STRANGE BELIEFS: ESSAYS ON DELUSION FORMATION

by

FEDERICO BONGIORNO

A thesis submitted to the University of Birmingham for the degree of
DOCTOR OF PHILOSOPHY

Department of Philosophy
School of Philosophy, Theology, and Religion
College of Arts and Law
University of Birmingham
September 2020

Strange Beliefs: Essays on Delusion Formation

Abstract:

This thesis is set out as a collection of self-standing essays. Throughout these essays, I try to illuminate a number of controversies surrounding the way in which delusions are formed, and relatedly, their nature and epistemic standing. In Chapter 2, after an introductory chapter, I flesh out a new 'endorsement' approach to the Capgras delusion, the main idea being that the delusion is formed by endorsing the content of a metaphorical-perceptual state in which a loved one is represented metaphorically as an imposter. In Chapter 3, I take issue with the idea, prevalent in the literature, that Bayesian accounts of inference in delusions are best characterised as versions of 'explanationism' (the view that delusions arise via an inferential process grounded in experience). Taking one such influential account as a test case, I argue that it is no more or less compatible with the endorsement alternative, which says delusions are acquired non-inferentially from experience. In Chapter 4, I examine a recent critique of predictive processing theories of delusions, according to which such theories suffer two limitations: they fail to explain why an agent believes something delusional as opposed to nothing at all or something else; they fail to explain how delusional hypotheses are generated in the first place. I suggest ways in which these concerns might be addressed. In Chapter 5, I consider how the Spinozan theory of belief fixation (the view that sees such processing as strictly unconscious and reflexive) can improve the prospects of doxasticism about delusions (the view that they are beliefs). Doxasticism has been criticised on the grounds that delusions typically do not abide by rational standards which we expect beliefs to conform to. If belief fixation is Spinozan, I argue, these deviations from rationality are not just compatible with, but supportive of, their status as beliefs. In Chapter 6, I ask whether misidentification delusions formed via endorsement processes (MDs) have any measure of perceptual justification. To give the best chance of getting a 'yes' answer, I address the question from the perspective of perceptual dogmatism, as this appears the best option if we are to grant at least defeasible justifiedness to MDs. I argue that, even so, MDs are completely unjustified.

Acknowledgements

Writing this thesis has been a long process, with many twists and turns along the way. I cannot possibly mention everyone who helped me see it through, but I will make my best attempt. I apologise in advance to those I have inadvertently left out.

My first thanks must go to the Arts and Humanities Research Council which financed my doctoral research, and granted travel funding for research training and conference attendance.

I am obliged to my friends and colleagues in the department of philosophy at the University of Birmingham, past and present, who have been throughout a constant source of insights and inspiration: Alex Miller Tate, Matilde Aliffi, Eugenia Lancellotta, Kash Sunghuttee, Thomas Davies, Valeria Motta, Helen Ryland, Casey Elliott, Will Sharp, Michael Roberts, and Tom Baker. Among present and former faculty members at Birmingham particular thanks are due to Will Davies, Maja Spener, Henry Taylor, Salvatore Florio, Nick Jones, and Nick Effingham, for their support, personal and professional, and for encouraging me to believe in my work.

Matthew Parrott deserves special mention for invaluable written comments that were instrumental in my preparing Chapter 2. I also thank him for sharing some of his own work in advance of publication, the reading of which prompted me to write Chapter 4.

I am grateful to Matthew Ratcliffe and Sam Wilkinson for teaching me how to write when I was a Master's student at Durham. Matthew first sparked my interests in delusions and offered early encouragement to pursue a PhD. Sam's reassuring presence kept me going through some tough times and restored my confidence when most needed.

In the past four years, I have been fortunate enough to present my research at different conferences and institutions, and have benefitted enormously from discussions with numerous individuals, including Peter Carruthers, Ryan McKay, Garry Young, Neil Van Leeuwen, Matthias Michel, Jorge Morales, Jake Quilty-Dunn, Adam Bradley, Quinn Gibson, and Greyson Abid.

My heartfelt thanks to all those who generously contributed their expertise in lengthy email conversations, which helped me get much clearer on several thorny issues: Elijah Chudnoff, Eric Mandelbaum, Karl Friston, Christopher Peacocke, and Katerina Fotopoulou.

I am much indebted to Phil Corlett and the members of his Belief, Learning, & Memory Lab at Yale University, where I spent some very memorable time as a visiting

researcher. Phil has been an outstanding adviser and mentor to me. I want to thank him for countless conversations that helped shape my thinking, as well as for initiating me to the delights of New Haven pizza. I should also like to thank Pantelis Leptourgos, Erin Reed, and Jane Garrison for offering many insightful suggestions, and for making New Haven feel like home.

Thanks to Anders Nes and several anonymous reviewers (especially two reviewers for *Mind and Language*) whose valuable feedback and recommendations on earlier drafts contributed considerably to the final product.

A warm expression of appreciation to my PhD supervisors Lisa Bortolotti, Ema Sullivan-Bissett, and Craig French, without whom this thesis could not have been brought to completion. Lisa, my primary supervisor, was an extraordinary interlocutor who sustained me in more ways that I can say or ever pay back. She has been abundantly generous with her time and offered plenty of sound advice in my journey across the pond. Not only did she put up with my persistent, at times obsessive preoccupation, but she did so graciously. Thank you for inspiring me as a role model in my academic and personal growth. Ema was the best secondary supervisor anyone could ever hope for: caring, engaging, patient, rigorous. She has been an ever-willing sounding board for (often unstructured) ideas, tireless reader, and enthusiastic supporter. Her unwavering confidence picked me up in the darkest moments, and for that I will always be thankful. Craig, my third supervisor, is one of the sharpest analytical minds I have had the pleasure to know. His meticulous attention to detail has saved me from many an error and inconsistency, prodding me to rethink several parts of the arguments advanced herein. Thanks to his efforts, Chapter 6 was made far better than it would have been otherwise. Needless to say, all mistakes that remain are mine and mine alone.

My most important debts are to my partner Anastasia, who patiently tolerated my frequent mood swings over the past few years, and to my parents, Francesco and Antonella Bongiorno, to whom this thesis is dedicated.

***

Chapter 2 is a slightly revised version of a single-authored paper published under the same title in *Mind and Language*, 35, 3, 2020. The work presented in Chapter 3 was co-authored with Lisa Bortolotti. It appears in almost identical form in *Inference and Consciousness*, edited by

Anders Nes and Timothy Chan, (New York: Routledge, 2020). Work presented in Chapter 4 was co-authored with Phil Corlett. It is now under review.

