion is bad believing, the anti-doxasticist thinks that something has gone so badly wrong (or wrong in such a way) that the subject is not longer believing at all.[2] In spite of her utterances, the delusional subject doesn't actually believe the content of her delusion.

To get a grasp of the distinction between evaluative and constitutive norms, consider an analogy with chess (a real favourite among philosophers).[3] If, during a game of chess, I move so that my queen can be captured by my opponent with impunity (and as a result she gains an enormous material advantage and goes on to win the game) that is a *bad* chess

[1] I mean "doing" in the weakest sense. It does not imply action or intentionally doing something.

[2] The parenthesis "or wrong in such a way" is to highlight that one might not think of this as a matter of degree. Some theorists, in a way analogous to asking "How irrational must one be in order to count as delusional?" may ask: "How irrational must one be in order for one to fail to be a believer?" Others, rightly in my opinion, will see the believing/not-believing boundary as a difference in type rather than degree.

[3] Wedgwood (2002), for example, uses the chess analogy.

move. An evaluative norm (the norm of playing chess well) is broken. If I move my bishop like a rook, however, that's not a chess move at all, let alone a bad one. A constitutive norm (dictated by the rules of chess) is thereby broken. The doxasticist will say that delusion is like the former. The anti-doxasticist will say that it's like the latter.

This analogy with chess, however, has limited utility and it is important to locate where the analogy breaks down. Belief is like chess in that it is subject to norms; some of the norms will be essential to what belief is, others will feature in evaluations of beliefs. In both cases, the constitutive norms being adhered to are a precondition of the evaluative norms coming into play. A move obviously can't be a bad chess move if it's not a chess move at all. And the same goes for belief. But the parallel stops there. Disanalogies, for example, include the fact that chess is a game, the playing of which is utterly deliberate. A move in chess is an *action*, rendered intelligible by beliefs and goals (by informational and motivational states). Given this fact about chess, you can deliberately play badly if your goal is to do so. As we will see, it is not obvious that this can be said of belief, at least not directly.

1.2. The Anti-doxastic Position

So why might, and why *have*, some theorists denied that delusions are beliefs? The main proponents of the anti-doxastic position are Gregory Currie and his collaborators (Currie 2000, Currie and Jureidini 2001, Currie and Ravenscroft 2002).[4] Their view comprises of two claims: a negative and a positive claim. The negative claim tells us that delusions aren't beliefs (Anti-doxasticism). The positive claim tells us what delusions are, namely, they are imaginings that are mistaken for beliefs (What Bayne and Pacherie 2005 call "The Metacognitive View"). In the next chapter I examine the positive claim, but here I will focus exclusively on the negative one.

[4] We will see that there have been others, e.g. Egan (2009), who have argued for similar positions along similar lines.

Currie thinks that delusions ought not to be counted as beliefs because, although they have the superficial trappings of belief, namely, sincere assertion (and perhaps some behaviour that is in keeping with the delusional assertion), they lack the right kind of functional role. Since, according to Currie, delusions

(i) are not supported by evidence in their initial formation,

(ii) are not open to review in the face of contrary evidence, and

(iii) do not fully guide action and reasoning,

they should not be counted as beliefs at all. (And, as mentioned, should be counted rather as imaginings that are mistaken for beliefs.) The negative claim that delusions aren't beliefs can, in principle, be opposed on the basis of two different kinds of considerations:

A. Conceptual – you can disagree with the constitutive norms of belief that are put forward by Currie and co.

B. Empirical – you can disagree that, as a matter of fact, patients infringe these norms.[5]

B would ideally require a fleshed-out story concerning the aetiology of the delusion and the behavioural dispositions of the deluded subject. Note also that a critique could clearly make use of both A and B-type considerations, both disagreeing with the norms, and with the portrayal of the deluded subject. This combination of the two is exactly what we'll do in the next chapter when we introduce B-type consideration about actual cases of delusional misidentification, on the basis of the aetiology argued for in Chapter 3. In this chapter, however, we focus on A-type considerations. We need to get clear on what makes something an instance of belief.

People with doxastic leanings might accept Currie's constitutive norms of belief and respond with exclusively B-type considerations. To (i) they respond that delusions may be based on evidence of a sort, namely, on strange experiences (e.g. as we saw, lack of affective

[5] This is a rather obvious, and general, point. If someone says that a certain phenomenon is F, you can disagree with him or her either by disagreeing with their characterization of the phenomenon or by questioning their concept of F.

response to familiar faces in the case of Capgras, cf. Maher 1974, Ellis and Young 1990).[6] In response to (iii) they claim that, although delusions often fail to generate the kinds of actions and emotional responses one might expect, the Capgras delusion (for example) leads to violence against the impostor in 18% of cases, and sometimes of a particularly gruesome sort.[7] Granted, it is much harder to explain away (ii) since the delusions are (and, as we saw in Chapter 1, some, including the DSM, say *by definition*) highly resistant to correction in the face of contrary evidence. However, perhaps one can claim that the experiential evidence in favour of the delusion is so strong that this resistance to correction is not irrational since the experience trumps all possible testimony; we just don't know how weird these subjects' experiences are (see Reimer 2009). But, in any case, these are all descriptive, B-type, issues about what these patients are actually like. Although in the next chapter, we will be concerned with B-type consideration (focusing in particular on the Capgras delusion) we will focus in this chapter on A-type considerations, namely, on the constitutive norms of belief. We are concerned with this question: What would *any* case have to be like in order to qualify, or fail to qualify, as a case of belief?

In order to answer this, we need to ask: "What is belief?"

2. Reflections on the question "What is belief?"

When asking the question "What is belief?", what are we doing, exactly? What kind of question are we asking, and how do we know when we have answered it correctly? It is very important for us to distinguish different versions of this question, and to isolate the one that is of most interest.

2.1. The meaning of the word "belief" as used in English

[6] Recall that Maher hypothesizes that "delusional belief is not being held "in the face of evidence strong enough to destroy it," but is being held because evidence is strong enough to support it." (1974, p.99).

[7] De Pauw and Szulecka (1988) tell of a young man who decapitated his stepfather, taking him to be a robot, in order to look for batteries and microfilm inside his head.

Analytic philosophy has held belief to be central to philosophy. As Bertrand Russell put it:

> Belief […] is the central problem in the analysis of mind. Believing seems the most "mental" thing we do, the thing most remote from what is done by mere matter. The whole intellectual life consists of beliefs and of the passage from one belief to another by what is called "reasoning". Beliefs give knowledge and error; they are the vehicles of truth and falsehood. Psychology, theory of knowledge and metaphysics revolve about belief, and on the view we take of belief our philosophical outlook largely depends. (Russell 1921, Lecture XII)

This eloquent paragraph seems to be a far cry from what ordinary people seem to be referring to when they use the word "belief". "Belief" and its cognates are in fact very rarely used on a daily basis (as Hacker 2004 points out), and when they are used they do not mean what Russell means. Here is a sample of phrases we might expect non-philosophers to utter in everyday life, and which differ from the philosophical use in illustrative ways:

(1) "I don't *believe* that this glass is in front of me: I *know* it."

(2) "I stand by my beliefs"

(3) "He believes that the keys are in the box on the mantelpiece, but they're not."

(4) "Is that an elm, you ask? I *believe* so."

From these we can ascertain the following. "Belief" is often used as an expression of conviction about something that is (1) uncertain, and/or (2) defining of one's character or values. Not only does this (as seen by (1)), therefore, pick out a narrower class of phenomena than Russell's term, insofar as it *excludes* cases of knowledge or certainty, but it is (as seen by (2)), potentially in conflict with it: insofar as "beliefs" about values can be seen as not being the “vehicles of truth and falsehood”, they are not beliefs in Russell’s sense.[8] Furthermore, when we use “belief” to talk about the epistemic states of others (see (3)), it is used to imply lack of truth or accuracy. We might therefore say that it is used to imply lack of objective accuracy in others (see (3)) and lack of subjective certainty in oneself (see (1) and (4)). We often use “belief” to mean “*mere* belief”. This is certainly subject to a plausible explanation in terms of pragmatics (see Kauppinen 2010 for a careful exposition of this). Roughly put, use of “belief” is only conversationally relevant when it means “mere belief”, because if you were certain you would either state the claim directly (“Yes, that is an elm”) or use “know” (see (1)). Similarly, in many everyday cases, when the beliefs of others are true you also use “know” (“He knows where the keys are”), or attribute correctness (“He’s right”).

So, “belief” as used by everyday English speakers does not pick out Russell’s “central” notion. There is, of course, another fact that suggests that Russell’s central notion has little to do with how the word “belief” is used. It is that there are many languages that don’t have direct analogues of the English word “belief” (German is one of them, and in French, translations are available but not always satisfactory (e.g. “croyance”)). It would hardly be charitable to think that Russell was not aware of this. Russell is clearly not implying that “the philosophical analysis of mind” should only be practiced by English speakers (or by speakers of a language that expresses an equivalent concept). Clearly what Russell has in mind is something altogether different. He is drawing our attention to a *phenomenon* of central importance, not merely to a *word*.

What we mean when we ask, “What is belief?” in this chapter, is something closer to what Russell had in mind. In any case, it certainly shouldn’t have as a desideratum that it should match our daily use of the word “belief”. Now that this is out of the way, let us move on.

[8] Of course, questions about the truth-aptness of evaluative beliefs and discourse take centre stage in the non-cognitivism (sometimes called “expressivism”) debate in meta-ethics. We will not need to take a stance on this.

2.2. Belief Attribution and the Metaphysics of Belief

So we've seen that when we ask whether delusions are beliefs, we shouldn't look at everyday usage of the word "belief". We should look for a phenomenon that self-conscious and theoretically informed language users, viz. philosophers, might call "belief". What philosophers take this phenomenon to be varies enormously, and varies depending on their theoretical concerns and the debates that they are engaged in. Frege, for example, was motivated to put forward a view of belief that could account for shared belief, and in large part, as a way of understanding how mathematics (as an illustrative subset of human communicative activity) was possible. As Blanchette (2012) puts it:

> The fact that thinkers from vastly different contexts can nevertheless agree or disagree, can prove the same theorems […] has to do, on the Fregean picture, with the fact that belief, disbelief, proof, and so on, are all essentially about timeless and language-independent thoughts. (p.9)

These "thoughts" now tend to be referred to as "Fregean propositions", and were taken by Frege to be the bearers of truth and falsehood. Frege's notion of belief plays a similar theoretical role to Russell's notion, but is fleshed out differently (Russellian propositions were literally composed of objects and properties in the world).[9]

More recently, especially in philosophy of psychology and cognitive science, belief has played a very different, and more psychologistic, explanatory role. Belief has been used to explain, not so much how people can agree, disagree, and express truths in language, but

[9] Frege famously expressed puzzlement, in his correspondence with Russell, at the latter's attribution of truth and falsity to propositions which he took to be literally composed of concrete worldly entities such as Mt. Blanc.

rather to explain how they reason and act.[10] Furthermore, not only is it theorists who use belief for these explanatory enterprises, these theorists also take it to be the case that healthy humans do so in everyday life (see Stone and Davies 1996). In other words, belief it is not a theoretical entity like a quark, that plays a role in a scientific theory, but which humans have no use for (or indeed understanding of) in daily life. It is a concept or phenomenon that plays a central role in pre-theoretical, everyday, human reasoning and activity. What is meant by this is that human beings perceive each other not just as other human-shaped objects, subject merely to the laws of mechanics, but as agents with their own informational perspective and set of motivations. This capacity for attributing mental states (and emotions) to others is called "theory of mind" or "mindreading", and develops in normal children at the age of 3 or 4 (there is dispute about whether it may develop earlier).[11] This attribution of belief, of course, has nothing to do with mastery of the English word "belief", and children from all linguistic communities, including those that have no word analogous to the English word "belief", still engage in the attribution of informational perspectives to others. The reason we call this "belief" attribution is simply because this informational perspective is put in terms of "belief" by theorists working in English (The motivational perspective is put in terms of "desire", hence why one often hears talk of "belief-desire psychology").

What philosophers take this "belief" to be differs enormously, along a number of dimensions. For example, there is dispute over the metaphysics of belief, or the metaphysical commitments of belief attributions. We can divide these into three rough groups. I will go over these very quickly, since I think that issues about the norms of belief (the conditions under which we should ascribe belief) are prior to, and largely independent of, such metaphysical considerations. However, as Bortolotti (2009, p.2) rightly points out, "a good

[10] One highly influential framework is the representational theory of mind. Encouraged by advancements in early computer science, some theorists thought that beliefs were to be thought of as stored mental representations that played a certain role. Different theories exist as to the format of these representations, but the most influential was certainly the Language of Thought Hypothesis. This gained support thanks to its ability to explain the "productivity" and "systematicity" of thought.

[11] There is also evidence of some cross-cultural variation in the development of theory of mind (Lillard 1998).

account of how belief ascription works will impose constraints on the type of things that can play the role of beliefs." So, I now quickly skim over the three broad views about the metaphysics of belief.

First, there are *realists* who think that beliefs literally exist. Particularly robust forms of realism take belief-attributions to be referential (in the sense that they refer to things called beliefs) and are literally true if they talk about beliefs that are actually there, and characterise them accurately. The arch-realist is Jerry Fodor, who takes beliefs to be sentence-like mental representations that are stored in the head. We'll see more on views like this in the next chapter.

Then there are *instrumentalists*, like Dennett and Davidson, who think that beliefs are part of an indispensable explanatory framework, but that there are not literally things called beliefs (Dennett 1969 claims that belief-talk is literally true, but non-referential). This also entails a certain indeterminacy about belief attribution. Belief attributions are accurate to the extent that they play an adequate explanatory role, but where two belief attributions play equivalent explanatory roles, there is no fact of the matter about which is correct (This is mostly associated with Davidson (1973), in whose work it is presented as an extension of Quine's indeterminacy of translation).

Then there are *eliminativists*, who think that our everyday belief attribution, which they call "folk psychology," is a useful shortcut, but it is rough-and-ready, and fails to pick out anything resembling a unified, scientifically respectable, kind. An upshot of this is that, although the folk will continue to use the rough-and-ready heuristics of folk psychology, and have no choice but to do so, *scientific* psychology should eliminate all talk of belief.

Realists, instrumentalists and even, in some sense, eliminativists, can in principle agree about the conditions under which it is appropriate to say of someone that they believe something (there is divergence, of course, in thought experiments).[12] What interests us in this

[12] A nice thought experiment from Schwitzgebel (2010) is as follows. " "Rudolfo", say, emerges from a spacecraft and integrates seamlessly into American society, becoming a tax lawyer, football fan, and

chapter are the conditions under which we can appropriately say of someone that they believe something, however this so-called believing is metaphysically underpinned. This is not a metaphysical issue.

2.3. Revisionism and Conservatism about belief

The divergence of greater relevance concerns the meta-philosophical desiderata for any attempt at spelling out the conditions for when someone believes something. This divergence might be phrased in terms of *revisionism* and *conservatism*. A conservative will take a desideratum for the conditions of accurate *theoretical* belief attribution, to be that they should align with our actual practices of belief attribution. A revisionist, on the other hand, will deny that this is the desideratum. Rather, belief, *in the relevant theoretical sense*, is something with a clearly definable nature that is independent of our practices of belief attribution. As a result, there may be cases where we attribute beliefs, but the subject doesn't really believe what our attributions suggest. This is not simply the claim that there may be inaccuracy in belief attribution. Even the conservative will grant that we make mistakes. Rather, unlike the conservative, the revisionist will leave it as an open possibility that some of the methods we *regularly* use to attribute beliefs are inaccurate. Indeed they may be systematically inaccurate. This clearly relies on belief being more than a mere "folk-psychological construct", since in order for revisionism to make sense, there has to be something, like a nature or correct theoretical characterisation, against which our folk-psychology is to be assessed (i.e. from which it can deviate or to which it can adhere). One therefore cannot be an eliminativist and a revisionist about belief. The extent to which one can be an instrumentalist while also being a revisionist will depend on the views one has

Democratic Party activist. Even if we know next to nothing about what is going on inside his head, it may seem natural to say that Rudolfo has beliefs much like ours."

concerning the explanatory role played by belief, and the metaphysical commitments of such explanatory frameworks.[13]

A very strong, and explicit, example of revisionism about belief, is to be found in the work of Ruth Barcan Marcus (most explicitly in her 1995 paper). She claims that belief, in the relevant sense of central theoretical importance, is a relation, not to sentences (or propositions conceived as sentence-like), but to possible states of affairs, and, as a result, she claims that one cannot believe the impossible.[14] This is deeply revisionist since we commonly ascribe impossible beliefs to others. For example, mistaken beliefs about identity are often ascribed as, "He believes that Batman is not Bruce Wayne", and it is impossible (true in no possible worlds) that something should fail to be self-identical.[15] But according to Marcus, no one can actually believe this, since you cannot stand in the belief relation to impossible states of affairs. We will not need to take a stance on this tangential debate about believing the impossible (for the record, I sympathise with Marcus, see Wilkinson (forthcoming)). However, it is a nice illustration of what I mean by revisionism, namely, of our belief attributions regularly getting things not quite right.

I said that the divergence between revisionism and conservatism has a crucial impact on our concerns. So let's ask: What impact might revisionism or conservatism have on the question "What is belief?" and, by extension, "Are delusions beliefs?"?

2.4. An Illustration of Conservatism

[13] Dennett is a prime example of an instrumentalist and a revisionist. This is why he is often accused of treading an untenable line between instrumentalism and realism. This accusation, in my view, is misguided once we understand his earlier views about the commitments of theoretical (explanatory) discourse (in short, he takes belief-talk to be true, but "non-referential", in the sense of referring to a concrete entity) (see Wilkinson (in press) for a discussion of this).

[14] She is open to thinking about propositions in terms of possible worlds. But although her view has all the same consequences of such a view, she claims that she is more comfortable with the vocabulary of states of affairs.

[15] Note, also, the potential implications for philosophy of mathematics: when a mathematician makes a mistake, we intuitively say that they used to believe something that was impossible (namely a mathematical falsehood). This is ruled out by such a view.

Lisa Bortolotti (2009) has recently written a book entitled, appropriately enough, *Delusions and Other Irrational Beliefs*, which is devoted to defending the claim that delusions are beliefs. Her tactic is to present the anti-doxasticist with a dilemma: if we are to deny belief-status to delusions, then we are going to have to do the same for many states that, intuitively, we are happy to think of as beliefs. Among these "mental states that we are happy to regard as typical beliefs" (p.57), she cites, in particular, the sort of biased hypothesis-testing we commonly get in scientific beliefs (p.148) and also the unrevisability of racist (p.150) and religious (p.152) beliefs. Note that this argument is based on the suggestion that denying belief-status to these non-pathological states, that we normally attribute as beliefs, is too great a theoretical cost. In other words, it takes as a desideratum for a theoretical characterization of belief, that it should align itself with the beliefs that we attribute on a daily basis. In other words, a characterisation of beliefs that denies belief-status to delusions also has as a consequence that religious, scientific and racist beliefs are not "really" beliefs, and this is an unacceptable consequence. This is clearly conservative in the sense just mentioned.

This is an intuitively appealing strategy. However, note that the anti-doxasticist can counter such a strategy by affirming a revisionism about belief, and simply question that denying belief-status to, e.g., certain religious "beliefs", is too great a theoretical cost. The anti-doxasticist might say, as Ruth Barcan Marcus has, that it may turn out that many of the times that we regularly attribute beliefs to others, we are wrong and they are not, strictly speaking, believing in the proper, revised, sense. Furthermore, one might think that this is a perfectly acceptable, or even attractive, consequence of a *philosophical* theory of belief. What is philosophy for, after all, if not to regiment and revise our pre-theoretical concepts? We may ascribe belief in a rough and ready way, but that doesn't mean that belief is a rough and ready phenomenon.