Contents

Chapter 1: Introduction

1.1. Some Background

From 1963 through 1976, Bert Yancey was one of the world's best golfers. In his most successful year, 1974, he was in Japan to promote golfing in the country. While there, he began to think he was the Messiah chosen by God to rid the Oriental world of Communism. A few months later, after playing in a tournament at Westchester, N.Y., Bert climbed a ladder inside LaGuardia Airport, shouting that the billionaire financier Howard Hughes had been in touch with him to dispense money for cancer research. He was then overpowered by security guards and taken into a quiet room where, as he later recalled, '[he] was spitting on a light bulb, thinking if [he] watched the saliva burn, the different colors and shapes, [he] could find the key to the cure for cancer' (Goodwin and Jamison, 1990, p. 27).

A 28-year-old, college educated, fast-food worker was presented to the emergency department in police custody after neighbours reported a disturbance. Upon arrival, officers found him kneeling in his mother's lawn, where he remained until he was escorted off. He said that the 'hit men' hidden across the street had their 'deadly ray guns' trained on him, ready to fire at a moment's notice. He later arranged for a transfer to the state hospital because the personnel at the academic medical center had been 'infiltrated by the hit men' (Lake, 2008, p. 1157).

YY, a 20-year-old college student, was referred for neurological evaluation after a catatonic episode. The first day after she got home, her father tried to open the front door only to find it locked from the inside. As she saw him ringing the bell, she phoned the police and told them that 'there was an imposter outside the house who was picking the lock and pretending to be her father'. Several months after onset she could still not identify her father. She persistently kept enquiring about her own life with the aim of outing him as an imposter (e.g., 'Where did I have my 10th birthday party?'). When asked about her real father's whereabouts, she replied that her mother and the imposter murdered him so they could be together (Brighetti et al., 2007, p. 191).

Each of the above individuals were clinically diagnosed as having *delusions*. Bert Yancey had delusions of grandiosity, in which one arrogates to oneself exaggerated worth and status, such as being handpicked by God. The 28-year-old fast-food worker had delusions of persecution, and YY had the Capgras delusion, the delusion that someone close to one (typically a spouse or family member) has been replaced by a duplicate or imposter.

What are delusions? And if there are different types of delusion, what do they have in common that makes them all types of delusion? Let me begin with some brief stage-setting. Delusions occur in a wide variety of medical and psychiatric conditions (including, but not restricted to, schizophrenia, bipolar disorder, delusional disorder, dementia, major depression), all with diverse underlying causes (such as brain lesions, genetic susceptibilities, endocrine disruptions, drug and alcohol abuse, and infections). Sometimes they persist in a person for years with intact or minimally impaired everyday functioning. In other cases, they are extremely frightening and interfere with all daily activities, social and family life. This is the case of people with persecutory delusions, who live in constant fear of being watched, followed, or harmed, and can refuse to leave the house, or even their bedroom, as a result. Depending on their type, delusions can aid in the diagnosis of particular disorders. For example, thought insertion (i.e., the delusion that foreign thoughts are being placed into one's mind) is generally a reliable marker of schizophrenia, whereas the Cotard syndrome (i.e., the delusion that one is dead, and that body parts/organs or rotting or missing) is typically associated with major depression.

There is no solid, unproblematic definition of delusion on the table. The definition in the latest version of the Diagnostic and Statistical Manual of Mental Disorders, fifth edition (APA, 2013, p. 87) differs considerably from that provided in the previous version, the fourth text-revised version (APA, 2000, p. 299). The latter is a more detailed definition. It says that a delusion is 'a false belief based on incorrect evidence 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'. But this definition has been found wanting in several respects (Coltheart, 2007, p. 1043; Davies et al., 2001, pp. 133–134).

To begin with, delusions need not always be empirically false, provided that they are based on no good reason. A person with delusional jealousy might speak truly when she says that her husband is having an extramarital affair. But when asked how she found out, she might say it is because a bird flew by and perched on a nearby tree. Occasionally, then, a cognition may count as both delusional and true, if only coincidentally. It is also not clear whether delusions are always or ever beliefs (more below). Nor is it the case that *all* delusions are based on inference and rooted in external reality. Thought insertion or the Cotard delusion do not cease to be delusional because they are not about external reality. Finally, it is not obvious why a delusion could not in principle be held by all members of one's

community, since there are already documented cases of delusions where more than two people are involved (Kelly, 2009).

To avoid these difficulties, the DSM-V characterises delusions in a minimalistic and deflationary way, not as false, but as 'fixed beliefs that are not amenable to change in light of conflicting evidence'. There is a core idea that has remained unchanged from one definition to another: delusions involve departures from the norms of epistemic rationality. Whatever evidence one has (if any at all) for a delusional proposition is outweighed by counterevidence. If we are to accept the bulk of DSM definitions, then, we are at least committed to thinking of delusions as irrational beliefs. This too, however, has been criticised on different grounds.

For one thing, it is dubious that violating epistemic norms is a sufficient condition on being a delusion, even in cases where violations result from gross negligence, e.g., cases where the delusion relies on no evidence at all, or there is overwhelming evidence against it. As many have pointed out, equally striking instances of irrationality can be observed in the non-clinical population (Bortolotti, 2010, 2012, 2018; Bayne and Fernández, 2009; Bayne, 2017). A survey conducted by Public Policy Poll in 2013 revealed that around one in four Americans believed Barack Obama was the antichrist.[1] Epistemically speaking, this belief is no better off than the delusion that one is the Messiah. The reason why we do not regard it as delusional is that conspiracy theories have long been intermingled in American political culture. So, if one wants to argue that being grossly irrational is sufficient for a cognitive state to count as delusional, then one must provide us with an effective way of demarcating delusional rationality from everyday irrationality (Bayne, 2017).