Bortolotti's work nicely shows that, given the continuity (in particular in terms of irrationality) between clinical delusions and non-pathological beliefs, if the anti-doxasticist position is going to be plausible, one has to be a revisionist about belief. What this means for us is that, in order to test the anti-doxasticist position, in its most acceptable form; we need to put forward a revisionist view of belief. Such revisionism may entail unpleasant consequences, like denying belief-status to certain religious or racist beliefs. I personally don't find this denial, in and of itself, unacceptable, as long as one is clear about what one means by belief, and why. However, as we are about to see, I do find the characterization of the phenomenon of belief in the light of which current anti-doxasticists frame their arguments generally unattractive.

Therefore, we need to not look at our practices of belief attribution, but rather look directly at what the phenomenon of belief, in the relevant sense, might look like. If it then turns out that some of the things that we think of as beliefs are not really beliefs, then so be it. Furthermore, note that *if* we can have an attractive, revisionist, view of belief, and still be doxasticists, then that is, *at least in purely argumentative terms*, more damaging to the anti-doxasticist than Bortolotti's position.[16] The anti-doxasticist will not have recourse to claims of revisionism, since we will already be on the same page (in that respect, at least).

So, let's recap quickly. We have looked at what we are interested in when we ask, "What is belief?" We are putting aside how we use the word "belief" in everyday life. We are also putting aside our pre-theoretical attributions of beliefs. We are rather looking at whether there is a phenomenon called belief that can be adequately characterised.

3. The Relevant Notion of Belief: A Downstream-Only View

In this section, I will intuitively characterise a phenomenon that we might have reason to call "belief", and in the light of which a revisionist view could be held. I don't want to claim that it is the only characterisation available. However, I do think, firstly, that it

[16] Bortolotti's conservatism may, of course, be attractive for other reasons.

is an attractive one, and, secondly, that it is the one that is most relevant to the question of whether delusions are beliefs. What is most interesting, and philosophically problematic, is whether delusions are beliefs in this sense of belief. Since I need this view of belief to be at least defensible, I will, after characterizing it, defend it against two major and influential threats.

3.1. Information vs. Motivation: Direction of Fit

What kinds of entities are at least candidate believers? One way of answering this question is by considering the teleological question: “What kinds of entities *need* something like belief?” One intuitive response is to say that organisms need information about the world, and hence are candidate believers. Not every organism is a believer, but the claim that organisms believe and, say, rocks don’t is at least a start. Then we might want to narrow this down a bit further to organisms that move. Entities that have no means of operating in the world have no need for an informational view of it. Then we might want to narrow it down further to organisms that move autonomously in the service of their goals. Animals navigate environmental features in the service of their goals, and presumably they do so by having some kind of information about the world, perhaps even something we could characterize as an informational perspective on the world.[17]

Frank Ramsey (1931) put belief precisely in these terms. He called beliefs “the maps whereby we steer”.[18] This metaphor of beliefs as being like maps that help us through the world in the service of our goals is extremely useful. In particular, it picks out an opposition between the phenomenon of belief (viz. the map) and the phenomenon of desire (viz. the

[17] Whether informational states in an artificial system, i.e. a robot, can be thought of as beliefs is an interesting question, but one that is not strictly relevant for our purposes. I would suggest that good reasons for opposition to this view would not be something intrinsic to the informational states themselves, but rather because the goals are not theirs. They are the goals of the programmer. Artificial systems only have derived goals, namely, the goals of their creators. Evolved beings have goals of their own.

[18] What Ramsey meant by “we” is not entirely obvious. He could have meant “human beings”. I’d like to think (and it fits in with his view, and the apparent modal strength of his claim) that he meant “any beings capable of belief”.

goal). The former involves *taking* the world to be a certain way, whereas the latter involves *wanting* it to be a certain way. This picks out a fundamental distinction between information and motivation.

This opposition between belief (information) and desire (motivation) has sometimes been put in terms of "direction of fit" (which in turn is derived from speech act theory, where it was used to elucidate the difference between assertions and commands). The first use of the term "direction of fit" was by J.L. Austin (1962), and was not in fact used to pick out this distinction at all.[19] His student, John Searle, picked up the terminology, and used it to pick out a distinction that he claims is best illustrated in Anscombe (who, in fact, did not use the "direction of fit" terminology). Searle's merging of Austin's terminology and Anscombe's example has stuck, and has been picked up by various philosophers since. The example Searle (1979) takes from Anscombe is one that distinguishes a shopping list, from a detective's record of what the man doing the shopping is buying:

> "…if the [shopping] list and the things that the man actually buys do not agree, and if this and this alone constitutes a mistake, then the mistake in not in the list but this man's performance (if his wife were to say: "Look, it says butter and you have bought margarine", he would hardly reply: "What a mistake! We must put that right, and alter the word on the list to "margarine"): whereas if the detective's record and what the man actually buys do not agree then the mistake is in the record." (Anscombe 1957, p.56)

So, the shopping list, which has a "world-to-word" direction of fit, is analogous to a desire, which has a "world-to-mind" direction of fit. The detective's inventory, which has a "word-to-world" direction of fit, is analogous to a belief, which has a mind-to-world direction of fit. It is in the nature of belief to fit with how things stand in the world. If there is

[19] Austin used it as a way of explaining different ways of asserting Fa (and given this difference, the logical form is insufficient to capture it). This is seen most clearly when we assert something of the logical form Fa which is false. Compare wrongly calling a triangle a square, which is "committing an act of violence to the language", with wrongly describing a triangular object as being square, which is "committing a act of violence to the facts". This difference is explained in terms of direction of fit since in the former the fault is in "fitting a name to the item", whereas in the latter it is in "fitting the item to the name".

a disparity, it is the "fault" of the belief, so to speak. Conversely, it is in the nature of desire to motivate one to behave in such a way that will make the world fit with it. If there is a disparity, it is the "fault" of the world, so to speak.

Although this example is nice for illustrating direction of fit, I find Ramsey's map metaphor more useful for understanding the essence of belief.[20] Although not much hangs on this, it is worth mentioning why I have this preference. Note that the shopping list and the inventory both have the same format: both are lists, with discrete linguistic items (and we will see, in the next chapter, further reasons to dislike this way of thinking). This suggests both a format for belief that I find unattractive, and a similarity in format between belief and desire, which I do not find plausible. Firstly, you have beliefs all the time, effortlessly, about everything you perceptually encounter. You can't help it. To put it another way, you are not (usually) informationally neutral about the things that you encounter. You are, however, *motivationally* neutral about the vast majority of what you encounter. We have appetites and goals. However, the vast majority of things (in the broadest sense to include objects and states of affairs) that we encounter aren't the objects of those goals. Some of them might well be relevant to attaining some goal or other; others won't even have *that* significance. Now consider the aptness of the map analogy. Suppose you want to get to point X, and realise that you'll need to take such and such a route, through a tunnel, over a bridge etc. These are things that impact on you getting to point X. However, in looking at the map, you will also expect to see a church here, a forest there. These may be irrelevant to you getting to point X, but you still take them to be there.

Secondly, and in a related vein, beliefs are holistic, and necessarily subject to consistency in a way that desire doesn't seem to be. If I believe that you are shorter than 6ft,

[20] Beliefs with modal or normative content may be hard to capture in literal maps, for example. However, they only qualify as beliefs if they guide my actions, and *in this respect* are map-like. If I truly believe that it might rain, I will, ceteris paribus, bring my umbrella. If I truly believe that killing babies is very wrong, I will, ceteris paribus, refrain from doing it. Perhaps, in any case, we can make sense of maps with normative and modal content. A street on a map with an arrow indicating one-way traffic has normative content. One with a warning sign that flooding at high tide is a risk has modal content (of sorts).

I will also by that same token believe that you are shorter than 7ft, or 8ft etc. Maps, like beliefs, are holistic in this way. If there is a boundary between a forest and a field, and I update it on the map, making the forest bigger, I *ipso facto* make the field smaller. We do however seem to have conflicting desires all the time. I want to taste my friend's delicious chocolate brownies, but I know that they'll make me fat and I don't want to get fat. That's when decision-making gets tough. There are, as we will see in the next chapter, cases where believing in humans isn't maximally *consistent*. It is arguably, however, not possible for it to be directly *contradictory* (and in cases where it appears to be so, this can be explained away).

Furthermore, note that a map also has a map-to-world direction of fit. If there is a mismatch, it is the map's "fault". So the map analogy preserves the insight of direction of fit, without the infelicities of the list analogy that we have just mentioned. This notion of direction of fit has been picked up in two slightly different ways, and it is important to disambiguate. Under one, direction of fit distinguishes the *informational* from the *motivational*, whereas in the other, it distinguishes the *cognitive* from the *conative*. Sometimes these two oppositions are taken to be equivalent, but an illustration of the two distinctions will clarify the difference. Let me first clarify the two uses, and their different consequences.

Velleman (1992) uses direction of fit in the second sense:

> The term "direction of fit" refers to the two ways in which attitudes can relate propositions to the world. In *cognitive* attitudes, a proposition is grasped as patterned after the world; whereas in *conative* attitudes, the proposition is grasped as a pattern for the world to follow. The propositional object of desire is regarded not as fact -- not, that is, as *factum*, having been brought about – but rather as *faciendum*, to be brought about: it's regarded not as true but as to be made true. (1992, p.8)

The sense of "regarded" in "regarded as true", is a weak sense. As a result, Velleman's sense of "cognitive" includes any representational states that present the world as being a certain way, *regardless of whether the subject actually takes it to be that way*. Thus, Velleman includes under the class of states or attitudes that are "cognitive", and hence have mind-to-world direction of fit, states like perceptions, hypotheses, fantasies, imaginings etc. Indeed an attitude can have that direction of fit, and persist, even if the subject is totally aware of the mismatch between the state and the world.

The other use of "direction of fit" distinguishes, not the cognitive from the conative, but the informational from the motivational. This is the sense of direction of fit that John Searle has in mind. One very nice characterisation of it is by Michael Smith (1987). I cite it in full, and it rewards close reading.

> …the difference between beliefs and desires in terms of direction of fit comes down to a difference between the counterfactual dependence of a belief and a desire that *p*, on a perception of *not p*: roughly, a belief that *p* is a state that tends to go out of existence in the presence of a perception that *not p*, whereas a desire that *p* is a state that tends to endure, disposing a subject to bring it about that *p*. Thus, we may say, attributions of beliefs and desires require that different *kinds* of counterfactuals are true of the subjects to whom they are attributed. We may say that that is what a difference in their direction of fit *is*. (1987. p.54)

Such a use of direction of fit seems to *define* belief, *qua* basic informational state, *in terms of* direction of fit. Therefore, unlike Velleman's use of "direction of fit", states like imaginings, hypotheses etc. do not have a mind-to-world direction of fit.[21] As Gendler (2011) nicely puts it:

> Since to believe that *P* is to take *P* to hold of the actual world, and since which world is the actual world is (in the relevant sense) not up to us, belief's task is to conform to some pre-ordained structure—its *direction of fit* is mind-to-world. By contrast, since to imagine *P* is to take *P* to hold of some particular (set of) non-actual world(s), and since which worlds are imagined *is* (in the relevant sense) up to us, imagining's task need not be to conform to some pre-ordained structure.

I will follow Searle, Smith, Gendler, and others, in using direction of fit in this way, and in characterising belief in terms of the informational, understood as that which has a mind-to-world direction of fit in his sense.

Another way of expressing the same thing as this sense of mind-to-world direction of fit is in terms of *transparency*. This is a term that gets used in various different ways, but the way in which I want to use it is as an expression of mind-to-world direction of fit. Beliefs are, by their nature, tied to (or simply are a function of) how the world is as far as the agent is concerned. As Gareth Evans puts it, belief is "directed outwards." If you discover that *p* isn't the case, you *ipso facto* cannot belief that *p*.[22] This is the same as direction of fit, since a belief that *p* will "go out of existence in the presence of a perception that *not p*". Imagination that *p* is not transparent in this sense, and can certainly remain in existence "in the presence of a perception that *not p*".

This has various revealing consequences for questions concerning whether you believe something. Let us contrast this with two other mental phenomena, for example, with

[21] Indeed it is not clear whether in this sense these states have an essential direction of fit at all - they may have elements of a mind-to-world direction of fit, to the extent that they have elements of *belief*, to the extent that they are *used as information*.

[22] See also Moore's famous paradox. Although the point is very similar, however, it is phrased in semantic rather than epistemic terms.

desire (which has a world-to-mind direction of fit) or with an emotion (which *prima facie* has no direction of fit). If I ask you whether you are angry at the moment, you will need to assess yourself, for example, what you are feeling. However, if I ask you whether you believe that *p* (in the philosophical sense of "believe") you don't, or at least shouldn't (unless you take me to be using "belief" in the everyday sense), look at yourself, as if trying to accurately describe an internal state (as you might with anger). Rather you simply ask yourself whether *p*.[23] This is obvious for perceptual beliefs, but it also extends to non-perceptual beliefs. If you ask me whether I believe that Paris is the capital of France, I ask myself whether Paris is the capital of France, not whether I believe it.[24] I examine the world, not myself. Belief is directed outwards, namely at the state of affairs that would make it true or accurate. These "outward" states of affairs can obviously include states of affairs involving one's body, and one's "self", e.g. one's personality.

So, belief is the basic informational state, believing is the basic informational phenomenon, and this involves taking the world to be a certain way. The question that interests us here is: What does one have to do (in the weakest sense of "do") in order to count as taking the world to be a certain way? It is with this question in mind that I present my revisionist view about the norms of belief. For lack of a better name, I call it a "Downstream-only View".

3.2. Introducing the "Downstream-only View"

[23] Velleman uses "transparency" to refer to a different phenomenon. He claims that the question "Whether to believe that p" is, "transparent to", namely, answered by answering the question "Whether p". We might call this *deliberative transparency*. The transparency I am talking about here is what we might call *transparency of self-attribution*. The later tells us something more fundamental about the nature of belief. Furthermore, although I don't need to get embroiled in this tricky debate, I don't agree with *deliberative transparency* since I don't believe in *doxastic deliberation*, namely, we never ask whether to believe that p, since belief is involuntary in the strongest sense (due to its nature, and in all possible worlds).

[24] Although in everyday speech, people talk about "beliefs" as the inconclusively supported commitments that help to define them as a person: e.g. "I stand by my beliefs". This is clearly not what I mean. People may well be asking questions about themselves when asked what they "believe" in these cases.

The best way to introduce the “Downstream-only View” is to draw attention to two things that we have already mentioned, and to see the connections between the two. First, recall our distinction between constitutive and evaluative norms, namely, the distinction between norms that dictate doing ϕ badly or well (evaluative) versus doing ϕ and not doing ϕ at all (constitutive). Second, recall the “functional idiosyncrasies” that lead Currie to claim that delusions aren’t beliefs. Delusions, according to Currie:

(i) are not supported by evidence in their initial formation,

(ii) are not open to review in the face of contrary evidence and

(iii) do not fully guide action and reasoning.

Now, let us think of the functional idiosyncrasies as being instances of breaking norms, namely, norms dictated by how a good belief ought (evaluative) or any belief must (constitutive) function. The norms that (i-iii) break, divide into “upstream” and “downstream” norms. That delusions are not supported by evidence in their initial formation, i.e. (i), and not responsive to review in the face of contrary evidence, i.e. (ii), both break upstream norms, namely, norms that are related to causal or evidential precursors of the belief ((ii) may seem to be breaking a downstream norm, since it is something that is exhibited after the belief has been formed, but since in attempting to correct the delusion, you are providing *input* for new judgement it is in fact upstream). That delusions don’t guide action or subsequent reasoning, i.e. (iii) breaks a downstream norm (namely, norms that govern the consequences of the subject’s belief). Note that any such belief-like state (like the religious commitment that “God is three and God is one”) can be divided in this way. We can talk of how the subject came to be in such a state (upstream), and the use that the subject makes of the state (downstream).

Given what we have said about belief’s transparent, map-like, nature, what can we say about upstream and downstream considerations and their role in determining whether something is a belief? In other words, which of (i-iii) break constitutive norms, and which break evaluative norms? I suggest that there is a very simple pattern. Upstream norms are evaluative, whereas downstream norms are constitutive.

This claim, as we are about to see, is certainly controversial among philosophers, but it is ultimately correct and defensible. To reiterate: the claim is that, whereas upstream considerations are relevant for telling us whether the subject is believing well or badly, they are not relevant for telling us whether the subject is believing at all. In other words, to move back from talk of norms to talk of functional role, these upstream issues about sensitivity to evidence describe the functional role of *good* belief, not belief *tout court*.

This is, rather obviously, called the "Downstream-only View" because it is the view that a mental state (or propositional attitude) can be formed in any old way, but as long as it has the right kinds consequences (namely for action and dispositions for action) the state qualifies as a belief.

The relevant notion of belief for our purposes is concerned with whether an organism (in this case a delusional subject) takes the world to be a certain way, and if it does, what that way is and how best to characterize it. It is not strictly relevant (although it is clearly an interesting question) how it is that the organism came to have this belief. Indeed the idea of a mad scientist manipulating your brain so as to "implant" a certain belief is not prima facie ruled out by our philosophical concept of belief.[25] This would be a prime case of irrational (or perhaps non-rational, depending on how you cash out epistemic irrationality) belief, but nonetheless it would count as belief, for example, if it were acted upon in the right way.[26] One might think that one doesn't need farfetched thought experiments to illustrate this point: the indoctrination of extremist cults, for example, is action-driving in the worst ways

[25] As we will see in the next chapter, for me, this would be not the implanting of a discrete sentence-like entity, but the "implanting" (or, better, instilling) of a particular disposition to action.

[26] Some would say that you needn't go as far as farfetched thought experiments: the brain-washing of extremist cults, for example, is behaviour-driving in the worst ways (e.g. suicide bombing), but the belief-formation process is far from rational.

(e.g. suicide bombing) and gives rise to strongly held beliefs, but the belief-formation process is far from rational.[27]

Returning to the map metaphor, if we want to know whether the subject is using a map-like drawing as a map, we need to look at the subject's dispositions to behave in accordance with the drawing. The presence of "disposition" is important, for, as we touched upon, there may be parts of the map, say distant parts, that are irrelevant to the subject's goal-directed navigation, but the subject can still take them to be accurate, and this is shown by the fact that *if* the subject *were* to be in those distant parts, he *would* use the map to navigate (e.g. would expect a river here, a mountain there). If, for some reason, the map was drawn using a bad method (analogue: unjustified belief) or a good method but which unfortunately went awry and was consequently inaccurate (analogue: unluckily false belief), then the map is a *bad* map, but still used as a map if the subject is inclined to navigate by it. Recall that it is merely a metaphor: I am not saying that belief is a map, or even necessarily map-like in format (indeed we will see in the next chapter that we need to be careful about talk of "representational formats" generally). However, the notions of maps and navigation are useful ways to think about belief and the metaphor is particularly useful here, in the sense that the sort of questions that we ask for ascertaining whether something is being used as map, are the same sorts of questions that we ask to ascertain if something is a belief.[28]

The consequence that the downstream-only view has for the question of whether delusions are beliefs is clear. If we want to claim that delusional patients don't really believe what they assert, it is *not* to resistance in the face of contradictory evidence, or to lack of

[27] Of course, there is the important of matter what they *intend* by this action, and hence what the *content* of their belief is. Do they believe that they will go to heaven? That they are doing the morally right thing?

[28] There is an illustrative limit to the map/belief analogy. Maps are typically things that you deliberately *use* to help you navigate. Beliefs are not things that you *use*: they are things that you *have* when you are in the process of navigating. Thus, it is more helpful to think of maps in the analogy *as models* rather than *as literal maps*. In other words, the subject isn't going through a process of using something map-like to navigate. Rather, the agent's informational state, as manifested in her goal-directed navigations, can be *modelled* (roughly, helpfully described) in terms of something map-like.

supporting evidence during formation, that we should turn. This tells us whether they are believing *badly* (and arguably presupposes that they are *believing*). Rather we need to show that the deluded patient is, all things considered, failing to act or reason in accordance with her professed beliefs. Only then should we say that she doesn't truly believe what she appears to. The deluded patient would then be (but for very different reasons) like the boss who claims that he believes in the equality of the sexes, but then goes on to employ a demonstrably sexist hiring policy. In short, we need to find downstream norm breaking, downstream functional idiosyncrasies, such as a disparity between action and utterance, between word and deed. The fact that delusions are unsupported (viz. i) and are resistant to counter-evidence (viz. iii) cannot serve to rob them of belief-status.