For another thing, it is not certain that every delusion is epistemically irrational, at least if one means by this that the evidence in their support is always overturned by evidence against them. Consider delusional parasitosis (DP), where the individual is convinced that they are infested by parasites, such as insects, mites, fungi, or other organisms. DP is often accompanied by tactile hallucinations of insects crawling in or upon the skin (formication) and related cutaneous symptoms like itching, burning, or stinging sensation. These experiences are very real, and can be extremely unpleasant, leading, in some cases, to self-mutilation, such as picking at one's skin with tweezers to uncover the parasites (Meehan, 2006). Now, suppose one believes that one has bugs underneath one's skin on the basis formication. Should we say that the belief in question is irrationally formed? Perhaps, but

[1] Public Policy Polling. (2013). 'Democrats and Republicans Differ on Conspiracy Theory Beliefs'. Web. https://www.publicpolicypolling.com/polls/democrats-and-republicans-differ-on-conspiracy-theory-beliefs/

perhaps not, since beliefs acquired non-inferentially from convincing hallucinations are intuitively justified. Also, can we be sure that it is irrational to weigh an experiential fact of the matter of what it is like for one (i.e., a sensation of bugs crawling beneath the skin's surface) against background beliefs (e.g., 'it is improbable that bugs should crawl below the skin') or the opinions of clinicians (e.g., skin samples testing negative for parasites)? After all, rare diseases may not affect many people, but they do affect some, and are notoriously hard to diagnose. As such, it seems reasonable (or at least not irrational) to seek out a second opinion if dissatisfied with a physician's explanation for symptoms.

1.2. Overview

In keeping with current orthodoxy, I assume that delusions are a unitary psychological kind—that whenever we attribute a delusion to someone, we attribute the same fundamental kind of mental state (Frankish, 2007, p. 2). Given this assumption, there are at least three respects in which delusions are exceedingly puzzling. First (i), there are questions about their mode of formation. How do they arise? Why do they acquire the precise content that they do? Second (ii), there are large issues concerning their nature. It is unclear whether delusions are a sui generis, previously undiscovered class of mental entities, perhaps hybrid states of some kind, or whether they can be reduced to folk-psychological attitudes we already countenance; and if the latter, whether they are beliefs or something else, e.g., imaginings misidentified as beliefs (Currie, 2000; Currie and Ravenscroft, 2002). Third (iii), there is the question of their epistemic standing. Is acceptance of delusional propositions ever underwritten by epistemic reasons? And if so, are these reasons ever strong enough to outweigh the counterevidence one may come to possess?

Each of these issues is, of course, connected with at least one another. Causal accounts of the origins of delusions raise questions as to their epistemic standing. For example, that a certain psychological process is involved in acquiring a delusion might have a bearing on its justificatory status. Understanding the psychological foundations of delusions might give us insights into what they are. Vice versa, knowing whether or not delusions are beliefs might give us insights into where to look for their causes. Lastly, some delusions may be epistemically irrational in a way that is incompatible with their being beliefs.

This thesis is a collection of self-standing essays revolving around (i), (ii), and (iii). My primary emphasis is on (i), because I think therein lie subtler but important questions

that either are neglected or have only recently come into focus. A related task is to uncover hitherto unexplored connections between (i) and (ii), and between (i) and (iii) respectively.

In what follows, I shall introduce the more specific themes that will run through the thesis by providing an overview of each of the subsequent chapters. On one extremely influential view (Maher, 1974, 1988), delusions are best thought as responses to unusual experiences. This view, which, following John Campbell (2001), we will call 'empiricist', has two variants. The first, which is Maher's own, affirms the *explanatory* character of the delusion-experience relationship, and for this reason it is known as 'explanationism' (Bayne and Pacherie, 2004). Perhaps the starkest illustration of this approach is a neuropsychiatric hypothesis of the origins of the Capgras delusion advanced by Hadyn Ellis and Andy Young (1990). Ellis and Young (1990) suggest that the impairments underlying Capgras are the reverse of those underlying prosopagnosia, a neurological disorder characterised by the inability to recognise familiar faces. Some people with prosopagnosia retain intact capacity for 'covert' recognition of faces, as shown by physiological signs from their autonomic nervous system, i.e., enhanced galvanic skin response (changes in hand sweating) at the sight of a familiar face. By 'covert' what is meant is that while these people may no longer be able to recognise familiar faces consciously, they continue to detect familiarity at an unconscious level. Some people with prosopagnosia, then, present a dissociation between impaired 'overt' recognition (conscious awareness of familiarity) and unimpaired covert recognition. This is possible because there are two neural pathways capable of face recognition, and damage to one (in this case, the pathway to overt recognition) can occur without affecting the other (Bauer, 1984).

Now, consider an exactly reversed situation, one where overt recognition is intact, but covert recognition is damaged. In such a situation, a person might consciously recognise a face as belonging to a familiar person, and yet lack (or show reduced) autonomic reaction of familiarity. Autonomic reactions, or absence thereof, do not register consciously. Still, it is plausible to think that the abnormal absence of autonomic activity could generate conscious experience, perhaps a generic sense of something being amiss with the familiar-looking face. This experience would likely be odd, and would prompt one to go searching for a way to explain it. One such possible explanation is that the face seen looks familiar but belongs to somebody unfamiliar, a replacer posing as the 'real' person. According to Ellis and Young (1990), this might be what happens in the case of the Capgras delusion.

A second important variant of empiricism is the 'endorsement' theory. Different from explanationists, who hold that delusions arise from an attempt to explain some irregular experience, endorsement theorists argue that delusions are formed by taking the experience at face value (Pacherie et al., 2006; Pacherie, 2009). According to explanationism, the content of the experience is considerably less specific than the content of the corresponding delusion. By contrast, the endorsement theory takes the content of the experience to be closely linked with, if not identical with, the content of the resulting delusion. In Capgras, the content of a person's experience would be something like 'That is not *y*' [where *that* is a perceptual demonstrative, and *y* is somebody intimate to the perceiver, say, their mum], or even more directly, 'That is a replacer of *y*'.

This is a promising theory, but it faces at least three challenges. First, one must indicate the neuropsychological anomalies that could produce an experience with either of the contents above. How do we get from the absence of autonomic arousal at the sight of Mum to an experience of Mum as being not really her, or worse, as being an imposter? Second, one must show that these contents can at all be contents of experience. The question of the admissible contents of experiences (i.e. which contents experiences can represent) is one of the most controversial questions in philosophy of perception; and it is especially controversial whether experiences have contents such as that someone is Mum, or that someone is an imposter. Third, and finally, one must explain how, if at all, an experience could have those contents without top-down loading from the delusional belief, for that would mean experience cannot be regarded as itself the source of the delusion.

The overall aim of Chapter 2 is to provide an adequate fleshing out of the theory. To achieve this, in the first part of this chapter, I consider some ways in which philosophers have tried to address the foregoing problems, and explain why I find their proposals unsatisfactory. In the second, I develop and defend an alternative model which I hope avoids the pitfalls of the other models.