3.3. Examples of Retrospectively Revisable Belief Attributions

One upshot of this particular revisionist, downstream-only, view of belief is that there are times when our belief attributions are, to use Marcus's terminology, "retrospectively revisable". This is not simply a case of the attributor correcting their attribution, but rather of the *theorist* (with the benefit of a God's eye view if you will) claiming that an attribution is inaccurate. Since belief attributions are usually expressed in language, and they are evidentially grounded in linguistic behaviour, in particular, assertive utterance, it is useful to look at certain examples where attribution, made by others or oneself, on the basis of assertion is retrospectively revisable, in other words, when what is asserted in not *truly* believed.

3.4.1. Straightforward Deception

The most obvious, indeed banal, case of such retrospectively revisable belief attributions is in a case of straightforward deception. In such a case, we attribute belief to someone on the basis of what they say. In a sense, we trust them. However, it is important to see that there are different levels of trust. The relevant trust here is that their assertions are an accurate representation of their belief state. We need not trust in the truth of the belief state.

Consider the following example. Suppose we are both standing in front of a rickety-looking bridge. I'm hesitating to cross, and you say: "Oh, go on. The bridge is safe." I can trust that you believe the bridge is safe, and also trust *what you say*, namely, that the bridge is safe. Alternatively, I can trust that you believe the bridge is safe, but have doubts about your judgment, namely, have my own doubts about the safety of the bridge. A third option is that I don't trust that your assertion "The bridge is safe" is an accurate representation of your belief-state.[29] Perhaps, you're trying to get me killed. I might then say "You go first", and if you were to show serious hesitation that might confirm my doubts.

No one has claimed that assertion is sufficient for belief. A more plausible sufficient condition is *sincere* assertion. The next two cases illustrate the insufficiency of sincere assertion, too.

3.4.2. Lack of self-knowledge

Since what I believe is simply how I will be disposed to act in the relevant circumstances, I can clearly be wrong about this. I don't seem to have any infallible access to my dispositions to action. I am, usually in a better position that you are to judge my own dispositions, but that's because I have more evidence to go on (feelings, rehearsals in mental imagery etc.), but I can certainly be wrong. There are cases when one is surprised or disappointed by one's own dispositions. A boss who claims that he believes in the equality of the sexes, and indeed actually believes that he believes this, but then who goes on to exhibit a demonstrably sexist hiring policy, may come to the realization: "Gosh, perhaps I am sexist after all.[30]"

[29] This could be built in part on my judgements about the bridge, for example, it could be because it seems so obvious that the bridge is not safe

[30] Indeed, there is evidence from brain-damaged patients that suggests considerably deeper and more shocking evidence of lack of self-knowledge. Gazzaniga (1995) describes the following case of a

One of Currie's important insights is not only that sincere assertion that *p* is insufficient to guarantee belief that *p*, but also that wrongly believing that one believes that *p* will be sufficient to produce the sincere assertion that *p*. The assertion "*p*", on a framework (which we will explore in more depth in Chapter 5) where linguistic action has no special status but it just another kind of action to be explained in terms of the informational and motivational states of the asserter (if sincerity is stipulated) only guarantees that the subject believes that they believe that *p*. It doesn't guarantee that they believe that *p*. In other words, the assertion that *p* is not an expression (in the sense where "ouch" is an expression of pain) of the belief that *p* (i.e. it does not reveal the belief that p). It is a self-attribution, which is liable to go wrong.

3.4.3. Lack of linguistic knowledge

One's vocalised self- attribution (which can take the form of "I believe that *p*" but more likely will simply take the form "*p*") will be conducted in natural, public, language, and as such will be open to the evaluations of public language. Thus one can be accurate in terms of self-knowledge, but still sincerely assert something one doesn't really believe, because one has an inadequate mastery of the words one has used. I can see that my neighbour has bought a harpsichord, but be labouring under the misapprehension that a harpsichord (which I can recognise just fine, I know what it *is*, I'm just mistaken about what it's *called*) is called a "clavichord". Thus I say, in all sincerity, "My neighbour has bought a clavichord". In an important sense, this is a mistaken self-attribution of belief, even though it

split-brain patient, namely someone who has undergone a surgical procedure (for the treatment of epilepsy) involving the severing of the *corpus collossum*, which connects the two brain hemispheres. With split-brain patients, certain stimuli can allegedly be presented to one hemisphere without it being "accessible" to the other. In this example, the subject fixated his eyes on a point straight ahead, while two cards were shown, one positioned to the left of fixation (which would be available only to the right brain) and one to the right of fixation (which would be available only to the left brain). Then the instruction, "Walk!" was shown to the right brain, and the subject got up and began to walk away. When he was questioned as to why he had got up, he replied, "I'm going to get a Coke from the house." What is striking is that this intention (which can be understood as the desire to get a coke and the belief that there is a coke in the fridge) is self-attributed with fluency and sense of obviousness. And yet it is clearly confabulated.

involves no lack of self-knowledge. I am perfectly right in attributing to myself the belief that my neighbour has bought *that thing*; I just wrongly think that that thing is called a clavichord.

3.4.4. Lack of understanding

There is another important example where belief attributions and self-attributions are retrospectively revisable. In such a case, the subject makes a claim that they don't understand. If believing something is to be disposed to act as if that something were the case, then one obviously can't be disposed to act as if something were the case if one doesn't know what that something is. So the Christian who claims that "God is three and is also one", may think that they believe this, think that they believe what the sentence expresses, but they don't since they don't know what it expresses. One cannot be disposed to behave as if a state of affairs obtained if one has no idea what that state of affairs is. A more accurate attribution would be that they believe that the *sentence* "God is three and is also one" is true. States of affairs can also involve linguistic entities, and their properties (including, in some sense, the property of being "true").

It may be useful to comment on religious beliefs, in passing. Religious beliefs, as Bortolotti (p.152) points out, "are a good example of mental states that play a role in people's mental economies and influence people's behaviour, but are not necessarily responsive to evidence". I agree with Bortolotti that many religious beliefs are indeed beliefs, and that their belief-status is not threatened either by the fact that they are formed in little evidence, or that they are resistant to correction. It is not upstream, but rather downstream considerations, that count in favour of, or against, belief status. However, I do think that many religious beliefs (common, healthy ones; not religious delusions) don't quite have the content that they are taken to have, and this is because of their consequences for action. Religious "beliefs", accepted articles of faith, may cause ritualistic behaviour, the assertions of certain sentences, but if the person isn't disposed to act as if that were the case, that is not literally believed. They believe that a sentence is true, but not what it expresses (or they believe that a certain ritual is somehow worth undergoing).

3.5. Two Objections to the Downstream Only View

We have presented the downstream-only view. However, it is open to two compelling, but ultimately misguided, objections. The first objection is an objection from strong normativism. This objection is essentially a straightforward contradiction of the downstream-only view, which is grounded in the intuition that belief "aims at truth". The second objection is the objection from alternative explanations. Belief can't simply be dispositions to actions, since there are other states that intuitively aren't beliefs but that explain action in the same way. Hence we need upstream considerations to tell these belief and non-belief states apart. I will present and respond to these two objections in turn in sections 4 and 5 respectively.

4. Objection 1: Strong Normativism and Aiming at Truth

The objection from strong normativism involves the claim that belief, by its nature, "aims at truth", or is constitutively subject to the "truth norm". How this applies to the question of whether delusions are beliefs is fairly obvious: in order for something to be a belief, it must aim at truth. Delusions typically don't aim at truth therefore they aren't beliefs.

We can break up this objection into three steps:

1. Beliefs by their nature, aim at truth.

2. Aiming at truth is an upstream consideration (failing to aim at the truth infringes an upstream norm).

3. Therefore the downstream-only view is mistaken

I will state my response to this first, and then flesh it out. It is as follows.

1 is ambiguous, in particular with regards to the notion of “aiming at truth”. In one sense of “aiming at truth”, 1 is true. However, in the sense in which 1 is true, 2 is false. Conversely, in the sense in which 2 is true, 1 is false. Namely, once “aiming at truth” is conceived as an upstream consideration, it ceases to be constitutive of belief. I will now flesh this out by looking at some interpretations of 1.

4.1. Believers “aim at truth” in believing

So, in what sense can we mean that belief, by its nature, aims at truth? Well, let’s start by highlighting what *cannot* be meant by this.

What cannot be meant by this is that the believer, in believing, literally, directly, tries to have beliefs that are true. This can’t be true because belief is involuntary. The believer can’t aim at *anything* in believing, or at least not *directly*. If one is not convinced that belief is involuntary (although in my view, one should), there is another argument against this reading of the truth-aim claim, which doesn’t rely on involuntarism (but which does provide indirect support for it). If the truth-aim claim, sometimes referred to as the “truth norm”, is seen as an aim or norm that can actually guide the subject’s beliefs then it suffers from the following problem (due to Gluer and Wikforss 2010). Consider other norms or aims that guide behaviour. For example, consider a rule that a stock-market trader might use: “buy low and sell high”. In order for this to actually guide the subject, the subject needs a belief about the current state of the stock market (namely, that what one is considering buying or selling is low or high). Now consider the norm: “Only believe something if it is true”. In order for this to guide belief formation, you already need to have a belief about whether the thing in question is true. But then you already have a belief about the matter in question. Therefore the norm “comes too late” to guide belief. Needless to say, I take this to fall out of transparency. And involuntarism also falls naturally out of this, since belief is not an action, not something that the agent does (and hence not something the agent can be guided in doing), but rather a *precondition* for action.

4.2. Believers "aim at truth" by regulating for truth

Perhaps what is meant is something more *regulative*. So, to think of an analogy, I can't directly control my cholesterol, but I can *aim* at lowering my cholesterol indirectly by observing a cholesterol-lowering diet. Similarly, with belief, I can't directly try to believe things that are true, but I can aim at truth indirectly by observing good epistemic practice, for example, by not interpreting or gathering evidence in a biased way etc. However, although this is now a coherent way of understanding the truth-aim claim as something that the *subject* aims for, namely, she aims for it in a regulative way, we now plausibly lose the *constitutive* force of the claim. And the claim is supposed to tell us what is constitutive of belief. What we now have is rather an evaluative, rather than constitutive norm. Biased evidence gathering, failure to "regulate" for truth, leads to *bad* believing, but not to failure to believe altogether.[31]

What we might say, however, is that it is *constitutive* of belief to have certain *evaluative* norms. This seems plausible. But then it is not constitutive that the believer aim at truth, but rather that she *ought* to. The consequences that this has for a belief-like state that is the product of biased reasoning, or that is poorly epistemically grounded, and even which is unresponsive to evidence, is that this is believing that is *bad*, that the subject ought not to have, but it is still belief. This is perfectly consistent with a downstream-only view.

[31] This regulative truth aiming may, however, be a constitutive norm for a more active evidence-gathering process, like "inquiry". In a strong, but not outlandish sense of "inquiry", somebody who treats evidence in a biased way may not be inquiring badly, but failing to inquire at all. Is someone who plays chess, put tries to lose, playing chess or not? Well, yes and no. This is not a substantive issue, but a terminological one: do you include the attempt to win in the concept of "playing chess"? Does it matter either way?

4.3 Belief *qua mental state* "aims at truth"

One last interpretation, which follows on from what we have just said, is that belief *itself*, the mental state, in some metaphorical sense *aims at truth*. This avoids problems about involuntarism, and about the norm coming "too late". So, what is meant by this metaphor? We can see it all the more clearly when we compare it to other cognitive mental states (by "cognitive" I mean, as before, as opposed to "conative": the world is represented as being a certain way rather than to be made that way). Consider Currie's own contrasting cognitive state of choice, viz. imagining. An imagining that misrepresents the world, or is not grounded in evidence, is not "bad" in any way. One can say that this is because imagining (the state itself) is under no obligation to "aim at truth". As Velleman (2000) puts it, unlike imagining, "belief is the state that goes right or wrong by being true or false". What this seems to suggest is that it is not the believer who aims at truth, but in some metaphorical sense, the *state itself*, which by its very nature "aims at truth" (and that in turn constitutively entails an evaluative norm on the part of the believer). This falls out of the state's direction-of-fit, understood in the transparent, informational sense. If this claim is true, it is trivial and perfectly in keeping with a properly understood downstream-only view. Belief aims at truth insofar as it is how the subject takes the world to be, but it still doesn't strictly matter for belief status how the subject came to take the world to be that way.

One final ambiguity that needs ironing out, which is present, both in talk of "transparency" and "aiming at the truth", is whether we are talking about objective, *de facto* transparency or truth-aiming, or merely *as far as the subject is concerned*. I think it is fairly obvious that we mean the latter.[32] If we meant the former, we would have to deny belief-status to mistakes (faultless, unlucky mistakes, as well as biases). And yet, mistakes get all their import from being doxastic, from being beliefs (which amounts to them being transparent). A mistake, a misrepresentation of reality in imagination, is not problematic because it isn't purporting to represent reality, and so typically wouldn't be acted upon in the

[32] And arguably this is captured in the metaphor of "aiming". It is not, for example,

service of relevant goals. So, to clarify, when we say that belief aims at truth, we mean truth as far as the subject is concerned. When we say that it is transparent, directed at the world, at reality, we mean the world, reality, as far as the subject is concerned. To use Ramsey's metaphor, we want to know what the subject's map is like, and an inaccurate map, even a systematically inaccurate one, is still a map. We shall see in the next chapter that this is vitally important for understanding delusions.

4.4. Recap of the Response to Objection 1

As I mentioned, "aiming at the truth" is ambiguous. If this means that the agent aims at the truth, then either this makes no sense, if the aiming is supposed to happen directly, or it does make sense and is regulative. If it is regulative, it fails to be constitutive of belief. Another interpretation of "aiming at truth" is that the state itself aims at truth. This in turn has an implausible normative reading, and a more plausible descriptive reading. This descriptive reading, is merely a consequence of its direction of fit, and is perfectly in keeping with the downstream-only view. Surely the normativist isn't saying simply that if a state hasn't got a mind-to-word direction of fit, then it isn't a belief. That's trivial, and no one would deny that. It's simply saying that something isn't a belief if it isn't a belief. What the downstream-only view allows is that there are states with that direction of fit that seriously infringe upstream norms. There are belief that are held for bad epistemic reasons, or even no reasons at all, and they are still beliefs because they have that direction of fit and will have the relevant consequences for disposing the subject to act.

And this leads me on to the next objection. There are states that explain action in the same way as beliefs do, but which aren't beliefs.

5. Objection 2: Non-belief states Explain Actions

This objection runs as follows. People often act in ways that aren't best explained in terms of beliefs, but in terms of states that aren't beliefs. How do we distinguish them from beliefs? The answer is: by using upstream considerations, such as their relationship to evidence etc. Therefore the downstream-only view is false. Delusions, as Currie and collaborators (Currie and Jureidini 2001), and Egan (2009), would have it, are more like these non-belief states, than like beliefs, and this is in large part due to their relationship to evidence.

What are these non-belief states? Let's go through them and their explanatory role. We will divide them into two groups. In the one group, let's place those that are referred to using pre-existing words in English. In another let's place those that are referred to using terms of art. We will examine these in order.

5.1. Imagining, Pretending and Believing

The criticism goes as follows. There has to be more to belief than mere dispositions to action, since there are states that fall short of belief, like imaginings and pretendings, but which have similar behavioural consequences. Thus, when a child is having a pretend tea party, we don't attribute the belief that she's pouring tea, even though the behaviour is similar. We say that she is imagining that there is tea in the teapot, and engaged in pretence. We say similar things about actors on stage. We might say that what makes these cases of imagining and not belief is the way that they are related to evidence. The child is not believing that there is tea since she can see that there isn't tea. If you were to try to correct the child's behaviour by pointing out that there is no tea in the tea pot, she'd quite rightly be somewhat confused by your attempt to correct her. This is because it imagination and pretence is not responsive to evidence. One might think that beliefs, contrary to this, are responsive. If the child did believe that it was tea, and you pointed out that it was merely water, she might taste it, realise that you are right, and revise her beliefs. But the child is not believing. She is imagining or pretending.

5.2. Alief and Bimagining

I'll introduce this with the opening example from Gendler's well-known paper. In March 2007, a glass walkway, suspended 4000 feet above the Grand Canyon, was opened to the public. Most tourists struggle to make it to the centre of the walkway; others are paralyzed with vertigo.

> How should we describe the cognitive state of those who manage to stride to the Skywalk's center? Surely they *believe* that the walkway will hold: no one would willingly step onto a mile-high platform if they had even a scintilla of doubt concerning its stability. But alongside that belief there is something else going on. Although the venturesome souls wholeheartedly *believe* that the walkway is completely safe, they also *alieve* something very different. The alief has roughly the following content: "Really high up, long long way down. Not a safe place to be! Get off!!" (Gendler 2008a, p.635)

Gendler has used this and several other examples to motivate putting forward the well-received view that there are belief-like states, that are not beliefs, but which have to be included into our taxonomy of mental states for explanatory reasons (of the same kind that motivate the presence of belief in such a taxonomy).[33] Roughly, to have an *alief* is to have "an innate or habitual propensity for a real or apparent stimulus to automatically activate a particular affective and behavioural repertoire, where the behavioral propensities to which an alief gives rise may be in tension with those that arise from one's beliefs" (Gendler 2008b, p.553). As we will see in more depth in the next chapter, Andy Egan (2009) posits a similar hybrid (or "in-between") state, "bimagination", to explain roughly similar phenomena

[33] Alief has been quite widely accepted since Gendler's introduction of it, e.g. McKay & Dennett 2009, pp. 499-500; Bloom 2010, chs. 6 and 7; Hodge 2011)

(including, as we have briefly seen, delusions). Bimagination has some of the functional role of belief and some of the functional role of imagination: it is both action-guiding (and hence belief-like) but resistant also to counter-evidence (and hence imagination-like.)

There has been some dispute about whether, and to what extent, *alief* is explanatorily necessary or helpful. In particular, critics, (see e.g. Mandelbaum forthcoming) think that the relevant phenomena can be explained by beliefs and desires coupled by interference from subpersonal processes (e.g. cued responses, conditioned or innate, to certain stimuli). Although I think my preferred view of belief (and its role in explanation) may be conducive to supporting such criticisms, such criticisms are not strictly relevant to defending the downstream-only view. The point is, rather, that Gendler takes these cases, and her solution to these cases by using aliefs, to oppose, and cause serious problems for something she calls a "Neo-behaviorist view" (2008a, p.653). Gendler characterizes such a view as follows:

> … to believe that *P* is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which *P* (together with one's other beliefs) were true. (2008a, p.652)

She says that the real problem for this view is that fact that "a wide range of attitudes – among them acceptance (Bratman), imagination (Currie, Velleman) and pretense (Doggett & Egan, Velleman) – may motivate P-concordant behavior" (2008a, p.653). She wants to add herself and alief alongside these philosophers and attitudes just cited. What distinguishes these states from being belief that P cannot simply be the way that they motivate behaviour, since, by stipulation, they motivate the same behaviour (namely "P-concordant" behaviour). What makes these states differ from beliefs is that belief has a certain relationship to evidence (it is, as Gendler puts it, *reality-sensitive*) that these other states lack. At this point it should be clear that accepting that belief requires a certain relationship to evidence, and not simply action, amounts to giving up the downstream-only view of belief. And, of course, that is the view that I am trying to defend.

I will respond to the objections from imagination and from alief in reverse order, so, starting with alief. The same central principle guides my responses to both, namely, an important distinction between action and behaviour.

5.3. Responses to Objection 2: Action and mere behaviour

Let Gendler have aliefs, and let's suppose that they are worth positing. What is my response to her critique of "Neo-behaviorism"? Firstly, a note on "Neo-behaviourism." If by "Neo-behaviourism" she means simply "behaviourism that happens to be held today", then I would not call the downstream-only view, "behaviourism". And yet, I entirely agree that her characterization of "Neo-behaviourism" (see quotation above) is the downstream-only view. So what's going on?

Behaviourism, as Chisholm (1957) rightly pointed out, is a reductive project. It aims to rephrase, *salva veritate*, claims about the mental in terms of claims about the non-mental, and, in particular, the behavioural. Thus, claims about belief are rephrased ("analysed", if you will) in terms dispositions to behaviour. However, the downstream-only view is not a reductive enterprise in this sense at all. We are not concerned here, for example, with a metaphysical project of trying to accommodate the mental in a physical world. We are simply trying to state the necessary conditions for someone to count as a believer. Thus we rephrase claims about believing in terms of claims about *action*. But action is a mental phenomenon. We are not being reductive. We are elucidating one mental phenomenon in terms of another mental phenomenon. We are picking out necessary relations between them. Well, what is the purpose of this here? So that we can tell when someone appears to be believing something when in fact they are not. To take an analogy, saying that, by knowing what someone desires and how they are disposed to act, you can ascertain what they believe, is like saying that you can ascertain velocity from knowing distance covered and time taken.

That is not a *reduction*. It does not "reduce" velocity to distance and time any more that one can "reduce" distance to velocity and time. You may have views about which is metaphysically more fundamental, but the relations between the phenomena will hold independently of this. Similarly, you may think that action has metaphysical primacy, or that belief has, or desire has. But that is a different issue.

Now note that Gendler's charaterization of "Neo-behaviourism" isn't reductive *either*. In it, she talks of disposition to *action*. But then she goes on to talk about P-concordant *behaviour*. She slips from talk of action to talk of behaviour, without marking the difference. And yet there is an enormous difference. If she had characterized neo-behaviourism in terms of behaviour it would be reductive and, I agree, erroneous. However, *action* is, at least in part, individuated by intention, not by mere bodily motion. Although there has been a great deal of debate surrounding the precise individuation conditions of actions, here we only strictly need to note something that all views agree on: that the same bodily motions performed with different intentions are different actions. (That is partly why action is explained, in the sense of "explain" we put forward in Chapter 2, rationally and not causally.) Behaviour, on the other hand, is individuated by bodily motion. A result of this is that two identical pieces of outward behaviour can be different actions. They will be different actions if the intentions are different, and the intentions will be different if the beliefs or desires are different. We can contrast cases where intentions are lacking, or when they are different.

Suppose, in Scenario1, John kills (in a neutral sense of being causally implicated in his death) James, because Joe, who wanted to frame him, implanted a microchip in his brain and controlled his behaviour. In Scenario 2, John kills James as a way of exacting revenge. We can stipulate that the outward behaviour is the same in both scenarios. However, the two scenarios are very different. In Scenario 1, we wouldn't think of John's causing of James's death as an action, at least not John's action. We might even plausibly call it Joe's action. This would be reflected in our moral and legal practices. Given all the relevant information, we would blame Joe for James's death and charge him with murder. In Scenario 2, we simply have the straightforward action of John murdering James. So, actions are different actions (sometimes, as we saw, attributable to different *individuals*) if the intentions are different. Bodily movement, without any intention, fails to be action at all.

One query, worth addressing now, however, is: how do we make space for the notion of unintentional action? This doesn't seem like a contradiction in terms, and yet it seems that on the view just sketched, this is ruled out. In order to respond to this, it is important to distinguish the action, from the way the action is described. As Davidson (1980) influentially put it, in order for something to be an action, it has to be "intentional under some description". Thus, to take an example from Searle (1983), it seems intuitively plausible that Oedipus's marrying Jocasta is an action. An action performed with the intention of marrying a beautiful woman whom he loved. It was not performed with the intention of marrying his mother, although that is *de facto* what he did. So we might say that he unintentionally married his mother. However, note that it would fail to be an action at all if there was *no* description under which it was intentional. Due to our ignorance, we often perform actions successfully, but with unintended consequences.

So here is my diagnosis of what aliefs show for the downstream-only view. Gendler is right to say that these states are not beliefs. However, this is not, as she says, because they are not supported by evidence or inflexible in the face of contrary evidence. There are, as I argue (and as Bortolotti (2009) was at pains to point out) many beliefs that are resistant to correction (although they should only be thought of as beliefs if the subject is disposed to act on the in the right way). Aliefs fail to be belief – or rather, subjects in the examples do not believe the "content of the alief" – because these states *causally explain behaviour* rather than *rationally explain action*. Aliefs are affective states that *interfere* with people's intentions and actions, rather that taking part in them in the way that beliefs typically do. The person reluctant to walk on the walkway, wants to walk on the walkway, believes that it is safe, but is hindered by a strong sense of fear. The person reluctant to drink from the bedpan

(Gendler 2008a, p.636), may want to drink from the bedpan (presumably the experimenter gives them some incentive to do so), believes that it is safe to do so, but is hindered by a strong feeling of disgust.

Now let's look at imagination. As we saw, Gendler's objection to "Neo-behaviorism" is confusing the individuation conditions of action and mere behaviour. The same response applies here. Part of what makes something imagination or make-believe, rather than belief, is not, as Currie emphasizes, that it's not responsive to epistemic reasons, but rather that the action on the basis of which you ascribe the merely imaginative episode in question is different from the action on the basis of which you would ascribe a state of believing (even if the *behaviour* is superficially the same). An actor on stage who runs away from something dangerous in the play doesn't believe that he is in danger, but rather believes that he is acting in a play, wants to give the audience the impression of being in danger, and that this is how that's done. He therefore isn't literally performing the action of running away from danger, however convincing he may be in his acting. Performing the action of running away from danger requires that one have the belief that one is in danger. Otherwise one is not performing *that* action, but merely something that superficially resembles it

Conclusion

In this chapter, I sought to delineate the notion of belief that is most relevant to the doxasticity debate. The relevant notion of belief is simply the basic (personal) information state, the state of "taking the world to be a certain way." What matters when we ask, "Are delusions beliefs?" is whether the subject is in the relevant informational state. Namely, we are asking whether the subject really takes the world to be the way in suggested by her delusional assertions. What will matter for that, as it does for when we ask of any subject whether they really believe something, is not how the belief was formed, but how she is disposed to act in the relevant circumstances. Such a view constitutes what I have labelled a downstream-only view of belief. We saw two powerful and influential, but ultimately misguided, objections to such a view of belief.

The first objection was from the fact that beliefs have to "aim at truth". This is ambiguous, and in the sense in which it is an upstream consideration (and hence an objection to the to downstream only view), it is merely evaluative and not constitutive (and hence fails to be an objection to the view). In the other sense, where it is merely a claim about direction of fit, this is not an upstream consideration at all, since, however the subject takes things to stand in the world will necessarily be taken into account in the relevant dispositions to action.

The second objection was from the claim that other states that fall short of belief (e.g. imaginings, aliefs) explain, and relate to, action in the same way. Therefore a state cannot distinguish itself as a belief purely on the basis of its relation to action. My response to this relied on a widely accepted point that actions are partly individuated by their intentions. Thus, the "actions" explained by aliefs aren't actions at all, and the actions explained by imagination and make-believe are very different actions from the ones that are explained by the relevant beliefs in the contrasting cases where the outward behaviour is the same.

This downstream-only view points towards a general lesson concerning evaluative and constitutive norms of belief. The doxasticity debate (DD) often talks about belief-status as if it were a matter of a threshold. If a belief is bad *enough*, irrational *enough*, it will cease to be a belief, and will be something else. But that is not how these norms work. If the constitutive norm is broken, then the evaluative norms for that phenomenon cease to apply. If only evaluative norms are broken, they can be broken as seriously as you like, and the phenomenon will still count as that phenomenon. Applied to belief we get: as long as the subject is disposed to *act* upon it in the right way, it is still a belief.

CHAPTER 5

Is the Capgras Delusion a Case of Belief?

Introduction

This chapter completes the task set up by the previous one, namely, that of examining and answering the question “Are delusions beliefs?” (viz. the “doxasticity debate”, or DD for short). In order to focus the argument, we will ask this question specifically about the Capgras delusion, although lessons of more general application will be extracted.

You will recall that in the previous chapter I made the claim that only downstream considerations can be appealed to in denying belief status. Upstream considerations tell us if something is a case of bad belief, and *nobody* denies that delusions, if they are beliefs at all, are bad ones. However, upstream considerations cannot tell us if something fails to be a case of believing at all. It’s not as if there’s a threshold after which something is a case of believing that is so bad, that it no longer counts as believing. What matters is whether the subject is in the state of taking the world to be a certain way; it doesn’t matter for belief status how the subject came to be in that state.

A result of this is that, in answering the question of whether a given case of delusion is a belief, we are less concerned with aetiology (e.g. with debates between bottom-up and top-down, one-factor and two-factor). If we took upstream considerations to matter for belief status, then aetiology would be potentially crucial since different aetiologies would, for example, render the subject epistemically irrational or reason-responsive to different extents. This would lead to belief status being attributable on the basis of some aetiologies (e.g. bottom-up one-factor accounts), but not others (e.g. top-down accounts). Indeed, as we mentioned in the last chapter, on many such views about the constitutive norms of belief, the aetiology of the Capgras delusion that I proposed in Chapter 3 would entail that the Capgras delusion is not a belief. However, given a downstream-only view, we are interested in how the delusional patient is disposed to act. Aetiology will have a role to play in determining these dispositions, but it won't have a direct role in helping us decide whether we grant or retract belief status. However the delusional state arises, we are interested in its *agentive effects*.

In the last chapter, I dismissed the importance of the metaphysics of belief for answering the question concerning the norms or conditions that make someone a believer. This question should indeed be answered first. However, once answered, it does have consequences of a metaphysical nature. Roughly, a downstream-only view about the constitutive norms of belief, although it doesn't strictly *entail*, is conducive to a certain view about the nature (viz. the metaphysics) of belief. Many labels could be used for such a view; I will call it *holistic dispositionalism.*

The aim of this chapter is to present and argue for *holistic dispositionalism* and see what happens when we approach the DD with it in mind. There are two consequences. One consequence is that the question "Are delusions beliefs" needs rephrasing, and, in particular, in a way that does not come naturally to the approaches of several influential participants in the debate. A second consequence is that, once the question is rephrased, the following can be said. Taking into account a distinction between egocentric and encyclopaedic believing, which I will characterise in more detail later, the relevant egocentric belief is more clearly attributable to the delusional patient than the commonly attributed encyclopaedic belief.

I will proceed as follows. I will show how the DD is approached by several (but admittedly not all) of the participants in the debate. For want of a better label, I will call this approach "The Standard Approach". The Standard Approach involves a fairly non-committal form of "discrete representationalism" about the nature of belief (we will see what I mean by this). I will then show how, within The Standard Approach, and given discrete representationalism, one can be a doxasticist, an antidoxasticist, and a semi-doxasticist. Then I will characterise an alternative view about the nature of belief, holistic dispositionalism, and show, firstly, how it is at odds with the Standard Approach, and, secondly, how it leads us to rephrase the central question of the DD (viz. "Are delusions beliefs"). Once rephrased (viz. to the question, "Does the delusional subject take the world to be the way her delusional assertions suggest?") I will explore how we might answer it with regards to the Capgras delusion. A key to answering it will require us to reflect more carefully on what the Capgras patient believes. This will be aided by a distinction between egocentric and encyclopaedic believing. I will end by drawing some conclusions about what the Capgras patient can be said to believe.

1. The Standard Approach

In this section we'll see that the Standard Approach either implies, or explicitly endorses, discrete representationalism about belief. We will also see that it encourages a certain way of approaching the DD, and that this allows theorists to be doxasticists, anti-doxasticist or semi-doxasticists with regards to "the delusional state."

1.1. The Standard Approach and Discrete Representationalism

The Standard Approach to the DD, although implicit in some of the aforementioned views, is very explicitly proposed, and endorsed, by Andy Egan (2009) in the following extract. It is such a concise and complete description of the Standard Approach that I cite it in full:

> "In what follows, I'll be concerned with characterizing (very partially) the attitude that delusional subjects bear to the contents of their delusions. I'm going to assume that the right way to go about this is to characterize the roles that particular, token mental representations play in the subject's cognitive economy. So I'll be assuming that

> some sort of minimal representational theory of mind is correct – that there is a medium of mental representation, and there are discrete representational items in the head. These representational items are operated on in various ways, and accessed by various systems, in order to regulate both our behavior and the maintenance of other representational items. Believing, desiring, imagining, and the bearing of propositional attitudes in general is a matter of having a representational item with the right kind of content, which plays the right kind of role in one's cognitive economy." (Egan 2009, p. 262)

So, the view of belief that we have here, discrete representationalism, and that we will abandon later, is that there is a "discrete representational item in the head", the delusional state, and if it plays the right kind of role, then it is a belief, and if it plays a different kind of role, then it is a different kind of state. The representational state in question is taken to have the content that is expressed in the standard characterization of the delusion, which in turn corresponds to what the patient sincerely asserts. So, for example, in the Capgras delusion, there is the representational state that has the content "My father has been replaced by an identical-looking impostor/stranger".[1] If that state plays a certain role, it is a belief; if it plays another role it is some other state. No surprise, then, that on this approach the question is phrased as "Are delusions beliefs?" or "Is this delusion a belief?" We will shortly see that this phrasing can be misleading.[2]

1.2. Doxasticism, Anti-doxasticism and Semi-doxasticism within The Standard

[1] See, for example, Ellis and Young (1990), among many others.

[2] Note the connection between a view of the constitutive norms of belief that includes upstream considerations, and a discrete representationalism concerning the nature of belief. If there is a state with a certain content, and it playing a certain role determines the kind of state that it is (viz. the "attitude"), then it might be theoretically appealing to build into that role (e.g. for belief) both upstream and downstream requirements. This amounts to building a kind of functionalism about belief into discrete representationalism (although there are plenty of theorists who are functionalists without being discrete representationalists (e.g. Dennett, Stalnaker, Lewis), but they tend to, like me, focus on the downstream functional role).

Approach

Accepting this approach for the moment, why might someone think that the representational state doesn't play the right kind of role? Within this approach, in order to say that a given mental state, say, a delusion, is not a belief, one has to, (i) get clear on the functional role of belief and (ii) show that the delusional state lacks that functional role. This is precisely what *anti-doxasticists* like Currie and his collaborators do. As we have seen, since, according to Currie, delusions

(i) are not supported by evidence in their initial formation,

(ii) are not open to review in the face of contrary evidence,

(iii) do not fully guide action, reasoning, or elicit the appropriate emotional responses,

they should not be counted as beliefs at all, and, as just mentioned, should be counted rather as imaginings that are mistaken by the subject for beliefs. This mistake, which gives rise to a false second-order belief, accounts for the sincere delusional assertion. As Currie and Jureidini put it:

> Imaginings seem just the right things to play the role of delusional thoughts; it is of their nature to co-exist with the beliefs they contradict, to leave their possessors unwilling to resolve the inconsistency, and to be immune to conventional appeals to reason and evidence. (Currie and Jureidini 2001, p. 160)

Here we see an approach that resembles Egan's in the following way.[3] Beliefs and imaginings are *things*, mental representations that play a certain functional role.

Egan, however, is what we might call a *semi-doxasticist*. He claims that the functional role of the delusional state is somewhere between that of belief and imagination, and

[3] It differs from Egan's view, of course, in explaining the affinities that delusions have with beliefs *metacognitively* (viz. in terms of false second-order beliefs).

therefore calls it “bimagination”.[4] Indeed he makes a strong general claim that “there seems to be no principled reason to think that we can’t get a spectrum of cases, from clear, totally non-belief-like imaginings to clear, full-blooded, paradigmatic beliefs, with intermediate, hard to classify states in the middle” (p.272).[5]

Doxasticists (Bayne and Pacherie 2005) either claim that the functional role is that of belief, or close enough to that of paradigmatic beliefs, or that “the belief” has somehow been “compartmentalized” (viz. it is a belief that has somehow been prevented, in certain contexts, from playing the role that it otherwise would (e.g. Reimer 2011, Bayne 2011)). What I suggest in the latter half of this chapter has some similarities to this last view, although it has very different metaphysical commitments.

As you can see, on all of these views it is at least implied, and at most explicitly asserted, that a belief is a kind of thing, namely, an internal representation that plays a certain role. These conflicting views within the DD all adopt (and indeed engage so successfully with each other precisely because they adopt) the Standard Approach. Before we adopt holistic dispositionalism, and are thereby forced to abandon the Standard Approach, let me formulate the Standard Approach.

> *The Standard Approach:* There is a discrete representational state, “the delusion”, which has the content that is commonly taken to characterise the delusion (e.g. “My father /wife/etc. has been replaced by an impostor/stranger” in the case of the Capgras delusion) and that state is a belief if (and only if) it plays a certain functional, and if it plays a different functional role, it is a different state.

Keeping this in mind, let’s look at holistic dispositionalism about belief. This contrasts

[4] Compare “bimagination” with Gendler’s aliefs, which we discussed in Chapter 4.

[5] In the previous chapter I tried to present an approach on the basis of which believing (*in the relevant sense*) does play a fundamental theoretical role. On this approach it is not so much that there is a “principled reason” to avoid what Egan describes, but rather that it makes no sense if you don’t think of beliefs as discrete mental representations that play a certain role.

sharply with the discrete representationalism that is characteristic of the Standard Approach.

2. Holistic Dispositionalism about Belief

Several philosophers have held versions of this view (Braithwaite (1932–1933), Ryle (1949), Price (1969), Audi (1972), Stalnaker (1984) Baker (1995), Schwitzgebel (2002)). My preferred formulation is from Ruth Barcan Marcus.

> An agent believes that S just in case (1) under agent-centred circumstances such as desires, needs, and other psychological states including other believings and (2) external circumstances (3) the agent will act as if S obtained, i.e., will act in ways appropriate to S being the *case,* where S is a state of affairs, actual or non-actual. (Marcus 1995, p.126)

Now, I say that Marcus's view, which is the view I want to endorse, is "a version" of a dispositional view because there are dispositional views that differ importantly from it. Most relevantly, note how this view differs from a dispositional view that takes belief to involve dispositions to act as if a sentence were true. First, such a view seems to (arguably implausibly) rule out that non-human animals and pre-linguistic infants can have beliefs. Second, there are a number of ways of acting as if a sentence were true, without actually taking the world to be the way that the sentence expresses. As we saw in the previous chapter (section 3.4), sometimes this is a result of lack of doxastic commitment (deliberate or otherwise). The subject *de facto* doesn't believe because although she is disposed to act as if the sentence were true (viz. by asserting it), she isn't disposed to act as if the state of affairs described by the sentence actually obtained. Alternatively it can be the result of a lack of linguistic understanding. In such a case, the subject *can't* believe, because she doesn't know what that state of affairs described is, or what it would be for it to obtain.

Instead of being disposed to act as if a sentence were true, what we have in Marcus's formulation is talk of acting as if *a state of affairs obtained*. Now, the state of affairs can *include* linguistic entities like sentences, and these sentences can be taken to have the property of being true.[6] Indeed, this is how Marcus accounts for some of the *prima facie* implausible consequences of her view. For example, it is through this that she explains away our apparent ability to believe impossibilities that she so controversially denies (viz. this involves wrongly believing that *sentences* that happen to express impossibilities are true, not believing *the impossibilities themselves*). It is also through this that she accounts for so-called Frege puzzles (viz. they involve belief about what individuals are *called*, or which individuals were encountered at certain junctures, not belief that an individual somehow fails to be self-identical).

I call this view "holistic dispositionalism" because it takes belief to involve dispositions to action of all varieties (viz. verbal, bodily, and even, perhaps, mental actions), and these dispositions are best understood holistically. (Indeed, taking belief to essentially involve dispositions towards sentences injects a certain atomism into the picture, namely, the atomism of the sentences themselves.) The best way of fleshing out holistic dispositionalism is in terms of views that it opposes. One can make out two related, but dissociable, such views:

Language-centred accounts: Belief is a relation to linguistic entities. In particular, "sincere, reflective, non-conceptually-confused" (Marcus 1983, p.333) linguistic assent is taken to be both a necessary and sufficient condition for belief.