My proposed model builds off of Christopher Peacocke's notion of 'metaphorical perception' (Peacocke, 2009), according to which it is sometimes possible to experience something metaphorically-as something else. Peacocke (2009) gives the example of a painting of four pots which many report seeing as a group of people. For a subject who enjoys this kind of experience, it does not look to her as though there are people in front of her, as it would for instance if she were hallucinating. Nor does she see the painting as a depiction of people. Yet seeing the pots as people seems like a unitary experience, not a composite of

parts, such as, say, a combination of visual experience and imagination. Peacocke (2009) calls this type of experience 'metaphorical', because he thinks it shares the same underlying structure as conceptual metaphors (e.g., Lakoff and Johnson, 1980), namely, an unconscious mapping from a source domain (pots, in this case) to a target domain (in this case people). I propose that the experience of seeing a loved one without the usual autonomic reaction involves a similarly perceptual process: a person has an experience in which someone looks like their loved one, but is perceived metaphorically as a replacer. On my view, the cognitive processing which initially prompts the thought 'That is not *y*' is unconscious. However, tokening such a thought creates an isomorphic mapping between the domain '*y*-lookalikes' and the domain 'replacers of *y*', which leads to an experience with the metaphorical content 'That is a replacer of *y*'. This view, I argue, offers a potential escape from each of the three sorts of problems above. I conclude the chapter with an explanation of why the metaphorical nature of the representation is not ultimately a bar to its being endorsed as belief.

The empiricist theories mentioned earlier share a common feature: they are pitched at the personal level. In the endorsement theory, delusions are consciously developed via a personal-level endorsement of experience. In Maher's variant of explanationism, delusions arise from a person's reasoned attempt to explain experience. Other existing theories depart considerably from the empiricist picture, as they consider delusion formation to unfold largely or wholly from subpersonal unconscious processes. Two such theories are the focus of Chapters 3 and 4 respectively: the two-factor theory (or at least a prominent version of the two-factor theory, which I call the 'Coltheart model', see Coltheart et al., 2010) and the predictive processing (PP) theory (e.g., Corlett et al., 2010, 2016; Hohwy, 2013; Clark, 2016).

These theories are in one respect akin to each other: they both hold that delusions arise through a process of Bayesian inference. There are three components to this 'Bayesian' approach. First, there is the notion of perception as 'unconscious inference': perception is a matter of unconsciously inferring causes in the world from sensory-driven patterns of neural activation. Second, there is the idea that this kind of inference proceeds in accord with Bayes's rule, which prescribes how to combine neural activation patterns with prior expectations in order to produce the 'best guess' as to what is in the world. Third, the neural activation patterns that the inference takes as input are, themselves, characteristically unconscious.

In other important respects, however, the Coltheart model is different from the PP theory. Here I will mention only three. First, the Coltheart model is developed in the context

of cognitive neuropsychiatry, and is chiefly concerned with monothematic delusions (i.e., delusions restricted to a single theme) of neuropsychological origins—the Capgras delusion, for example (though see Coltheart, 2013). On the other hand, the PP theory is developed in the context of computational psychiatry, and is primarily about schizophrenic delusions, which are typically polythematic (though see Corlett et al., 2010). Second, and most importantly, the Coltheart model posits two neurocognitive impairments (factors) to explain delusions, one in perception (which explains the content and the adoption of a delusional hypothesis), and one in hypothesis evaluation (which explains the persistence of a delusional hypothesis after its initial adoption). Instead, the mechanisms of delusions formation and maintenance under the PP theory involve a single deficit in the processing of prediction errors (i.e., mismatches between predictions and input).[2] Finally, the Coltheart model refuses any explanatory role to conscious experience prior to the formation of the delusional hypothesis (Coltheart et al., 2010, p. 264), whereas the PP theory allows that consciousness may be engaged at some stage in the process.

In Chapter 3 I ask where the Coltheart model stands with respect to the two varieties of empiricism discussed earlier, namely, explanationism and the endorsement theory. The Capgras delusion is the paradigm case here, so I will focus specifically on this. Strictly speaking, the Coltheart model is not a Maher-type empiricism, at least not if one treats experience as by definition conscious. It does, however, suggest that delusions arise via inferential responses to abnormal 'data' (i.e., explananda that are not available to personal-level consciousness). In Capgras, reduced autonomic activity at the sight of a familiar face constitutes an abnormal datum, and the delusional hypothesis 'That is an imposter' is inferred to explain this datum. Because the hypothesis serves an explanatory function, the Coltheart model has been interpreted as modern version of explanationism (Parrott, 2019; Young, 2014). I argue, however, that an explanatory function as understood here is no less compatible with an explanatory picture than an endorsement one.

Chapter 4 is connected with Chapter 3 in that it considers two objections (proffered in Parrott, 2019) that can be put equally forcefully for any Bayesian approach to delusions. The first objection concerns hypothesis selection: why are delusional hypotheses selected

[2] The idea that the PP theory involves just one deficit—namely, disrupted prediction error processing—is often, but not always, accepted. Some scholars (Miyazono and McKay, 2019; cf. McKay, 2012) have resisted the idea, arguing that a second deficit is involved in the process of precision-weighting, which reflects the expected reliability of prediction errors. I am not persuaded by this argument, because I am not convinced precision-weighting is separable from prediction error processing in the way a two-deficit picture would assume. I will return to this briefly in Section 4.3.

over more plausible candidates, or none at all? The second objection concerns hypothesis generation: why are delusional hypotheses thought of in the first place? I argue that both objections may be fruitfully addressed from within the PP framework. In rough outline, my response to the first objection is that faulty prediction error signalling may not only lead to the adoption of delusional hypothesis, but also prevent suspension of judgment and the selection of more viable alternatives. As concerns the second objection, I point out three possible ways that implausible hypotheses may enter an agent's candidate set, each of which is consistent with PP.

All of the chapters outlined so far deal with some unresolved issues facing theories of delusion formation. Chapter 5 and 6 go on to examine specific ways in which delusion formation may have implications for whether delusions are beliefs and whether they have any degree of epistemic justification.

The guiding idea behind Chapter 5 is that existing defences of doxasticism about delusions are incomplete in one fundamental way: while they construe delusions as beliefs, they are noncommittal about the nature of belief that underpins their construal. My primary aim here is to find a model of the normal formation and revision of beliefs that could provide a basis for a more robust defence of doxasticism. Throughout, I argue that the so-called 'Spinozan theory' (Gilbert, 1991; Gilbert et al., 1993; Mandelbaum, 2014) provides just such a model. Briefly, this is the view that we by default believe whatever propositions we entertain, regardless of whether we have evidence for or against believing them; it is only during a second processing stage, and only if we have the mental energy, that we can label a false proposition as 'false', and cease believing it. If beliefs are as the Spinozan theory describes them, I argue, anti-doxasticism is a non-starter, and the features that are often interpreted as telling against their belief status are easily accommodated within a doxastic framework.