Discrete representationalist accounts: Belief consists in having a certain internal

[6] So one can believe that a sentence is true, without being able to believe what it expresses (for example, if one fails to understand it, but trusts the validity of the source).

representation that plays a certain role.

These two views are dissociable in the sense that one can consistently hold both or either. Namely, one can think of linguistic assent as being central to belief, *and* as being the expression of ("giving voice to") an internal representational state (viz. the belief). Or one can think of linguistic assent as being crucial to belief, but be neutral, or even skeptical, about internal representations (Davidson clearly holds something like this view). Alternatively, one can think of linguistic assent (in public language) as not being necessary for belief, but insist that belief is a matter of having an internal representational state that plays the right functional role. It is this commitment to discrete internal representations (which we saw so clearly in the passage quoted from Egan 2009) that is at the heart of the Standard Approach.

But first let's look at how holistic dispositionalism differs from language-centred accounts.

2.1. A Critique of Language-Centred Accounts: Belief without Language

We have already seen that holistic dispositionalism differs from a dispositional view that appeals to dispositions to act as if sentences were true. But why would one adopt holistic dispositionalism over language-centred views *generally* (i.e. whether dispositional or otherwise)? Following Marcus herself (1983, 1990, 1995), we can use the intuition that non-human animals have beliefs, namely, that there can be belief without language, as a springboard from which to critique language-centred accounts. Marcus gives us a nice example:

> A subject, call him 'Jean,' and his dog, call him 'Fido,' are stranded in a desert. Both are behaving as one does when one needs and desires a drink. What appears to be water emerges into view. It is a mirage […] Both hurry toward it.

> On what ground can we deny Fido a desire to drink, a belief that there is something potable there? […]
>
> The important kernel of truth in such a linguistic view is that arriving at a *precise* verbal description of another's beliefs and desires is difficult, and especially so when the attribution cannot be verbally confirmed by the subject. We will not go far wrong in attributing thirst to the dog Fido, or the belief that there is the appearance of something potable. Whether we can attribute to the dog the recognition of *water* would depend in part on whether dogs can select water from other liquids to roughly the extent that we can. That is an empirical question. (1990, pp. 134-135)

In essence, the view is as follows. Many animals exhibit autonomous goal-directed behaviour, and use their sensory and cognitive apparatus to inform themselves about the world. Language, as a public, communicative, symbolic medium is certainly vital for talking about belief (as it is for talking about anything), and hence is vital for *verbalized* belief *attribution*, but not for belief *per se* (nor, indeed, for non-verbalized belief attributions). These attributions will be accurate to the extent that they capture the animal's informational perspective. As a result, it is much easier to attribute beliefs to other human beings, not only because they can correct us, but also because their informational perspective is more likely to resemble our own. And insofar as our language relies on our informational perspective, it is unlikely to be able to capture the informational perspectives of animals.[7] However, in using a less fine-grained content ascription with non-human animals ("something potable" instead of "water") we minimize our risk of getting it wrong.[8]

As we've seen, language is not only useful for talking about and attributing belief, but can supplement the kinds of beliefs we are capable of having (among other things, we can

[7] Wittgenstein's remark "If a lion could speak, we would not understand him" might spring to mind here. I would extend this insight to the claim that, if a lion could speak, he would not be able to ascribe beliefs accurately to us.

[8] Note, also, that we may have similar, but less extreme, difficulties in accurately attributing beliefs to other human beings from very different cultures.

have beliefs *about* linguistic entities). However, that does not mean that without language there cannot be belief. To claim this, and thereby a discontinuity between humans and non-human animals (and indeed pre-lingual infants), is, as Marcus rightly states, "anti-naturalistic" (1995). Animals clearly take the world to be a certain way, viz. have an informational perspective, and can be wrong about how things stand in the world. Furthermore, there is no doubt that we naturally attribute informational and motivational perspectives to non-human animals. (Why else would we hide from a prowling lion, other than because we believe that it might see us, and that it might be motivated to harm us?)

2.2. A critique of beliefs as discrete internal representations

What if one grants that non-human animals have beliefs, but that belief still involves internal representations? They may even be language-like in some sense. Animals may not have a *public* language, but, if they are cognitively sophisticated enough to believe, they have a "language of thought" (Fodor 1975), and beliefs are simply sentences in the language of thought that are stored in the subject's head. Others have spent far more time directly criticizing Fodor's position (e.g. Dennett 1981). But one doesn't have to criticise the details of such a view if one finds that it simply misses the point. On such a view, postulating discrete representational vehicles that correspond to beliefs misunderstands the nature of belief, and confuses two very different questions we can ask about human believing. As Stalnaker 1993, neatly puts it:

> There are questions about what is said in claims that attribute thought to thinkers and questions about the psychological mechanisms in virtue of which those claims are correct; that is, there are questions about what it is to think, as contrasted with questions about how, in fact, we do it. (p.424)

If we narrow down "thought" to "belief" in this quotation, the point is as follows. Even if human beings, and animals, *happen* to have a language of thought, this does not seem to be central to what belief essentially is.[9] We can, it seems, imagine beings that exhibit autonomous goal-direct behaviour and react intelligently to their environment in the service of these goals, without anything like a language of thought. These beings would arguably still be believers.[10]

One might retort that there must be internal representations of some sort, even if they need not be language-like. They could be, for example, map-like. Egan, indeed, leaves this possibility open: "This sort of view takes no stand on just what the representations are like, so, in particular, they needn't be like sentences. They might be more like maps, or models, or something else altogether" (2009, p.262). But tweaking the format also misses the point.[11] There are perhaps all sorts of ways that subjects can be enabled to take the world to be a certain way. And, insofar as the subject can be wrong about how the world is, we might say (in a trivially broad sense) that there is representation "going on". But that does not mean that there are discrete things in the head called representations that are to be *identified with* beliefs. Belief is not that which enables a system to take the world to be a certain way, but is the phenomenon of taking the world to be a certain way *itself*. And we can think about that without any commitment to internal representations at all.

Belief is not, at least not essentially, a relation to a sentence, or to an internal representation (of whatever format), but is rather the taking of the world to be a certain way.

[9] Schwitzgebel (2006) makes a similar point: An alien, ""Rudolfo", say, emerges from a spacecraft and integrates seamlessly into American society, becoming a tax lawyer, football fan, and Democratic Party activist. Even if we know next to nothing about what is going on inside his head, it may seem natural to say that Rudolfo has beliefs much like ours—for example, that the 1040 is normally due April 15, that a field goal is worth 3 points, and that labor unions tend to support Democratic candidates."

[10] Indeed perhaps the best treatment of the Fodorian view is to take a historian's viewpoint. The language of thought hypothesis arose amidst advancements in computer science and solid symbol systems, and a related and ill-founded optimism that matter (the brain) could be given mind in a way analogous to the way in which computers were enabled to perform intelligent tasks, namely, by rule-guided operations on concretely stored internal representations. Even bracketing the question of the essential nature of belief, the computationalist paradigm has been shown, firstly, to not be the only way of getting matter to be intelligent (see e.g. Connectionist approaches), and, secondly, to not be a good imitation of human intelligence (e.g. it doesn't account for learning).

[11] Recall how the map analogy I used in Chapter 4 for not a point about representational format.

This will necessarily be manifested in dispositions to act (not simply *behave*) in certain ways in relevant circumstances. We can use different symbolic or formal tools to talk about or model these ways of taking the world, and we can do so with varying degrees of success. This success will depend on our explanatory concerns. Stalnaker (2007) expresses precisely this view:

> There is, alas, only a brain in the head — no pieces of information, no contents, wide or narrow. Propositions, whether they are Fregean thoughts, Russellian complexes, or sets of possible states of the world, are abstract objects that we [theorists] use to represent [namely, model] certain capacities and dispositions of people, and certain kinds of social relations between them [e.g. verbal communication].[12] (2007, p.109-110)

2.3. Talking about, and modelling, belief

So, subjects take things to be a certain way, they have an informational perspective, and we can talk about relevant elements of that perspective by using sentences in English. For daily purposes that usually works fine.[13] However, for certain theoretical tasks (e.g. addressing technical issues in philosophy of language, philosophy of mind or epistemology) using sentences in English (or any natural language) to talk about what people believe can be too vague or ambiguous to accurately capture the relevant elements of their informational perspective. Thankfully, better, sharper, tools can be developed. One popular such tool is the

[12] People often talk of the Fodorian sentences-in-the-language-of-thought as "propositions", in a way similar to the way that Frege's "thoughts" can be thought of as propositions. This suggests that Fodor and Frege are *competitors*, giving competing accounts of what propositional attitudes like belief *are*. But the Fodorian representations are not abstract entities used to *model* all the things that one might call belief (note the glaring omission of "Fodorian propositions" in this quote). They are taken to be real things that are stored in the head that deserve to be called beliefs. This is a claim of such a different status, that it doesn't even compete with different models of belief. Furthermore, as we are about to see, a Fodorian discrete representationalist approach certain doesn't capture what is interesting or important about the DD.

[13] In any case, it works well enough that a more precise method of talking about belief didn't naturally develop, and doesn't seem to be under any pressure to do so.

tool of possible worlds developed by Hintikka (1962), and perhaps more famously picked up by Stalnaker (1984) and David Lewis (1973). Sets of possible worlds can be used to precisely model *informational content*, and, by extension, a subject's informational perspective.

We can think of information (informational content) in terms of ruling out possible worlds, namely, the possible worlds that are no longer compatible with that information. The more information you have, the more possible worlds you have ruled out. So, for example, omniscience, on this model, involves narrowing down all the possibilities down to one, viz. knowing exactly which world it is that you inhabit.

You or I are far from omniscient, but we have ruled out enough relevant possible worlds to navigate successfully through life. Taking present examples, I have ruled out, among countless other possible worlds, the possible worlds where I am not typing at this computer, where I am not currently wearing a blue shirt, where David Cameron is not the current prime minister of the UK. It can get more fine-grained than words can accurately express (except by means of demonstratives). I have ruled out the possible worlds where this table is not *this* shade of brown, where I don't have *this* precise kind of mild stomach ache, etc.

Possible worlds can be used to characterise the informational content of all sorts of things. They can be used to characterise linguistic phenomena: the conventional meaning of a sentence (i.e. the possible worlds that would be ruled out by an ideally competent language user believing that sentence), the intended meaning of a declarative utterance (the possible worlds the speaker hoped would be ruled out by their interlocutor), the meaning as interpreted by the hearer (the possible worlds taken to be ruled out by the hearer) and so on. They can be used to characterise a given experiential episode (the possible worlds that would be ruled out by judging on the basis of that which is experienced). Most importantly for our purposes, they can be used to characterise entire "belief-sets" (Stalnaker 1984), namely, the totality of what a subject believes, their whole informational perspective, at a given time.

This is, in principle, the set of possible worlds that is compatible with how they take things to stand in the world at a given moment.[14]

It is vital to understand that the possible worlds framework is not a claim about the nature of belief: it is rather a way of modelling belief. However, insofar as holistic dispositionalism draws a sharp distinction between talking about (and modelling) belief and the phenomenon of believing itself, it accommodates the possible worlds framework as one of potentially several ways of modelling belief, and a potentially very accurate one.

Of course, there is a trade-off when we decide to use possible worlds to talk about belief. What the possible worlds framework gains in precision, it loses in manageability. We certainly wouldn't, on a daily basis, use sets of possible worlds to talk about what others believe. However, since it is more precise than natural language attributions, we can use the possible worlds framework to evaluate the accuracy of belief attributions in natural language. In particular, a belief attribution is accurate (under a certain interpretation, if it is ambiguous) to the extent that it doesn't rule out any possibilities that aren't ruled out in the subject's belief-set. So, returning to Fido, attributing the belief that there is *water* is inaccurate because Fido's informational perspective hasn't ruled out that it could be something other than what we humans call *water* (viz. H_20). Fido might simply take there to be – and would be satisfied by – something suitably thirst-quenching.

2.4. Two Consequences of Holistic Dispositionalism

Let me now enumerate two important consequences of holistic dispositionalism. Only the second consequence is strictly incompatible with the discrete representationalism implicit

[14] Along very similar lines, Marcus thought that belief can be nicely modelled in terms of possible (though not necessarily actual) states of affairs. This it "along very similar lines", since this can be expressed as the set of possible worlds in which the states of affairs obtain. On the sense of "state of affairs" intended here, it can clearly include the state of affairs that a certain occurrence has, for example, a degree of likelihood. Some might be tempted to us "state of affairs" to mean something more static. This is how Marcus builds in degrees of belief into her account.

in the Standard Approach.

2.4.1. The relationship between language and belief: "Actions speak louder than words"

As we have seen, sincere assertion is only one indicator of belief, and not an over-riding one. In fact, since language is a symbolic medium, it can cast distance between the subject and that which is asserted. Physical action, on the other hand, cannot do this and therefore *is* an over-riding indication of belief. There are two things to note here. First, we are talking about an "indication" from a God's-eye view. We are not talking about evidence present to the casual everyday observer. Second, and in a related vein, we are talking about *action* not mere *behaviour*. The everyday observer may mistake mere behaviour for action, or may mistake action A for action B, and therefore may misattribute belief on the basis of observable evidence.

Physical actions trump verbal actions because they are more of a commitment, and hence more revealing of how the subject takes things to stand in the world. If I merely *say*, "The bridge over the ravine is safe", I don't personally risk anything if I am wrong (although I may be misinforming somebody else). If I manifest that belief by crossing the bridge, I am risking rather a lot. I can say one thing and physically act contrary to what I have said. I can claim the bridge is safe, but refuse to use it (even though, we may suppose, it is much in my interest to get across it). But I cannot physically act in two opposing ways; I cannot perform an action whose success depends on one and the same state of affairs both obtaining and not obtaining.[15] To put it another way, when people freely act (not just behave) in ways that conflict with their professed beliefs, we take their actions to betray what they "really" believe. As Marcus writes, "actions might belie the agent's words and sincere assenting

[15] I can be hesitant, because I am unsure about whether the bridge will break or not. But here I unambiguously manifest the belief that the bridge *might* break. Whether epistemic modals are built into the content or the attitude is a contentious issue that I won't get into here. For a neat and plausible treatment of epistemic modals within the possible worlds framework see Yalcin (2011). Alternatively, I can be certain of the bridge's safeness, but my agentive manifestation of that can be interfered with by affective or emotional states of the kind that Gendler calls "aliefs" (see section 5 of the previous chapter).

might not be the privileged marker of believing" (1990, p.145).

Unlike the language-centred view, this allows, plausibly, that we can discover what it is that we really believe: "Verbalization as a necessary *condition* of believing precludes [the intuitively commonplace phenomenon of] our discovering and then reporting what we believe" (Marcus 1990, p.139). It also allows that others may at times be better placed to know one's beliefs than oneself.

All of this conflicts with language-centred views, but not necessarily with discrete representationalism. Currie and Egan, for example, clearly don't think that sincere assertion is sufficient for belief. They also allow that one can be mistaken about one's beliefs, and that others can be better placed to know what one believes than oneself. Indeed they think that they themselves are better placed than the delusional subjects that they write about. However, note that being wrong about one's own beliefs on their view amounts to failing to correctly identify one's own mental representations (and on their view, it is the attitude rather than the content that has been misidentified). On the holistic dispositionalist view, on the other hand, it's simply a failure to be aware of one's dispositions to action (in the relevant circumstances), and/or to put them accurately into words.

2.4.2. The totality of what a subject believes

On the dispositional view that we have been sketching, the question "How many beliefs has a subject got?" makes little sense and has no determinate answer. There is perhaps, in principle, for an ideal agent, a fact about *what* is believed, how the subject takes the world to be (including what they have ruled in, ruled out, and are neutral or ignorant about), namely, what I have been calling their "belief-set". However, that is not the aggregate of individual beliefs: it is *holistic*. Rather, as we have seen, a subject can be truly said to have "a belief" if a certain belief attribution accurately captures relevant elements of the subject's belief-set. In stark contrast, on a discrete representationalist view, the question "How many beliefs does a subject have?" does make sense, and does have an answer for an ideal subject. One would simply count up the "discrete representational items in the head" that play a belief-role.

To put it another way, according to holistic dispositionalism, *what is believed* has primacy over *discrete things we call beliefs*. Belief (believing) is a phenomenon that subjects exhibit; it is not a thing that they have. It does make sense to talk about "a belief", but only when we attribute beliefs by using sentences: "*the* belief" in question, inherits its discreteness from the discreteness of the sentence used in the attribution (whether it is an attribution to others or oneself).

2.5. The "Rationality Constraint" as Coherence, not Consistency

Unlike the kind of epistemic irrationality that we examined in Chapters 1 and 4, which tells us when a belief may be bad in some way (because, e.g. it is unsupported) we can talk of an altogether different kind of irrationality, *robust irrationality*, which raises problems for our ability to attribute belief *at all*. This kind of rationality, which enables belief attribution, and which we might call minimal rationality, is the sense of "rationality" used when theorists talk about "the rationality constraint", most famously associated with Davidson (1985).[16]

Any view that takes belief to be, in some way language-centred will derive the rationality constraint from those of logical *consistency*. Beliefs, like well-formed sentences, will be attributed strict logical properties like "consistent, contradictory, logically true, or related by deducibility and the like" (Marcus 1990, p. 138). These are, roughly, properties that exist between symbol-types and properties abstracted from symbol types (although now is not the time to look into foundational issues in philosophy of logic!).

On our view, since belief is neither discrete nor sentential, (and, in a related vein, we have been at pains to emphasise the difference between belief and belief attribution) the relevant kind of rationality should be thought of in terms of *coherent action* of both physical

[16] To clarify: robust irrationality is a failure to exhibit minimal rationality.

and linguistic varieties, instead of strict relations of "consistency" and "deducibility".[17] Conversely, the relevant kind of irrationality is incoherent *behaviour* (since the incoherence precludes the attribution not just of belief, but also of *action*).

Let me present an illustrative example. Suppose another human being, brandishing a poisonous mushroom, announces,

1. "This mushroom is deadly poisonous"

and,

2. "I have no intention of killing myself"

but then,

3. Proceeds to eat the mushroom.

This is certainly a perplexing series of actions (or *apparent* actions) for us to witness. Given this, and only this, information, what could be going on here?

Let us first assume that this person is being minimally rational here (viz. not robustly irrational); these are indeed actions and we can therefore attribute belief, attribute a way the world is as far as the subject is concerned, that is relevant to explaining these actions (namely, rendering them intelligible). This option has to involve re-interpretation of one of the actions (note that I take 1 and 2 to be actions, namely, communicative, verbal ones). This re-interpretation requires the assumption that the actions have a significance (viz. are to be explained by underlying information or motivation) that deviates from their usual or conventional significance. First, we might consider a re-interpretation of his utterances (i.e. the beliefs, desires and intentions behind them). So, for example, 1 or 2 might be uttered with sarcasm or insincerity. Or the utterances are sincere, but he is ill informed about the meaning of one or several of the words uttered (and hence ill informed about the effect that

[17] Bortolotti 2009 talks about "attitude-behaviour inconsistency": "If I sincerely state a preference for chocolate ice-cream over crème caramel, but systematically choose crème caramel over chocolate ice-cream when I order dessert, which preference should an interpreter ascribe to me?" (p.172). I would rather call this "utterance-action inconsistency." And unless there was some further reason for the choice of crème caramel (e.g. suppose you think it's healthier), I would simply say that, since actions speak louder than words, this person is failing to accurately self-ascribe preference. (More, of course, needs to be said about the relationship between belief and preference, but I don't want to digress.)

her utterances will have on a competent English-speaking interlocutor). Second, we might re-interpret his bodily actions (i.e. the beliefs, desires and intentions behind them). Thus we might take utterances 1 and 2 at face value (viz. to be both sincere and not conceptually confused), but think that he is, for example, demonstrating a powerful antidote. Either of these options relieves the incoherence, and renders the subject minimally rational.

However, now let us imagine (or, rather, *try* to imagine) that this being is not doing any of these things. 1 and 2 are uttered sincerely, and are, fully understood (and accurate self-ascriptions), and he is not proving an antidote. Might it not be possible that there is such a degree of cognitive malfunction or disruption that something like this could happen? I would say the following. This *behavioural sequence* is certainly not a metaphysical impossibility. But can we accurately describe this scenario *agentively*, namely, by using terms like "sincerity" or "action"? I'd be tempted to say "No".[18] Granted, only one element needs to be seen as *non-agentive* in order to relieve the incoherence. For example, if we see the eating of the mushroom (or even one of the utterances) as brute behaviour, caused by some freak spasm, then we can take the two utterances at face value. But that is not re-interpretation in the relevant sense. It is not attributing further mental states, but simply removing agency.

When we witness something that looks incoherent, we are forced to re-interpret the agent's actions on pain of denying agency to them in one way or another. *Inconsistency*, on the holistic dispositionalist view, does not force a re-interpretation in this way (although that's not to say that an apparent inconsistency cannot be alleviated by such re-interpretation).

[18] This is not just making the *epistemological* point that in order for us to *access* the way in which an organism takes the world to be we need to pay attention to its behaviour, and in particular make the assumption that this behaviour is *action*. It is rather the *ontological* point that what it *is* for an organism to take the world to be a certain way is at least partly (perhaps fully) constituted by its dispositions to action, to autonomous goal-directed behaviour. Part of what it is to believe that something is lethally poisonous is that you are disposed to not ingest it if you want to stay alive.

Let me illustrate with a mundane and minor cases of inconsistency, that I adapt from Egan (2009).[19] Suppose I'm watching TV and there's a power cut, and the TV goes off. I correctly believe that there has been a power cut. But then I get bored, and think that, instead of watching TV, I can check my email. I try to turn on my computer, and then, of course, instantly realise that it too has been affected by the power cut. Did I *believe* that the computer was not affected by the power cut? If so, have I thereby demonstrated, e.g. that I do not understand what a power cut is? If we put these questions to one side and simply try to describe the case accurately, we can do so in an unproblematic (and un-philosophical) way. I was just being thoughtless, forgetful. In a moment of stupidity, I thought that the computer would work, because, although I hadn't forgotten about the power cut, I had failed to "see" its implications (perhaps because I was very tired). Going back to talk of beliefs, there is a case for saying that I *momentarily* believed that the computer hadn't been affected by the power cut. After all, I did turn on the computer (an action) with the intention of checking my email, while knowing that there was a power cut. However, I can also be said to believe that computers are affected by power cuts, and that my computer is a fairly paradigmatic computer. And yet I still tried to switch it on.

Generally speaking, just because the relevant belief attributions may be difficult to articulate, it doesn't mean that I lack an informational perspective, or even that my turning on the computer was not an action. Note also that the inconsistency here doesn't need re-interpreting in order to portray me as an agent. Inconsistencies, reasoning flaws, and fleeting moments of stupidity, are part of what it is to be human.

The antidoxasticists that we are dealing with here (unlike, e.g. Berrios 1991, who literally takes delusional utterances to lack intentionality) are not saying that the delusional patient needs to be seen as, to some extent lacking in agency (like in the second version of

[19] Egan (2009, p.269) takes this example to show that "the different bits of the stereotypical role – the dispositions to guide certain aspects of behavior in certain circumstances – are separable from one another. A single belief can be disposed to guide one sort of behavior in one sort of context, but not disposed to guide another sort of behavior in a different context". This is clearly all framed with discrete representationalism in the background.

the poisonous mushroom example). Rather, the point is that her delusional assertions need reinterpreting. In particular, they should not be explained in terms of beliefs, but something else.

3. "Incoherence" in Capgras Patients

So, do Capgras patients display the "incoherence" that we have just mentioned? Do they, in a perplexingly direct manner, fail to act in accordance with their claims, in such a way that they need to be re-interpreted? It is, after all, often claimed, and widely accepted, that Capgras patients often fail to show concern for the replaced loved one. Let's look at some examples.

3.1. Some apparent examples of incoherence in Capgras patients

Alexander et al. (1979) report a 44 year-old man, who misidentified his entire family. "At no time after the injury were suspiciousness, paranoia […] noted and he never displayed agitation or anger to anyone, including his wife" (p.334). The patient "described positive feeling toward "both wives", showed no anger or distress about the first wife's desertion" (p.335).

Another example, which we have already encountered, is patient DS (Hirstein and Ramachandran 1997) who exhibits a striking lack of concern for the welfare or whereabouts of his real father. Speaking of his father's impostor, he calmly says: "He's a nice guy, doctor, but he's not my father" (p. 438).

It is not always the case that *no* concern is shown for the replaced loved one. However, even in some cases where there is concern, there is still the problem of how fleeting this concern is. Lucchelli and Spinnler (2007) give us the case of patient "Fred" and his wife "Wilma" (both, of course, fictitious names):

> On another occasion, he urged her [his wife, Wilma] to go with him to report Wilma's disappearance. *Most of the times*, however, he was quite pleased to see her as the "double" Wilma and addressed her in a very gentle way. His wife [viz. Wilma] described his manner as "courting as when we were dating" (p. 189, emphasis added).

So, although Fred occasionally acted as if his wife was missing, he was mostly perfectly happy with the double Wilma. But if the double Wilma isn't the "real" Wilma, then where is the real Wilma? Surely these are not the actions one expects from a man who asserts that his wife, whom he loved very much, is missing.

Stone and Young (1997) sum up what they can glean from these and similar case-studies as follows:

> So, although in some cases of Capgras delusion patients act in ways that seem appropriate to their beliefs, in many other cases one finds a curious asynchrony between the firmly stated delusional belief and actions one might reasonably expect to have followed from it. (p. 334)

Let us note, in passing, that there is mention of "the firmly stated delusional belief" as if the delusional claim, firmly stated, guarantees belief, and the task is then to explain why it fails to have the consequences that it has. But putting this to one side, the main point can be rephrased unproblematically as follows. Just as it is partly constitutive, for example, of your believing that something is deadly poisonous that you should refrain from eating it if you want to stay alive, so it seems to be partly constitutive of your believing that a loved one has been replaced by an impostor that you should be concerned for the welfare or whereabouts of the loved one in question.

From here there are two options for re-interpretation. First, one might say that if you are not concerned then it either can't be a "loved one" that you believe is missing (and here there are two possibilities: either you don't really believe that it is *that person* that's missing, or you can't have *loved* them very much). Second, you can't really *believe* that they are missing (i.e. it is some other attitude that falls short of belief, like imagining). In other words, you can locate the deviance at the level either of content or of attitude. Currie does the latter and that is why he proposes that delusions are actually imaginings mistaken by the subject for beliefs (his "metacognitive view"). In a sense, I'll be doing the former, although the approach I take, and the theoretical background it builds on, is extremely different.

3.2. Accounting for Incoherence within the Standard Approach

The Standard Approach correctly points out that the subject's utterance, however sincere, is not a guaranteed indicator of belief.[20] In particular, according to both Currie (and collaborators and Egan the Capgras patient's sincere utterance is a case in point.

However, it seems to be assumed that if the utterance is taken to not be an indicator of belief, it must still be explained in terms of *some other discrete representational state with the same content*. For Currie this is an imagining, whereas for Egan it is a *sui generis* state, a "bimagining". In other words, if that utterance doesn't give voice to a belief, it must be giving voice to something else. But on our holistic dispositional view there is less pressure to think in this way, since the relationship between an utterance (or belief attribution, including self attribution) and what is believed is much less direct. Granted, insofar as it is action (and not mere behaviour, like a nervous vocal tick) it will be in principle explainable in terms of what the subject believes and desires, but neither of these need, by conceptual necessity, correspond in content to the utterance itself. People are motivated to make certain utterances,

[20] If I believe that I believe that *p* that will suffice to produce a sincere assertion: "*p*". Or, to view it from the other angle, if I assert sincerely and without conceptual confusion, "*p*", that is (given that we are *stipulating* the sincerity and lack of confusion) only a guaranteed indication that I believe that I believe that *p*. It is not a guaranteed indication that I believe that *p*.

including assertions, by all sorts of things. Someone might say to me, "You're looking well" merely as a matter of politeness. Does this mean that, since she doesn't truly have a belief that I'm looking well, she has something else, like an imagining, a hypothesising, a considering, which has the content that I'm doing well? Clearly it is enough to say that she said it because the context motivated her in some way to say it. We generally try to do something (however small or vague) with our actions, and our verbal actions are no different. Of course, I am not saying that the assertions of Capgras patients are anything like this example of harmless insincerity. In fact, as you will see, I think that they are genuine and accurate attempts to communicate how they take the world to be. I am just using this example to illustrate the general relationship between utterances and what is believed, within holistic dispositionalism.

Furthermore, another issue I'd like to raise for both Currie (and collaborators) and Egan's approach to explaining the Capgras cases, which I take to be an upshot of the Standard Approach, is as follows. Aside from being a troublesome and multifarious phenomenon (as Bayne and Pacherie 2005, rightly warn), imagination is also *categorically* different from belief as holistic dispositionalism conceives of it. Unlike believing, imagining is episodic and conscious; you don't have dispositional or unconscious imaginings. Moreover, sometimes we can come to believe on the basis of an imaginative episode. Imagining can be *input* for belief-formation. It's not a *replacement* for it. Consider the following exchange between a hallucinating subject and his friend.

A: Why are you cowering in the corner?

B: Because there's a Lion in the room

A: No there isn't; you're just imagining it.

Here the subject both imagines *and* believes, or rather, believes on the basis of an imaginative episode. A has attributed the imaginative episode to B because something has been conjured up by B's mind ("it's only in his head"). However, belief can also be attributed because the imaginative episode is being taken seriously, and B actually takes there to be a Lion in the room (hence the cowering in the corner). And of course, this imaginative "conjuring up" doesn't always have to be taken seriously, viz. what it presents needn't be believed. You can come to know that you are just imagining something ("Phew, I am only imagining it!"). Furthermore, imagining can come under voluntary control, in which case (thanks to belief's direction of fit, see Chapter 4) it won't be believed.[21]

3.3. Rephrasing the central question of the DD

On the holistic dispositionalist view, belief and imaging are not exclusive states of the same kind that exclude one another (when attributable at the same time, with the same content). As a result of this, it is confused to ask (with an exclusive use of "or") whether a given state is a belief *or* an imagining (or indeed, whether a given content is believed or imagined – it could be both, viz. believed on the basis of having been imagined). Nor should we, like Egan (2009), say that it is somewhere in between (a "bimagining"). Rather, in the DD, we want to know *what* is believed, how the subject takes the world to be, in a way that is relevant to how the delusion is commonly attributed.[22]

The Standard Approach asks: Is this state a belief? And then, if this is answered in the negative, asks: What state is it, if not a belief? Is it an imagining or something in between belief and imagining? On the dispositional view, the relevant question is rather: Does the subject take things to be the way that her assertions suggest? And then if this is answered in the negative, the relevant question is: So how *does* the subject take things to stand in the world? Options for answering *this* question are what we will examine now.

[21] If you ask me to imagine what I would look like with a beard, I can't thereby believe that I have one. (In fact, I must, in order for the request to make sense, believe that I haven't got one.)

[22] Perhaps in advanced stages of psychosis in schizophrenia patients there are various imaginative episodes going on, some of which are being taken seriously, others which are not. The patient suffering from auditory verbal hallucinations may both imagine and, on that basis, believe that he is being spoken to. Alternatively, he may learn to dismiss his voices as being "only in his head. As Lera et al. 2011, write, "Patients who hear the hallucination inside their head rather than outside show better insight, possibly because such patients can understand the voice as being created by their own mind" (p.701). In these cases, the imaginative episode is not accompanied by belief. However, given the nature of the available aetiologies, there is no good reason to think that this is going on in the cases of the Capgras delusion that occur only in the context of brain damage (viz. that don't occur in the context of schizophrenia).

4. What is Believed by the Capgras Patient?

Let's start by stepping back and looking at the DD. What *exactly* is it that the Capgras patient is supposed to believe and that the anti-doxasticist is denying that she believes? I think that there are two importantly different options, of which the second is somewhat ambiguous. I present them as how the delusional subject would express them (viz. as implicit self-ascriptions):

> *Egocentric belief*: "This man is not my father"
>
> *Encyclopaedic belief*: "My father has been replaced by an impostor"

To put my cards on the table, I think that, of these two options, we can correctly attribute the egocentric belief. As for the second option, the encyclopaedic belief attribution is ambiguous. On one reading, it is more problematically attributed to the Capgras patient.

As a first step towards fleshing this out, I now clarify this distinction between encyclopaedic and egocentric believing.

4.1. Encyclopaedic and Egocentric Believing

We humans have the capacity to believe certain things, regardless of where we are and what we are experiencing.[23] I can believe that the Battle of Hastings took place in 1066, while I am locked in a dark cupboard. This "encyclopaedic" believing can be general such as my believing that salt dissolves in water, that lions are dangerous, or they can be singular, about individuals (or events), such as my belief that David Cameron is the current prime

[23] I am, of course, not ruling out that non-human animals have something resembling encyclopaedic belief. However, it seems clear that, since it is more demanding, a larger set of them will be capable of egocentric belief.

minister of the UK (or that the Battle of Hastings took place in 1066). I call this "encyclopaedic" believing because the beliefs that you correctly attribute to me are a bit like entries in an encyclopaedia (which, of course, is not to say that the beliefs themselves are representational items discretely stored in my head like entries in an encyclopaedia). Many of these beliefs we have acquired through communicative or linguistic channels. I can change the way I take the world to be on the basis of what somebody, or some other source of information (e.g. a book or a website) tells me. Linguistically acquired belief tends to fall under the category of encyclopaedic belief. I say, "tends to" because, for example, verbally guiding a blindfolded person – e.g. "To your left there is a chair" – is giving them linguistically mediated yet non-encyclopaedic (what I will call egocentric) information. The information depends on the context in which the subject finds herself.

However, not all of our encyclopaedic beliefs are acquired through communication. We have encyclopaedic beliefs that are *based* on a perceptual event, an event that itself is highly dependent on context.[24] I can believe the generalization that salt dissolves in water, not because my science teacher has told me, but because I have seen it dissolve in water on numerous occasions. I can have the encyclopaedic belief that my friend Anthony is taller than me, without anybody having to tell me this: I can see it for myself. Coming to believe these things *in a context-independent way* will require (among other things) that I have some conception of salt and of my friend Anthony, so that my beliefs come to be about these entities (which will manifest itself in terms of the relevant dispositions towards them). After the initial perceptual encounter with these kinds or individuals of which I have a conception, once I have gleaned the information, I can in principle (e.g. memory permitting) have those beliefs anywhere, namely, while not in perceptual (or iconically/episodically recollective) contact with salt or with Anthony. This is what I mean by encyclopaedic.

[24] Although some beliefs cannot be based on perception due to the nature of the properties involved: my belief that Paris is the capital of France cannot be directly based on perception because you cannot perceive "Capitalhood".

Egocentric belief is not, as one might think, belief that involves reference to believer in the content of the belief attribution (so, it's not reflexive, or at least not necessarily so). Rather, it is believing that *cannot occur* independently of a perceptual or recollective episode. These perceptual and recollective episodes are implicitly egocentric in the sense that they are from a particular perspective (but the subject needn't in any way feature in the content of the belief attribution).

Egocentric belief will typically be expressible using indexicals ("That man is [or was, in the recollective case] wearing a blue shirt"), but not all belief attributions that use indexicals attribute egocentric beliefs. For example, "My father is a nice man" is indexical it will have different truth-conditions when asserted in different contexts (viz. by different people, or at least people who have different fathers), but it is an encyclopaedic belief. This is because, as we shall see, I have a stable conception of my father, and, can have that belief anywhere.

Borrowing Stalnaker's (2007) terminology, I will say that the transition from egocentric belief to encyclopaedic belief requires "detachment", namely, "detachment" from the context in which the initial judgment is made. Detachment applies to both general and singular beliefs. "This substance dissolves in water" detaches to "salt dissolves in water" when you know that this substance is salt. Similarly, "This man is taller than me" detaches to "Anthony is taller than me" when you know "who" this man is, namely, Anthony. Given that we are dealing with singular judgments, and eventually with *delusional* ones, we will focus on the latter case.

4.2. "Detachment" and judging "who someone is"

Let's look at a standard case of detachment for a belief with singular content. As we will see, detachment comes in degrees, depending on the stability and nature of the conception you have of the person the judgment is about (or, to put it another way, the significance they have for you). In this example, there will be three informationally relevant steps.

Step 1: You see someone at a party. You notice that she has green eyes. You believe something you would express as: “This woman has green eyes”.

Step 2: You start talking to her. She tells you that her name is Alexandra, and that she just came back from an expedition where she successfully climbed Everest. You now believe: “This woman [whom I have never met before] is called Alexandra, and has climbed Everest”.

Step 3: Later at the party, your friend Tom tells you that the woman you were talking to is Alexa, an old school friend of yours. Her appearance has changed so much that you failed to identify her.

If we use the metaphor of identity files that I introduced in Chapter 3, what happens at Step 3 is that the file for “This woman here present” (which was formed at Step 1, and stabilized at Step 2) merges with a longstanding file for “My friend Alexa”: the information in those files is taken to pertain to one and the same person. That girl you knew at school went on to climb Everest and is standing before you now. As we saw in Chapter 3, judgments of identification are not of the form Fa, but of the form a=b. You don’t predicate a property of an individual: you simply *draw a connection* between two individuals, taking them to be numerically one and the same.

Correctly identifying someone (“knowing who someone is”) involves drawing a connection between someone currently perceived and someone one has previously encountered (it needn’t be someone encountered in person, it could be someone you are familiar with from the television, e.g. “Is that Brad Pitt?!”). It requires the retrieval or association of the correct file. This describes what is going on at the psychological level.

This can be *modelled* at the abstract, informational, level, in terms of ruling out the possible world where these two individuals are numerically distinct.

The crucial point is that this identification (and the resulting detachment), although we take it for granted, is a clever cognitive achievement. It requires careful management of information, and can in principle go wrong (and often does, even in healthy people) in a number of ways. For example, you can confuse somebody encountered with someone else (retrieve the wrong file). Alternatively, you can simply fail to identify someone (viz. think you've never met them before, fail to retrieve their file, and erroneously create a new file). Note that failing to identify someone and creating a new file is comparatively cognitively *undemanding*. In other words, if the task that the cognitive system faces is simply to open a new file for the person you are currently looking at, this is an easier task than having to retrieve the correct file out of a plethora of other files.

Now, think about the Capgras delusion. The delusional misidentification, at least initially, involves failure to identify an individual (rather than confusing them with somebody else), namely, erroneously opening a new file for an individual who actually already has a file. This amounts to taking there to be two individuals when in truth there is only one (that is why Alexander et al. (1979) are spot on when they call Capgras "a reduplicative phenomenon"). So, keeping in mind what identification amounts to, and the role that it plays in detaching perceptually acquired beliefs that are about individuals, let's look more closely at what the Capgras patient's informational perspective might look like.

4.3. Consequences for the Capgras Delusion

When you meet Alexandra and fail to identify her as your old school friend, you have failed to relate this person here present to the relevant individual from your past (your old school friend Alexa), in particular to relate the two as being one and the same individual. Something analogous (although far more deviant from normality) happens in the Capgras case. The patient fails to identify this man here present with the relevant salient individual from her past (i.e. her father). Admittedly, in the party case you fail to make the connection because Alexa has changed in appearance. The Capgras patient, on the other hand, admits that the person looks just like the person she denies that it is. This inability to provide publicly accessible reasons for belief is, as we saw in Chapter 1, in large part why it is classified as a delusion. Your mistake, on the other hand, is perfectly understandable.

How are we to give the delusional misidentification a characterisation in terms of how the person, all things considered, takes things to stand in the world? One might think it ought to be as easy to do this for the Capgras patient as it is for the case when you simply fail to identify Alexa at the party.[25] However, whereas at the party we can suppose that you act perfectly in accordance with your beliefs (you might reminisce with Alexa about your school days), the patient, as we know, fails to worry about the welfare or whereabouts of the replaced loved one. How can we make sense of this agential inertia?