Chapter 6 begins with the assumption that at least some delusions, misidentification delusions in general, and the Capgras delusion in particular, form in the manner envisioned by the endorsement theory, namely by taking perceptual experience at face value. The question underlying this chapter is whether such delusions (MDs) enjoy any degree of epistemic justification. To give a 'yes' answer the best shot at being right, I address the question from the standpoint of a popular theory of perceptual justification, often known as 'perceptual dogmatism' (PD), since this seems, at first pass, best suited to allow for at least the defeasible justifiedness of MDs. Indeed, according to PD, having a perceptual experience

that *p* alongside with the appropriate phenomenology is sufficient to immediately prima facie justify a corresponding belief that *p*, regardless of the experience's aetiology. We get two different versions of PD depending on what is required for an experience to have the appropriate phenomenology. One version says that what is required is that the experience represents *p* as true with *assertive* force (Pryor, 2004; Tucker, 2010; Tolhurst, 1998); in another version, what is required is that the experience represent *p* as true with *presentational* force (Chudnoff, 2011, 2012, 2013, 2016). I argue, first, that MDs' contents are not propositions with respect to which the relevant experiences have presentational force, and second, that even if an experience has assertive force with respect to MDs' contents, it could still not make you justified in believing them. The upshot of this is that MDs are thoroughly unjustified even from the viewpoint of PD.

1.3. Methodology

Many of the questions raised in this thesis (with the exception of some of those raised in Chapter 6) fall under the umbrella of what we might call 'empirically informed philosophy of mind' (e.g., Irvine, 2014). There was a time, not so very long ago, when philosophers of mind thought their role consisted primarily, if not exclusively, in a priori investigations of folk concepts like 'belief' and 'perception' (e.g., McDowell, 1994). That time is past; most would now recognise that a thorough-going philosophical theory about the mind should be answerable to empirical constraints, rather than based purely on armchair reflection.

This shift occurred largely as a consequence of the development of cognitive science, a cross-disciplinary field that brings together results from diverse disciplines such as psychology, philosophy, artificial intelligence, linguistics, and neuroscience, all aimed at greater understanding of the nature of the mind. With the success of cognitive science came the realisation that many allegedly a priori truths were demonstrably empirically false (e.g., Burge, 2010). Some even go so far as to say that we cannot know where there is room for a priori conceptual analysis unless we know something about the objective sciences of the mind (Block, 2014, p. 570). This does not mean purging philosophy of mind of intuition-based assessments; rather, it means ensuring that the intuitions we derive a priori from reflecting on, say, the folk concept of perception are not in conflict with what we know about perception from psychology and neuroscience.

I have so far been talking about empirically informed philosophy of mind in very general terms. Now I need to be a little more precise. Since my theme is delusion, I shall

focus on a relatively narrow range of mental phenomena that are commonly taken to be the most relevant to the matter at hand, namely, perception, inference, and belief. Such phenomena are subjects of lively interest by many researchers in cognitive science, so there is ample room for combining empirical research with philosophical speculation, as we will see shortly. In the same vein, I devote particular attention to the subfields of cognitive science whose work has been applied very fruitfully to the study of delusions, namely, cognitive neuropsychology, a branch of cognitive psychology which uses data from cognitive disorders as a means to test normal models of cognition, and cognitive neuroscience, a field of study concerned with studying the neural mechanisms that underlie cognition.

It will help to further clarify my approach to draw on the now familiar distinction between philosophy *in* cognitive science and philosophy *of* cognitive science (Brook, 2009). The former is what I mean by the phrase 'empirically informed philosophy of mind', a kind of philosophical work that is integral to cognitive science. The latter is a meta-theoretical research programme in the philosophy of science (Samuels et al., 2012). It is concerned with such questions as what types of explanation are better suited for cognition, how much evidence is needed to accept or reject a hypothesis, how to separate genuinely causal from merely accidental correlations, and how to integrate diverse findings into a unified framework of wider explanatory scope. While the role of philosophy *of* cognitive science is well-understood, it is less clear what role philosophy should play *in* cognitive science, because it is difficult to see where philosophy ends and science begins or vice versa. As some have pointed out (van Gelder, 1998), there is not a sharp division of labour between cognitive scientists trained in philosophy and those trained in scientific disciplines. I do think, however, that philosophy plays a particular role in cognitive science. Mention can be made here only of the few areas to which I take my own work to be a contribution.

The first is the imaginative process of generating new hypotheses. Obviously, this is hardly distinctive of philosophy as such. At the core of the scientific method is the interaction between hypothesis generation and empirical testing. Unlike scientists, philosophers often generate hypotheses without any testing other than conceptual coherence. Thus, some worry that, in at least this respect, philosophy is a second-rate method of doing the same sort of thing as can be done more reliably by science (Brook, 2009). To be sure, I do not take myself to be doing experimental science here. I do not collect empirical data in the way scientists do. Nor do I engage in designing and running laboratory experiments. Still, most of the hypotheses I advance are empirical hypotheses about hitherto complex and puzzling data.

They are, if not supported by the evidence, empirically testable, or at least capable of opening up a space for hypotheses that have not yet been thought of, let alone tested. As such, I have no pretensions to be doing something else, more valuable than science does when devising hypotheses. I would be happy to settle for doing as well as science does, while using a different background framework.

A second area of contribution is the interpretation of tested hypotheses or theories, which includes comparing them with others and integrating them into bigger pictures (e.g., ask whether the Coltheart model is a version of explanationism, see Chapter 3), but also working out their implications and limitations (e.g., ask how delusional hypotheses could even be candidate explanations for some experience, see Chapter 4). Philosophers' interpretations provide a vantage point that allows some perspective to be taken with respect to results obtained by others. From this vantage point, we can achieve more objectivity, as well as identify new links between different fields of research.