4.4. Accounting for agential inertia with "competing states"

Agential inertia is one example of incoherence (in particular, it involves *failing* to act as one would expect, rather than acting as one would not expect). Like other forms of incoherence, agential inertia can be accounted for by a re-interpretation in terms of other states that render the subject's behaviour intelligible. To take an example, I might believe that somebody is a dangerous enemy, but treat them with overt kindness. This could be explained by a competing state, for example, an intention to give them a false sense of security and trick them later on.

This is why the likes of Quine (1960) and Davidson (1973) talk about *ceteris paribus* (other things being equal) clauses when ascribing beliefs. Note that using the *ceteris paribus* clause in cases where there are such competing states doesn't tell against the

[25] Another potential difference is that, unlike the Capgras patient, I may have not necessarily "ruled out" that this person is Alexa. I just haven't "ruled it *in*".

presence of the belief in question, but rather presupposes it. It presupposes that the subject still believes something in situations where that may go un-noticed by the casual ascriber.

Now, it is important to see that it is highly implausible that something of this nature can be appealed to in explaining the agential inertia of the Capgras patient's state. It is hard to think of some competing state (e.g. a motivational state) that would render the Capgras belief intelligibly inert.[26] So instead of attempting a "competing states" explanation of the agential inertia, we would do well to re-examine what it is that the delusional patient might plausibly believe, in other words, to be more precise about the patient's informational perspective.

4.5. Egocentric doxasticity and encyclopaedic ambiguity

Let's start by noting one belief that, it seems we cannot make in the light of the case studies mentioned above. It seems that the Capgras patient cannot be said to believe something they might express as:

(2) My father, whom I love, is missing and could be in danger.

In fact, we might say that precisely what perplexes people about Capgras patients is that this isn't believed, in spite of the fact that the following is thought to be:

(2) My father has been replaced by an impostor (or stranger).

So is (2) believed? This encyclopaedic belief attribution is ambiguous. In particular it has a *de dicto* and a *de re* reading. If it is a *de dicto* belief attribution (namely, attributed as the *subject herself* conceives of things) then the subject believes something that they themselves would express as "My father has been replaced by an impostor (or stranger)". Now this implies a number of things. First, it implies that the subject believes that *some*

[26] Granted, one might think that the Capgras patient may have a fear of not being believed. However, this would explain reluctance to a certain kind of verbal action (namely, telling others about it) but it would not prevent actually doing something about the situation, especially if the situation were serious enough (e.g. a missing loved one). We get the opposite in Capgras cases. They make the delusional assertions, but don't do what we would expect in terms of doing something about the situation.

process of replacement (benevolent, malevolent or neutral) has taken place. Second, it implies that the emphasis, in the subject's informational perspective, is at least as much on the *absence of the father* as it is on the *presence of the stranger*. Neither of these seems to be very well supported by the cases. Sometimes patients reason their way into an explanation of the current situation. For example, patient DS (Hirstein and Ramachandran 1997), when asked why this man was pretending to be his father, replied:

> That is what is so surprising, doctor; why should anyone want to pretend to be my father? Maybe my father employed him to take care of me … paid him some money so that he could pay my bills. (p.438)

But note that even here, the "real" father is only appealed to in an explanation of the presence of the impostor, not of the absence of the father. So, I would suggest that the *de dicto* reading of (2) is somewhat misleading given the overall agentive profile of the patients. However, it is important to note that this is not because it would violate a strict rationality constraint. It is rather because it doesn't seem to fit the cases. In other words, if (let us stipulate) the Capgras patient did believe this, but still failed to be concerned, this would have the status of being very perplexing, rather than something that would strictly undermine the initial belief attribution. In other words, it is not strictly incoherent to believe that your father has been replaced while at the same time feeling no concern for his welfare. Concern is not constitutive of that content. It is simply something that we strongly *expect* to follow from that content. For example, we can imagine that a neurologically emotionally stunted person would be able to have that belief without exhibiting concern. This is not to say that, as a result of this stunting, they would not be restricted in the beliefs they could have. They might not be in a position to believe that a *loved one* had been replaced, since they may not be capable of conceiving of anyone as their "loved one".

Alternatively, if (2) is a *de re* attribution (namely, attributed as *the attributor* conceives of the things that feature in the attribution) it seems more accurate. How such an attribution would arise might be as follows. The subject expresses the egocentric judgment, “That man is not my father” (or the erroneously detached encyclopaedic belief “The man I live with is not my father”) and the theorist or clinician attributes, *de re*, the belief “That her father (or “a loved one”) [viz. the man *we* know is her father (“a loved one”)] has been replaced by an impostor/stranger [viz. is not being judged to be that individual]”. I am happy to accept that the *de re* attribution is accurate. But it is much less informative, and falls quite trivially out of the initial egocentric delusional misidentification.[27] And in this case it is not nearly such a serious omission in reasoning that the subject fails to worry about the father’s welfare or whereabouts. Whereas it is a small abductive step (since people tend not to be kidnapped for their own wellbeing) from the *de dicto* attribution:

(2) “My father [whom I love] has been kidnapped and replaced by an impostor”

to something like:

(3) “I hope my father is OK.”

It is a rather larger step from the egocentric judgment

(4) This man is not my father (or the *de re* version of (2))

to (3). This requires a number of steps that, granted, we might *expect* from a normal person, but they aren’t strictly obligatory. It requires the subject to think, for example, “If this man isn’t my father, then where is my father?” and “If my father isn’t here, he might be in danger”.

5. Why the lack of concern?

[27] It’s analogous to somebody observing your behaviour at the party, prior to Stage 3, and saying “She/he thinks that Alexa is somebody else [namely somebody you’ve never met before]”.

Although the lack of concern does not in itself exclude the attribution of a genuinely doxastic delusional misidentification, it is still perplexing and in need of explanation. This is speculative, but two things that may be appealed to are, first, how the subject perceives the "double" (in what "light" she sees him) and, second, how she conceives of the "missing father".

When the patient lays eyes on her father at the initial moment of misidentification, she fails to identify him (in mental file terms, she fails to retrieve the correct file) and judges that this man is not her father. She fails to make the link between "this man" and "my father", and takes herself, so to speak, to be in a possible world in which the man in her presence is a stranger, someone she has never met before, and hence not her father. At this point one can reasonably suppose that she is neutral about *who* this person *is* (indeed, not just neutral, but takes herself to be ignorant of that fact, since she is adamant that she has never met this person before); she is only explicitly committed to who this person *isn't*.[28]

However, how she perceives this man might be relevant for explaining her lack of concern towards the "missing father". There are at least two options for how this man (*de facto* her father) is seen by the delusional subject (viz. her conception of him).

One option is that the file retrieval interferes with the identity judgment cognitively speaking, but leaves something positively emotional towards that individual, something below the level of identity judgment.[29] This remaining emotional element might even be something subpersonally indexed to the "missing" individual. The flirtation of Fred toward the "double Wilma", the patient (in Alexander et al 1979) who describes positive feelings

[28] One may even want to put this, as Bayne and Pacherie (2004) in their "endorsement theory" do, in terms of the content of her perception. In other words, what the patient's visual experience is "telling her" is something like "This person is not my father". Alternatively, we can think of it in terms of the aetiology I put forward in Chapter 3, in terms of a perceptual judgment that is unavoidably caused by a specialised tracking mechanism going wrong.

[29] Although it seems in conflict with the view that Capgras delusion involves a lowered SCR (Ellis et al. 1997) to the misidentified loved ones. I wouldn't attach too much importance to this, however, since, firstly, SCR is a surface indicator, and secondly, Capgras patients clearly do seem capable of feeling normal levels of emotional warmth towards others.

towards “both wives”, and DS, who calls his father’s double “a nice guy” all seem to support something like this.

Alternatively, this could be explained not because an emotional bond remains, so to speak, but because an emotional bond may be quickly built up (especially given the similarity not only in appearance, but also in mannerisms, to the loved one). In mental file terms, another mental file, the new one for “The guy I now live with”, becomes stable and is filled with more and more information (in a way analogous to a scenario where, suppose, I fail to identify Alexa, but she acquires a new salience, and I think about her, that individual, under a robust but new mode of presentation, e.g. “My *new* friend Alexa”).[30] According to either of these explanations (viz. residual or created emotional bond with “the double”), we might say that the father, although “*cognitively* missing” is not “emotionally missed”. We might even speculate that this is because the subject is getting the “emotional dose” she would get from his presence.

How does this account for cases, not of fondness, but of *violence*, towards the double? One thing to say in response to this is that these violent cases are less problematic for explaining agentive inertia, because violence is what one might expect to be exhibited towards somebody pretending to be a loved one. Furthermore, we might hypothesise that for these patients the residual emotional component in the “stranger-friendly” patients was damaged along with the identificational component. Another possibility is that the cases of violent delusional misidentification are substantially different in their underlying nature. Nestor et al. 1995, claim that “the so-called Capgras […] syndromes evident in some of the severely violent patients could not be related to a specific organic insult, but rather occurred against a backdrop of other significant paranoid delusions” (p.338).

What about the delusional subject’s conception of the missing loved one; the “real father”? As I said, we can model her informational perspective in terms of her taking the

[30] Although, as we saw in Chapter 3, this sometimes doesn’t happen. The Capgras patient sometimes duplicates over and over (i.e. the doubles themselves have doubles).

world to be such that there are two men who look the same, one is her father, the man who (among other things) played a major role in her upbringing, and the other is the man she now lives with, who looks like her father, and who she may have grown quite fond of. In spite of this, she seems to not only show lack of concern, but also be informationally neutral about the whereabouts of her real father. We might even say that, as a result of her brain damage, she has informationally "lost track" of him. Note, however, that this doesn't prevent us from attributing the egocentric belief that this man is not her father. Nor does it even prevent us from attributing the encyclopaedic belief that the world has two individuals who look the same.

It is a virtue of this position that it is in keeping with the following peculiarity in the patient's agential profile. Capgras patients, although they (often) fail to worry about their "missing" loved ones, do behave differently towards their (actual, present) loved ones and sometimes, as we saw (in approximately 18 % of cases) inflict violence on them. That is often, after all, why they come to the attention of clinicians. In other words, this combination of egocentric belief, and lack of the standardly attributed encyclopaedic belief, accords nicely with those cases where there is a change of general behaviour towards the loved one ("the double") (egocentric), coupled with a strange lack of concern for the welfare or whereabouts of the "real" loved one (encyclopaedic). The immediate, constitutive, consequences of perceptually based delusional misidentification play out, but more distant ones, which though expected, are not constitutive of the delusional misidentification, do not.

Conclusion

The Standard Approach asks: "Are delusions beliefs?" By asking this there have been a number of implicit assumptions, the clearest of these being that there is a discrete, belief-like state that has the same content as the standard clinical characterization of the delusion: "that a loved one has been replaced by an identical-looking impostor". The challenge is then to ascertain whether the state in question has the right functional role. A consequence of this is that the *content* of the delusional state is not questioned (it is treated a bit like a sentence in the head, with a fixed content); rather it is the attitude that the subject bears towards the content that is up for dispute. This leads some, like Currie and his collaborators (e.g. Currie and Jureidini 2001), to claim that delusions are imaginings that are mistaken for beliefs by the subject (the patient imagines that *p* and so asserts that *p*, but fails to genuinely believe that *p*). Others, like Egan (2009) – who himself openly admits (in the long paragraph I quoted) that he is *assuming* discrete representationalism – blur the traditional boundaries between attitudes and claim that delusions are in-between states (bimagining; somewhere between belief and imagining).

If we give up discrete representationalism in favour of holistic dispositionalism, the issue is no longer whether "the delusional state" plays the right role, but rather whether the delusional patient really takes things to be the way we think she does. If instead, we ask: "How does the patient take things to stand in the world?" we get to the heart of the problem, without subscribing to any presuppositions about what happens to underpin human believing. The question of whether a fixed representational state, "the delusion", is a belief or not is secondary, and comes with a great deal of theoretical baggage. The really interesting problem with delusional patients is that we find it hard to characterise how they take the world to be. However, just because we find elements of their informational perspective hard to characterise, it doesn't mean that they haven't got one.

It is not only because of the pathological nature of these subjects, that we find it hard to attribute beliefs to them. It is also partly because we have been using rather blunt tools in attempting such attributions, in attempting to capture their informational perspective. In particular, we have been using rather vague and ambiguous sentential attributions (viz. "The subject believes that a loved one has been replaced by an identical-looking stranger"). Even the rather simple refinement of distinguishing egocentric from encyclopaedic believing, and reflecting on the relationship between the two, sheds valuable light on these cases.

Furthermore, reflecting on whether our belief attributions to the delusional patient are *de re* or *de dicto*, stimulates us to attempt more accurate and useful attributions *de dicto* (namely, attributed as the subject herself conceives of things).

Bibliography

Abed, R.T. and Fewtrell, W.D. (1990) “Delusional misidentification of familiar inanimate objects. A rare variant of Capgras syndrome”. *British Journal of Psychiatry*.157:915- 7.

Achinstein, P. (1983) *The Nature of Explanation*. Oxford University Press

Ahles, Scott, R. (2004). *Our Inner World: A Guide to Psychodynamics and Psychotherapy*. Johns Hopkins University Press.

Alexander, M. P., Stuss, D. T., & Benson, D. F. (1979). “Capgras syndrome A reduplicative phenomenon”. Neurology, 29(3), 334-334.

Alloy, L.B., Abramson, L.Y. (1979). “Judgment of contingency in depressed and nondepressed students: Sadder but wiser?”. *Journal of Experimental Psychology: General* 108: 441–485

American Psychiatric Association (2000). *Diagnostic Statistical Manual of Mental Disorders*, Fourth edition, Text Revision (DSM-IV-TR).

Anscombe, G. E. M. (1957). Intention, Oxford: Basil Blackwell.

Audi, R. (1972). “The concept of 'believing'”. *Personalist* 53:43-52.

Austin, J. L. (1962) *How to do Things with Words*. The William James Lectures delivered at Harvard University in 1955. (ed. J. O. Urmson and Marina Sbisà), Oxford, Clarendon Press

Bach, K. (1994). *Thought and Reference.* Oxford University Press.

Baker, L. R. (1995), *Explaining Attitudes*. Cambridge University Press: Cambridge.

Bauer, R.M. (1984). Autonomic recognition of names and faces in prosopagnosia: a neuropsychological application of the guilty knowledge test, *Neuropsychologia* 22/4: 457-69.

Bayne, T. and Fernández, J. (2009) (eds.) *Delusion and Self-deception: Affective and Motivational Influences on Belief Formation*, Hove: Psychology Press

Bayne, T. and Pacherie E. (2004a). “Bottom up or top down?,” *Philosophy, Psychiatry, & Psychology*, 11 (1): 1–11.

Bayne, T. and Pacherie, E. (2004b). “Experience, belief, and the interpretive fold,” *Philosophy, Psychiatry, & Psychology*, 11 (1): 81–86.

Bayne, T. and Pacherie, E. (2005). “In defence of the doxastic conception of delusion,” *Mind & Language*, 20 (2): 163–188.

Bechtel, W. P. and Wright, C. D. (2007). “Mechanisms and psychological explanation”. In Paul Thagard (ed.), *Philosophy of Psychology and Cognitive Science*. Elsevier.

Becker, G. and K. Murphy (1988) "A Theory of Rational Addiction". *Journal of Political Economy*, 96, 675-700.

Bermudez, J. L. (2005) *Philosophy of Psychology: A Contemporary Introduction*, Routledge Contemporary Introductions to Philosophy

Berrios, G. E. (1991). "Delusions as 'wrong beliefs': a conceptual history," *British Journal of Psychiatry*, 159 (suppl. 14): 6–13.

Blanchette, P. A. (2012). "Frege on shared belief and total functions". *Journal of Philosophy* 109 (1).

Bleuler, E. (1950). *Dementia Precox, or the Group of Schizophrenias*, New York: International University Press.

Bhatia, M.S (1990). "Capgras syndrome in a patient with migraine". *British Journal of Psychiatry*.157:917–918.

Bortolotti, L. and Mameli, M. (2012). "Self-deception, delusion and the boundaries of folk psychology". *HumanaMente* 20:203-221.

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

Bortolotti, L. and Broome, M.R. (2007). "If you didn't care, you wouldn't notice: recognition and estrangement in psychopathology," *Philosophy, Psychiatry, & Psychology*, 14(1): 39–42.

Braithwaite, R.B. (1932–1933), "The nature of believing", *Proceedings of the Aristotelian Society*, 33, 129–146.

Breen, N., Caine, D., & Coltheart, M. (2000). Models of face recognition and delusional misidentification: a critical review. *Cognitive Neuropsychology, 17(*1–3), 55–71.

Breen, N., Caine, D., Coltheart, M., Hendy, J. and Roberts, C. (2000). "Towards an understanding of delusions of misidentification: four case studies," in M. Coltheart and M. Davies (eds.) *Pathologies of Belief*, Oxford: Blackwell, 74–110.

Broome, M.R. (2004). "Rationality in psychosis and understanding the deluded," *Philosophy, Psychiatry, & Psychology*, 11(1): 35- 41.

Bullot, N. (2009). "Toward a Theory of the Empirical Tracking of Individuals: Cognitive Flexibility and the Functions of Attention in Integrated Tracking". *Philosophical Psychology* 22 (3):353-387.

Butler, R. W., & Braff, D. L. (1991). "Delusions: A review and integration." *Schizophrenia Bulletin*, 17, 633-647.

Campbell, J. (1999). "Schizophrenia, the space of reasons and thinking as a motor process," *The Monist*, 82(4): 609–625.

Campbell, J. (2001) "Rationality, meaning and the analysis of delusion," *Philosophy, Psychiatry, & Psychology*, 8 (2–3): 89–100.

Campbell, J. (2009). “What does rationality have to do with psychological causation? Propositional attitudes as mechanisms and as control variables,” in M. Broome and L. Bortolotti (eds.) *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*, Oxford: Oxford University Press, 137–150.

Capgras, J. and Carette, P. (1924). “Illusion de sosies et complexe d'Oedipe”. *Annales Medico-*
Psychologiques, 82, 48-68.

Capgras J., Reboul-Lachaux J. (1923). “Illusion des « sosies » dans un délire systématisé chronique”. *Bulletin de la Société Clinique de Médicine Mentale* 2: 6–16.

Carey, S. and Xu, F. (2001). “Infants’ Knowledge of Objects: Beyond Object Files and Object Tracking”, *Cognition* 80 179-213
Castillo P.M. and Berman C.W. (1994). “Delusional gross replacement of inanimate objects” *British Journal of Psychiatry*.164(5):693-696,

Chisholm, R. M. (1957), *Perceiving* (Ithaca: Cornell)

Clark, A. (2013). “Whatever next? Predictive brains, situated agents, and the future of cognitive science” *Behavioural and Brain Sciences*. 36 (3)

Coltheart, M. (2005). “Conscious experience and delusional belief,” *Philosophy, Psychiatry & Psychology*, 12 (2): 153–157.

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

Coltheart, M., Langdon, R. and McKay, R. (2007). “Schizophrenia and monothematic delusions,” *Schizophrenia Bulletin*, 33 (3): 642–647.

Corlett, P.R.. Taylor J.R., Wang X.J., Fletcher P.C., Krystal J.H.. (2010), “Toward a Neurobiology of Delusion” *Progress in Neurobiology* 92(3):345-69

Currie, G. (2000). “Imagination, delusion and hallucinations,” in M. Coltheart and M. Davies (eds.) *Pathologies of Belief*, Oxford: Blackwell, 167–182.

Currie, G. and Jureidini, J. (2001). “Delusions, rationality, empathy,” *Philosophy, Psychiatry and Psychology*, 8 (2–3): 159–162.

Currie, G. and Ravenscroft, I. (2002). *Recreative Minds: Imagination in Philosophy and Psychology*, Oxford: Oxford University Press.