A third contribution is in clarifying what kinds of concepts should be used to describe delusions and their behavioural manifestations. If delusions are types of beliefs, then there are types of beliefs that are unresponsive to evidence and disconfirmation, badly integrated with one's other beliefs, and weakly behaviour guiding. Conceptual clarification as I understand it here is not just a matter of introspecting and examining our folk intuitions about belief to see if they match these surface features (Bortolotti, 2010, 2012). Nor is it empirical merely in the sense of collecting data concerning the folk practice of belief ascription in cases of delusions (Rose et al., 2014). Instead, the thought is that closer scrutiny of the specifics of belief fixation in light of experimental findings stands to benefit our understanding of the concept of belief, and potentially of delusion as well. If it should turn out that the surface features of delusions can be explained by appeal to belief fixation processes, this would make a strong case for them being grouped under the concept of belief (see Chapter 5).

Finally, a fourth contribution is in using theories about normal cognitive processes, some of which underexplored, to help understand those involved in the genesis of delusions. Most essays in this thesis move in this direction of explanation, but the link between normal and delusional cognition can be explored also in the opposite direction, suggesting a promising avenue forward. For instance, learning about belief fixation in delusional individuals can help us discern which models are viable candidates for understanding belief fixation in neurotypical individuals. Take the Spinozan view as an example. (Recall that

according to this view beliefs are acquired automatically upon the mere representation of a truth-apt proposition, regardless of the evidence). What do delusions tell us about its viability? If acquiring beliefs happens automatically, why do not people become delusional just by entertaining fanciful thoughts? Is there some sort of mechanism for epistemic vigilance that normal individuals have but delusional individuals lack? If yes, does this invalidate the Spinozan view? And so on.

To sum up, the research I present in these pages aims to move beyond a passive consumption of science—whereby empirical work is just used to support and revise armchair philosophical claims—to an active reception—whereby philosophy plays a role within scientific practice itself. In recent decades, scholars have made great strides towards understanding delusions. Nonetheless, many questions remain open. Finding answers to some such questions would constitute a further, important step forward. What follows is an attempt in that direction.

One final remark. My use of the term 'belief' reflects a lingering ambiguity in the literature on delusions. Some key aspects of the concept of belief are generally uncontested. Beliefs are attitudes toward propositions; they answer a direction of fit from mind to world (they aim at fitting the world, rather than getting the world to fit them); and they are poised to guide reasoning and behaviour (e.g., Frankish, 2007). However, the literature distinguishes between two uses of the term 'belief', a binary use and a graded use (e.g, Christensen, 2004). On the one hand, we ordinarily think of belief as an on-off attitude: for each proposition *p*, either one believes that *p* or one does not, with no possibility in between. A binary use of 'belief' is implicit in many accounts of delusions, including prominent versions of the endorsement theory (e.g., Pacherie, 2009) and explanationism (e.g., Maher, 1999). On the other hand, beliefs (often otherwise called 'credences') come in varying degrees of strength, which may take any value between 0 (representing certainty of falsehood) and 1 (representing certainty of truth). Under this interpretation of belief, it is possible for one to believe *p* even as one is not fully convinced of *p*'s truth. For instance, suppose *p* is defined on the proposition 'the Miami Heat will win the next NBA championship'. If I believe that *p*, I may take it to be close to certain that *p* is true (in which case my credence in *p* is marginally lower than 1), or take *p* to be merely more likely true than not (in which case my credence in *p* is marginally greater than 1/2). The degree-of-belief concept is pervasive in Bayesian models of delusions, where an agent's beliefs are expressed in terms of probabilities the agent assigns to propositions about reality.

As we have just seen, there is no consistency in the way writers about delusions use the term 'belief'; while some writers speak of believing as a binary state, other conceive of it as a graded matter. For this reason, I will switch from one meaning to another depending on the particular account that is under consideration. I will not attempt here to grapple with the question of how to reconcile these two notions of belief, although doing so would be a very worthwhile project. The worry might be that 'binary' and 'graded' are not just two variants of a single core state, but two categorically distinct types of belief with different operating characteristics. If this were true, which of these, if either, should delusions be?

Keith Frankish (2007) interprets the binary/graded distinction as coextensive with several others, and he takes this as suggesting that binary and graded beliefs reside in two separate systems requiring different levels of cognitive resources. Beliefs of the first kind (call it type 1) are, in addition to being graded, unconscious (i.e., action-guiding without being consciously entertained), passive (i.e., formed automatically without any deliberate consideration), and dispositional (i.e., organised in clusters of behavioural propensities). Beliefs of the second kind (call it type 2) are, in addition to being binary, conscious (i.e. action-guiding only when consciously entertained), active (i.e., deliberately formed), and functionally discrete (i.e., susceptible to selective activation).

Frankish is certainly right that what we call 'beliefs' are a diverse and heterogeneous bunch. But it is far from settled whether we can group their different properties into two fundamentally separate kinds. Among other things, it is unclear whether Frankish's bipartite distinctions exactly overlap one another. For instance, it is conceivable that one could have a conscious degree of belief in *p*, perhaps by feeling more or less confident in *p*'s truth (Railton, 2013). Likewise, it is conceivable that consciously entertained beliefs could manifest themselves in spontaneous, non-deliberative actions. Suppose you make plans to meet your date at a restaurant at 8 p.m. As you wait for her to arrive, you entertain a consciously occurrent belief that she will arrive soon. Still, it seems possible that your belief will become apparent in behaviour that is not consciously controlled, such as adjusting the neck of your shirt or instinctively checking your zipper.

Regardless, even if we accept that there are two different kinds of mental state that we designate when we say 'belief', it would be an over-simplification to think delusions fit neatly into one or the other kind. It is true that the content of delusions is usually accessible to conscious reflection and verbal report, and it is also true that the attitudes they manifest are communicated in a binary fashion. That is why Frankish thinks delusions fit the profile

of type 2 beliefs (Frankish, 2013). But for one, we cannot exclude the possibility that non-binary attitudes are simply re-expressed in binary terms. Second, it is not at all clear that delusion formation is always active (i.e., occurring through an act of deliberate judgment) as a type 2 statute would dictate. I will have a bit more to say about these matters in Chapter 5, but, for now, suffice it to say that delusions defy classifications into clear-cut types of belief.

1.4. Future Directions

In closing, I shall highlight two considerations for future research. First, the notion of metaphorical perception, as discussed in Chapter 2, is admittedly speculative and further work is called for in order refine and improve upon Peacocke's initial characterisation. We need more evidence that what Peacocke calls 'metaphorical perception' is genuinely perceptual (i.e., and not just a convenient label given to a complex of more standard mental states (e.g., perception plus spontaneous imagination). This is important because if metaphorical perception really is a kind of perception, then we can adopt useful insights from perceptual psychology. In particular, we can ask questions such as whether there are inferences involved in metaphorical perception, and if so, in what way are these inferences related to, or different from, those involved in ordinary perception (assuming both are drawn from things known by previous experience).

My thought at the moment is that metaphorical perception could be helpfully understood as located on a continuum with common illusory perceptions (i.e., non-clinical everyday illusions).[3] Illusions are most likely mediated by top-down priors. That is, prior expectations modulate perceptual content so it conforms to what is expected. Our prior bias for certain objects (e.g., people, faces) can cause us to see them at times of perceptual uncertainty, such as when in a dark field a tree is mistaken as a person. The same effect occurs with the pots, but to a lesser degree. Our priors automatically detect something human-like in the way the pots look (e.g., shape, clustering together), which is why we see them as pots. Unlike with illusions, however, we can partly counter the cognitive penetration of perception, with the result that we end up seeing the pots in two different ways at once (the pots way and the people way), rather than just one. How might this happen? An interesting possibility is that the influence from prior expectations is strong enough to alter our experience, but not strong enough to completely offset the evidence from visual input.

[3] I owe this idea to Jane Garrison.

Another intriguing area for future research will be to monitor advances in the burgeoning new science of belief (for a review, see Porot and Mandelbaum, 2020). Standard defences of doxasticism do not take a stand on whether there is any mental phenomenon of scientific study to fill the role set by the folk concept of belief. The main reason for this neutrality is that we as yet do not have a well-developed model of normal believing. At best then, 'belief' is a temporary expedient, a stopgap, to be used until a suitably scientific notion becomes available. At worse, it reduces to, or is nothing over and above, the kind of thing we call 'belief' in our folk psychological explanations for behaviour.

It is true that there is still no scientific consensus on how beliefs are fixed, stored, or changed—the science of these issues is in its early stages, with a large mass of empirical findings crying out for interpretation. That said, some generalisations have already been attempted. Here I focus only on one such generalisation, the Spinozan theory, according to which belief fixation is mandatory, automatic, and prior to any kind of evaluation or appraisal (again, see Chapter 5). My argument turns on two claims: (a) the Spinozan view is promising enough to make it worth considering its implications for doxasticism, and (b) appealing to this view offers a robust means of cashing out doxasticism. Despite the promise, much more needs to be done before the Spinozan view can be regarded as a successful theory of belief. On the one hand, there is disagreement over whether findings actually support the view. For instance, while some theorists take findings on fictional persuasion as tending to confirm that we form beliefs automatically and reflexively (Pennycook et al., 2019), others interpret the same findings as being compatible with some degree of selectivity in belief fixation (Steglich-Petersen, 2017). On the other hand, recent studies have shown results that seem to speak against the Spinozan view. Perhaps most significantly, Nadarevic and Erdfelder (2019) purport to have demonstrated that the results of one highly cited pro-Spinozan study (Gilbert et al., 1990) fail to replicate under identical conditions. In sum, additional work is warranted to fully assess the theory's degree of empirical support and to investigate why original findings were not replicated. That, however, should not deter us from exploring the implications it has for the status of delusions. At the very least, the discoveries we make when we study delusions provide an important check for candidate theories of belief. All things being equal, preference is given to theories that explain a greater range of facts. So, if I am right that the Spinozan view can successfully explain why delusions are beliefs, I would consider that progress.

1.5. References

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders-TR: Fourth Edition*. Washington: American Psychiatric Press.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders: Fifth Edition*. Washington: American Psychiatric Press.

Bauer R. M. (1984) 'Autonomic Recognition of Names and Faces: A Neuropsychological Application of the Guilty Knowledge Test'. *Neuropsychologia*, 22, pp. 457–469.

Bayne, T. (2017). 'Delusion and the Norms of Rationality'. In T.-W. Hung and T. J. Lane (eds.) *Rationality: Constraints and contexts* (pp. 77–94). Amsterdam: Elsevier Academic Press.

Bayne, T. and Fernández, J. (2009). 'Delusion and Self-Deception: Mapping the Terrain'. In T. Bayne and J Fernández (eds.) *Delusion and Self-Deception: Affective Influences on Belief Formation* (pp. 1–20). Hove: Psychology Press.

Bayne, T. and Pacherie, E. (2004). 'Bottom-up or Top-down? Campbell's Rationalist Account of Monothematic Delusions'. *Philosophy, Psychiatry, and Psychology*, 11, pp. 1–11.

Block, N. (2014). 'Seeing-As in the Light of Vision Science'. *Philosophy and Phenomenological Research*, 89(1), pp. 560–572.

Bortolotti, L. (2010). *Delusions and other irrational beliefs*. Oxford: Oxford University Press.

Bortolotti, L. (2012). 'In Defence of Modest Doxasticism About Delusions'. *Neuroethics*, 5(1), pp. 39–53.

Bortolotti L. (2018). 'Delusions and Three Myths of Irrational Belief'. In L. Bortolotti (ed.) *Delusions in Context*. Cham: Palgrave Macmillan.

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

Brook, A. (2009). 'Introduction: Philosophy in and Philosophy of Cognitive Science'. *Topics in Cognitive Science*, 1(2), pp. 216–230.

Burge, T. (2010). *Origins of Objectivity*. Oxford: Oxford University Press.

Campbell, J. (2001). 'Rationality, Meaning, and the Analysis of Delusion'. *Philosophy, Psychiatry, and Psychology*, 8(2–3), pp. 89–100.

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

Chudnoff, E. (2011). 'What Intuitions are Like'. *Philosophy and Phenomenological Research*, 82(3), pp. 625–654.

Chudnoff, E. (2012). 'Presentational Phenomenology'. In S. Miguens and G. Preyer (eds.) *Consciousness and Subjectivity* (pp. 51–72). Berlin: Ontos Verlag.

Chudnoff, E. (2013). *Intuition*. New York: Oxford University Press.

Chudnoff, E. (2016). 'Epistemic Elitism and Other Minds'. *Philosophy and Phenomenological Research*, 96 (2), pp. 276–298

Clark, A. (2013). 'Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science'. *Behavioral and Brain Sciences*, 36, pp. 181–204.

Clark, A. (2016). *Surfing Uncertainty*, Oxford: Oxford University Press.

Coltheart, M. (2007). 'The 33rd Sir Frederick Barlett Lecture: Cognitive Neuropsychiatry and Delusional Belief'. *The Quarterly Journal of Experimental Psychology*, 60(8), pp. 1041–1062.