Dalgalarrondo, P., Fujisawa, G., & Banzato, C.E.M. (2002). Capgras syndrome and blindness: Against the prosopagnosia hypothesis. *Canadian Journal of Psychiatry*, 47(4), 387-388.

David, A. (1999). “On the impossibility of defining delusions”. *Philosophy, Psychiatry and Psychology*. 6(1): 17–20.

Davidson, D. (1963), ‘Actions, Reasons and Causes’, *Journal of Philosophy*, 60: 685–700

Davidson, D. (1970), 'Mental Events', in Lawrence Foster and J. W. Swanson (eds.), *Experience and Theory*, London: Duckworth

Davidson, D. (1973), "The Material Mind", Reprinted in Davidson (1980), *Essays on Actions and Events*, Oxford: Clarendon Press

Davidson, D. (1982). "Paradoxes of irrationality," in R. Wollheim (ed.) *Philosophical essays on Freud*, Cambridge University Press, 289–305. Reprinted in D. Davidson (2004) *Problems of Irrationality*, Oxford: Clarendon Press, 169–188.

Davies, M. (2000) "Interaction without reduction: The relationship between personal and subpersonal levels of description". *Mind and Society* 1 (2):87-105.

Davies, M. (2008). "Delusion and motivationally biased belief: self deception in the two factor framework," in T. Bayne and J. Fernàndez (eds.), *Delusion and Self deception: Affective and Motivational Influences on Belief Formation*, Hove: Psychology Press, 71–86.

Davies, M. and Coltheart, M. (2000). "Introduction," in M. Coltheart and M. Davies (eds.), *Pathologies of Belief*, Oxford: Blackwell, 1–46.

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

Dean, C. and Surtees P.G. (1989) "Do psychological factors predict survival in breast cancer?" *Journal of Psychosomatic Research* 1989;33(5):561-9.

Dennett, D. C. (1969), *Content and Consciousness* (London: Routledge).

Dennett, D. C. (1971). "Intentional Systems". Journal of Philosophy 68 (February):87-106.

Dennett, D. C. (1978), Brainstorms (Cambridge, MA: MIT).

Dennett, D. C. (1981). "True Believers : The Intentional Strategy and Why It Works:. In A. F. Heath (ed.), Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford. Clarendon.

de Pauw, K.W., Szulecka, T.K. (1988) "Dangerous delusions. Violence and the misidentification syndromes." *British Journal of Psychiatry* 152: 91–6

Egan, A. (2009). "Imagination, Delusion and Self-Deception" in T. Bayne and J. Fernàndez (eds.), *Delusion and Self deception: Affective and Motivational Influences on Belief Formation*, Hove: Psychology Press

Her. U. (1999) "A zoocentric Capgras syndrome". *Psychiatr Prax.* 1999; **26:** 43 – 44.

Ellis, H. (1998). "Cognitive neuropsychiatry and delusional misidentification syndromes: an exemplary vindication of a new discipline," *Cognitive Neuropsychiatry*, 3 (2): 81–89.

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., and de Pauw, K. W. (1997). Reduced autonomic responses to faces in Capgras delusion. *Proceedings of the Royal Society of London: Biological Sciences, B264,* 1085-1092

Evans, G., (1982). *The Varieties of Reference*, Oxford: Oxford University Press.

Faye, J. (2007). "The pragmatic-rhetorical theory of explanation." In: Persson J, Ylikoski P (eds) *Rethinking explanation*. Series: Boston studies in the philosophy of science, vol. 252.

Feinberg, T., Eaton L. Roane D., Giacino J. (1999). "Multiple Fregoli delusions after traumatic brain injury". *Cortex* 35 (3): 373–87

Fine, C., Craigie, J. and Gold, I. (2005). "Damned if you do, damned if you don't: The impasse in cognitive accounts of the Capgras delusion," *Philosophy, Psychiatry & Psychology*, 12: 143–151.

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.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Freeman, D. (2008). "The assessment of persecutory ideation," in D. Freeman, R. Bentall, and P. Garety (eds.) *Persecutory Delusions. Assessment, Theory and Treatment*, Oxford: Oxford University Press, 23–52.

Friedman, M. (1974). "Explanation and Scientific Understanding". *The Journal of Philosophy*, Vol. 71, No. 1.

Frith, C. (1992). *The Cognitive Neuropsychology of Schizophrenia*, Hove: Psychology Press.

Gallagher, S. (2009). "Delusional realities," in M. R. Broome and L. Bortolotti (eds.) *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*, Oxford: Oxford University Press, 245–268.

Garety, P. (1991). "Reasoning and delusions," *British Journal of Psychiatry*, 159: 14–18.

Garety, P. A. and Freeman, D. (1999). "Cognitive approaches to delusions: A critical review of theories and evidence," *British Journal of Clinical Psychology*, 38: 113–154.

Garety, P. and Hemsley, D. (1987). "Characteristics of delusional experience," *European Archives of Psychiatry and Neurological Sciences*, 236: 294–298.

Garety, P. and Hemsley, D. (1997). "Delusions: Investigations into the Psychology of Delusional Reasoning," Hove: Psychology Press.

Garety, P.A, Kuipers, E., Fowler, D.G, Freeman, D. and Bebbington, P.E. (2001). "A cognitive model of the positive symptoms of psychosis," *Psychological Medicine*, 31: 189–195.

Gazzaniga, M. (1985). *The Social Brain*, New York: Basic Books.

Gendler, T. (2011). “Imagination”, *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2011/entries/imagination/>.

Gendler, T. (2008a) “Alief and Belief,” *Journal of Philosophy*, 105(10).

Gendler, T. (2008b) “Alief in Action (and Reaction),” *Mind and Language*, 23(5): 552–85.

Gerrans, P. (2012) “Dream experience and a revisionist account of delusions of misidentification”, *Consciousness and Cognition* 21, 217-227

Graham, G. (2013) *The Disordered Mind: An Introduction to Philosophy of Mind and Mental Illness (Second Edition)*. Routledge.

Hacker, P. M. S. (2004). “On the ontology of belief”. In Mark Siebel & Mark Textor (eds.), *Semantik Und Ontologie*. Frankfurt: Ontos Verlag.

Halligan, P. and Marshall, J. (1996). “The wise prophet makes sure of the event first: Hallucinations, amnesia, and delusions,” in P. Halligan and J. Marshall (eds.) *Method in Madness*, Hove: Psychology Press, 235–266.

Halligan, P. and Marshall, J. (eds.) (1996). *Method in Madness: Case Studies in Cognitive Neuropsychiatry*, Hove: Psychology Press.

Hamilton, A. (2007). “Against the belief model of delusion,” in M.C. Chung, K.W.M. Fulford, and G. Graham (eds.) *Reconceiving Schizophrenia*, Oxford: Oxford University Press, 217–234.

Heal, J. (1998). Consciousness and content. In Anthony O'Hear (ed.), *Contemporary Issues in the Philosophy of Mind*. Cambridge University Press.

Heckers, S. 2009. Who is at Risk for a Psychotic Disorder? Schizophrenia Bulletin 35(5) 847-850.

Hempel, C. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press.

Hintikka, J. (1962). *Knowledge and Belief*. Ithaca, N.Y.,Cornell University Press.

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 B: Biological Sciences, 264,* 437-444.

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

Hohwy, J. (2007). “The sense of self in the phenomenology of agency and perception,” *Psyche*, 13(1).

Hohwy, J. and Rosenberg, R. (2005). “Unusual experiences, reality testing and delusions of alien control,” *Mind & Language*, 20(2): 141–162.

Hosty, G. (1992) “Beneficial delusions” (correspondence). *Psychiatric Bulletin*, 16, 373

Humphreys, N. and Dennett, D. (1989). “Speaking for our selves: An assessment of multiple personality disorder,” *Raritan*, 9(1): 68–98.

Huq, S., Garety, P., Hemsley, D. (1988). “Probabilistic judgements in deluded and non-deluded subjects,” *Quarterly Journal of Experimental Psychology*, 40(A): 801–812.

Jaspers, K (1963). *General Psychopathology*, J. Hoenig and M. Hamilton (trans.), Manchester: Manchester University Press.

John, C.H., Dodgson, G. (1994). Inductive reasoning in delusional thinking. *Journal of Mental Health*, 3:31-49.

Johns, L.C., van Os, J. (2001). “The continuity of psychotic experiences in the general population,” *Clinical Psychology Review*, 21 (8): 1125–1141.
Jones, E. (1999). “The phenomenology of abnormal belief : A philosophical and psychiatric inquiry”. *Philosophy , Psychiatry and Psychology*, 6, 1-16.

Kapur, S. (2004). “How antipsychotics become anti-‘psychotic’ – from dopamine to salience to psychosis,” *Trends in Pharmacological Sciences*, 25: 402–406.

Kauppinen, A. (2010). “The pragmatics of transparent belief reports”. *Analysis* 70 (3):438-446.

Landis T, Cummings JL, Christen L, Bogen JE, Imhof HG (1986). “Are unilateral right posterior cerebral lesions sufficient to cause prosopagnosia? Clinical and radiological findings in six additional patients”. *Cortex*;22:243 – 52.

Langdon, R. and Coltheart, M. (2000). “The cognitive neuropsychology of delusions,” in M. Coltheart and M. Davies (eds.) *Pathologies of Belief*, Oxford: Blackwell, 183–216.

Langdon, R., Ward, P. and Coltheart, M. (2008). “Reasoning anomalies associated with delusions in schizophrenia,” *Schizophrenia Bulletin*, [Available online].

Leader, D. (2011). *What is Madness?* Penguin Books.

Lera G, Herrero N, et al. (2011). “Insight among psychotic patients with auditory hallucinations”. Journal of Clinical Psychology 67 (7): 701-8.

Lewis, D. (1973) *Counterfactuals*, Oxford: Blackwell

Lucchelli, F. and Spinnler, H. (2007). “The case of lost Wilma: a clinical report of Capgras delusion,” *Neurological Science*, 28(4): 188–195.

McDowell, J. (1985). “Functionalism and anomalous monism”. In Brian P. McLaughlin & Ernest LePore (eds.), *Action and Events*. Blackwell.

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

Maher, B.A. (1988). “Anomalous experience and delusional thinking: The logic of explanations,” in T.F. Oltmann and B.A. Maher (eds.) *Delusional Beliefs*, New York: Wiley, 15–33.

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

Mandelbaum (forthcoming). “Against Alief”. *Philosophical Studies*

Marcus R. B. (1983). “Rationality and Believing the Impossible”. *Journal of Philosophy* 80 (6):321-338.

Marcus R. B. (1990) “Some Revisionary Proposals About Belief and Believing”. *Philosophy and Phenomenological Research* 50:133-153.

Marcus R. B. (1995) “The Anti-naturalism of Some Language Centered Accounts of Belief”. *Dialectica* 49: 113-130.

Marr, D. (1982). *Vision*. Freeman.

Marshall, J. and Halligan, P. (1996). “Introduction,” in P. Halligan and J. Marshall (eds.) *Method in Madness*, Hove: Psychology Press, 3–11.

McKay, R. and Cipolotti, L. (2007). “Attributional styles in a case of Cotard delusion,” *Consciousness and Cognition*, 16: 349–359.

McKay, R., Langdon, R. and Colheart, M. (2007). “Models of misbelief: Integrating motivational and deficit theories of delusions,” *Consciousness and Cognition*, 16: 932–941.

McKay, R., Langdon, R. and Coltheart, M. (2005a). ”Sleights of mind: Delusions, defences, and self deception,” *Cognitive Neuropsychology*, 10: 305–326.

McKay, R., Langdon, R. and Coltheart, M. (2005b). “Paranoia, persecutory delusions and attributional biases,” *Psychiatry Research*, 136 (2–3): 233–245.

Millikan, R. (1989). “Biosemantics”. *Journal of Philosophy* 86 (July):281-97.

Moor, J., and Tucker, G. (1979) “Delusions: Analysis and criteria”. *Comprehensive Psychiatry,* 20:388-393.

Murphy, D. (2012) “The Folk Epistemology of Delusions”. *Neuroethics* 5 (1):19-22.

Murphy, D. (2013). “Delusions, Modernist Epistemology and Irrational Belief”. *Mind and Language* 28 (1):113-124.

Nasar, S. (1994). *A Beautiful Mind: A Biography of John Forbes Nash, Jr., Winner of the Nobel Prize in Economics*. Simon & Schuster.

P.G Nestor, J Haycock, S Doiron, J Kelly, D Kelly (1995). “Lethal violence and psychosis: a clinical profile” *Bulletin of the American Academy of Psychiatry and the Law*, 23, pp. 331–341

Nozick (1993) *The Nature of Rationality.* Princeton University Press.

Pacherie (2009) “Perception, Emotions and Delusions: Revisiting the Capgras Delusion” in T. Bayne and J. Fernàndez (eds.), *Delusion and Self deception: Affective and Motivational Influences on Belief Formation*, Hove: Psychology Press

Palermo, R. and Rhodes, G. (2007). “Are you always on my mind? A review of how face perception and attention interact”. *Neuropsychologia*. 2007 Jan 7;45(1):75-92.

Papineau, D. (2012). “There Are No Norms of Belief”. In T. Chan (ed.), *The Aim of Belief.*

Pollock, J. (1986). *Contemporary Theories of Knowledge*. (Towota, NJ: Rowman And Littlefield Publishers). 1st edition.

Perring, C. (2010) “Mental Illness”, *The Stanford Encyclopedia of Philosophy* (Spring 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2010/entries/mental-illness/>.

Perry, J. (2001). *Reference and Reflexivity*. Stanford, CA: CSLI Publications.

Pitt, D. (2004). “The phenomenology of cognition, or, what is it like to think that *P*?” *Philosophy and Phenomenological Research* 69 (1):1-36.

Price, H. H. (1969). *Belief: The Gifford Lectures Delivered at the University of Aberdeen in 1960*. New York, Humanities Press

Proust, J. (2012). “The norms of acceptance”. *Philosophical Issues* 22 (1):316-333.

Pryor, J. (2000). “The skeptic and the dogmatist”. *Noûs* 34 (4):517–549.

Quine, W.V.O., (1960), *Word and Object*, Cambridge, Massachusetts: MIT Press.

Radden (2012), *On Delusion,* Thinking in Action Series, Routledge

Ramachandran, V.S. and Blakeslee, S. (1998). *Phantoms in the Brain: Human Nature and the Architecture of the Mind*, London: Fourth Estate.

Ramsey, F. P. (1931). *Foundations: Essays in Philosophy, Logic, Mathematics, and Economics*. Humanties Press.

Ratcliffe, M. (2004). “Interpreting delusions,” *Phenomenology and Cognitive Sciences*, 3: 25–48.

Ratcliffe, M. (2008). “The phenomenological role of affect in the Capgras delusion”. *Continental Philosophy Review* 41 (2):195-216.

Recanati, F. (1993). *Direct Reference*. Cambridge University Press.

Reimer, M. (2009) “Is the Impostor Hypothesis Really so Preposterous? Understanding the Capgras Experience.” *Philosophical Psychology* 22 (6): 669 – 686.

Reimer, M. (2011). “Only a Philosopher or a Madman: Impractical Delusions in Philosophy and Psychiatry”. *Philosophy, Psychiatry, and Psychology* 17 (4).

Reznek, L. (1987). *The Nature of Disease*. Routledge & Kegan Paul.

Reznek, L. (2010). *Delusions and The Madness of the Masses.* Rowman & Littlefield

Roberts, G. (1991). “Delusional Belief Systems and Meaning in Life: A Preferred Reality?” *British Journal of Psychiatry,* 159 (suppl. 14), 19-28.

Russell, B. (1921). *The Analysis of Mind.* George Allen and Unwin Ltd.

Searle, J. (1983). *Intentionality*. Oxford University Press.

Searle, J (1979). “What is an intentional state?” *Mind* 88 (January):74-92.

Samuels, R. (2009). “Delusions as a natural kind,” in M.R. Broome and L. Bortolotti (eds.) *Psychiatry as Cognitive Neuroscience: Philosophical Perspectives*, Oxford: Oxford University Press, 49–82.

Sass, L. (1994). *The Paradoxes of Delusion: Wittgenstein, Schreber, and the Schizophrenic, Mind*, Ithaca: Cornell University Press.

Scholl, B. (2007). Object Persistence in Philosophy and Psychology. *Mind and Language* 22 (5):563–591.

Schwartz, S., Maquet, P. (2002). Sleep imaging and the neuro-psychological assessment of dreams. *Trends in Cognitive Science*, 6(1), 23-30.

Schwitzgebel, E. (2002) “A phenomenal, dispositional account of belief”, *Nous*, 36, 249–275.

Schwitzgebel, E. (2006) “Belief” *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2011/entries/belief/>.

Siddle R, Haddock G, Tarrier N, Faragher EB. (2002). “Religious delusions in patients admitted to hospital with schizophrenia”. *Social Psychiatry and Psychiatric Epidemiology.* 37:130-138.

Silva, J. A, Leong, G. B. (1997) *Canadian Journal of Psychiatry*; 42 (6) 665

Smith, M. (1987). “The Humean theory of motivation”. *Mind* 96 (381):36-61.

Stalnaker, R. (1984). *Inquiry*. Cambridge University Press.

Stalnaker, R. (1993) “What is the representation theory of thinking?: A comment on William G. Lycan”. *Mind and Language* 8 (3):423-430.

Stalnaker, R. (2007). *Our Knowledge of the Internal World.* Oxford University Press.

Stone, T. & Davies, M. (1996). “The mental simulation debate: A progress report”. In Peter Carruthers & Peter K. Smith (eds.), *Theories of Theories of Mind.* Cambridge University Press.

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

Stone, A., Valentine, T. (2004). Better the Devil You Know? Non-conscious processing of identify and affect of famous persons. *Psychonomic Bulletin & Review*, 11, 469-474.

Strawson, P. (1959) *Individuals: An Essay in Descriptive Metaphysics*. Routledge.

Taylor, S.E.; Brown, J. (1988). "Illusion and well-being: A social psychological perspective on mental health". *Psychological Bulletin* 103 (2): 193–210.

Taylor, S. E. (1989). *Positive illusions: Creative self-deception and the healthy mind.* New York: Basic Books

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

Velleman, J. D. (1992). "What Happens When Someone Acts?" *Mind* 101 (403):461 - 481.

Velleman, J. D. (2000). "On the aim of belief". In David Velleman (ed.), *The Possibility of Practical Reason*. Oxford University Press.

Vainik, E. (2002). "Emotions, Emotion Terms and Emotion Concepts in an Estonian Folk Model". *Trames*, 6 (4), pp. 322–341.

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

Wedgwood, R. (2002), "The Aim of Belief", *Philosophical Perspectives* 16, 267–297.

Williams, B. (1973) "Deciding to believe" in Bernard Williams (ed.), *Problems of the Self.* Cambridge University Press

Wittgenstein, L. (1969) *On Certainty*, G.E.M. Anscombe and G.H. von Wright (eds.), G.E.M. Anscombe and D. Paul (trans.), Oxford: Blackwell.

Wilkinson (in press) "Dennett's Personal/Subpersonal Distinction in the Light of Cognitive Neuropsychiatry" in: *Consciousness and Content Revisited*, (de Brigard and Munoz-Suarez eds.) *Studies in Brain and Mind* Series (Springer))

Wilkinson (forthcoming) "The Status of Delusion in the Light of Marcus's Revisionary Proposals", *Theoria*

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Yalcin, S. (2011). "Nonfactualism About Epistemic Modality". In A. Egan & B. Weatherson (eds.), *Epistemic Modality*. Oxford University Press.

Young, A., Reid, I., Wright, S., Hellawell, D.J. (1993). "Face-processing impairments and the Capgras delusion," *British Journal of Psychiatry*, 162: 695–698.

Young, A.W. and Leafhead, K. (1996). "Betwixt life and death: Case studies of the Cotard delusion," in P. Halligan and J. Marshall (eds.) *Method in Madness: Case Studies in Cognitive Neuropsychiatry*, Hove: Psychology Press, chapter 8.