Coltheart, M. (2013). 'On the Distinction between Monothematic and Polythematic Delusions'. *Mind and Language*, 28(1), pp. 103–112.

Coltheart, M., Menzies, P. and Sutton, J. (2010). 'Abductive Inference and Delusional Belief'. *Cognitive Neuropsychiatry*, 15, pp. 261–287.

Corlett, P. R., Honey, G. D. and Fletcher, P. C. (2016). 'Prediction Error, Ketamine and Psychosis: An Updated Model'. *Journal of Psychopharmacology*, 30, pp. 1145–1155.

Corlett, P. R., Taylor, J. R., Wang, X.-J., Fletcher, P. C. and Krystal, J. H. (2010). 'Toward a Neurobiology of Delusions'. *Progress in Neurobiology*, 92, pp. 345–369.

Currie, G. (2000). 'Imagination, Delusion and Hallucinations'. *Mind and Language*, 15(1), pp. 168–183.

Currie, G., and Ravenscroft, I. (2002). *Recreative Minds: Imagination in Philosophhy and Psychology*. Oxford: Oxford University Press.

Davies, M., Coltheart, M., Langdon, R. and Breen, N. (2001). 'Monothematic Delusions: Towards a Two-Factor Account'. *Philosophy, Psychiatry, and Psychology*, 8(2/3), pp. 133–158.

Ellis, H. D. and Young, A. W. (1990). 'Accounting for Delusional Misidentifications'. *British Journal of Psychiatry*, 157, pp. 239–248.

Frankish, K. (2007). *Mind and Supermind*. Cambridge: Cambridge University Press.

Frankish, K. (2013). 'Delusions: A Two-Level Framework'. In Broome, M. and L. Bortolotti (eds.) *Psychiatry as Cognitive Neuroscience* (pp. 269–285). Oxford: Oxford University Press.

Gilbert, D. T. (1991). 'How Mental Systems Believe'. *American Psychologist*, 46, pp. 107–119.

Gilbert, D. T., Krull, D. S. and Malone, P. S. (1990). 'Unbelieving the Unbelievable: Some Problems in the Rejection of False Information'. *Journal of Personality and Social Psychology*, 59,

pp. 601–613.

Gilbert, D. T., Tafarodi, R. W. and Malone, P. S. (1993). 'You Can't not Believe Everything You Read'. *Journal of Personality and Social Psychology*, 65(2), pp. 221–233.

Goodwin, F. K. and Jamison, K. R. (1990). *Manic-Depressive Illness*. Oxford: Oxford University Press.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Irvine, E. (2014). 'Problems and Possibilities for Empirically Informed Philosophy of Mind'. In. M. Sprevak and J. Kallestrup (eds.) *New Waves in Philosophy of Mind* (pp. 185–207). New York: Palgrave Macmillan.

Kelly, BD (2009). 'Folie à Plusieurs: Forensic Cases from Nineteenth-Century Ireland'. *History of Psychiatry,* 20, pp. 47–60

Lake, C. R. (2008). 'Hypothesis: Grandiosity and Guilt Cause Paranoia; Paranoid Schizophrenia is a Psychotic Mood Disorder; A review'. *Schizophrenia Bulletin*, 34, pp. 1151−1162.

Lakoff, J. and Johnson, M. (1980). *Metaphors We Live By*. Chicago: Chicago University Press.

Maher, B. A. (1974). 'Delusional Thinking and Perceptual Disorder'. *Journal of Individual Psychology*, 30, pp. 98–113.

Maher, B. A. (1988). 'Anomalous Experience and Delusional Thinking: The Logic of Explanations'. In T. F. Oltmanns and B. A. Maher (eds.) *Delusional Beliefs* (pp. 15–33). Chichester: John Wiley and Sons.

Maher, B. A. (1999). 'Anomalous Experience in Everyday Life: Its Significance for Psychopathology'. *The Monist*, 82(4), pp. 547–570.

Mandelbaum, E. (2014). 'Thinking is Believing'. *Inquiry*, 57(1), pp. 55–96.

McDowell, J. (1994). *Mind and World*. Cambridge, MA: Harvard University Press.

McKay, R. (2012). 'Delusional Inference'. *Mind and Language*, 27, pp. 330–355.

Meehan, W. J., Badreshia, S. and Mackley, C.L. (2006). 'Successful Treatment of Delusions of Parasitosis With Olanzapine'. *Archives of Dermatology*, 142(3), pp. 352–355.

Miyazono, K. and McKay, R. (2019). 'Explaining Delusional Beliefs: a Hybrid Model'. *Cognitive Neuropsychiatr*y, 24(5), pp. 335–346.

Nadarevic, L. and Erdfelder, E. (2019). 'More Evidence Against the Spinozan Model: Cognitive Load Diminishes Memory for 'true' Feedback'. *Memory and Cognition*, 47, pp. 1386–1400.

Pacherie, E. (2009). 'Perception, Emotions and Delusions: Revisiting the Capgras Delusion'. In T. Bayne and J. Fernandez (eds.) *Delusions and Self-Deception* (pp. 107–126). Hove: Psychology Press.

Pacherie, E., Green, M. and Bayne, T. (2006). 'Phenomenology and Delusions: Who Put the 'Alien' in alien control?'. *Consciousness and Cognition*, 15(3), pp. 566–577.

Parrott, M. (2019). 'Delusional Predictions and Explanations'. *British Journal for the Philosophy of Science*, 0, pp. 1–32.

Peacocke, C. (2009). 'The perception of Music: Sources of Significance'. *The British Journal of Aesthetics*, 49(3), pp. 293–297.

Pennycook, G., Tranel, D. and Asp, E. (2019). 'Beyond Reasonable Doubt. Cognitive and Neuropsychological Implications for Religious Disbelief'. In A. Coles and J. Collicut (eds.) *The Neurology of Religion* (pp. 115–129). Cambridge: Cambridge University Press.

Porot, N. and Mandelbaum, E. (forthcoming). 'The Science of Belief. A progress Report'. *Wires Cognitive Science.*

Pryor, J. (2004). 'What's Wrong with Moore's Argument?'. *Philosophical Issues*, 14, pp. 349–378.

Railton, P. (2013). 'Reliance, Trust, and Belief'. *Inquiry* 57 (1), pp. 122–50

Rose, D., Buckwalter, W. and Turri, J. (2014). 'When Words Speak Louder than Actions: Delusion, Belief, and the Power of Assertion'. *Australasian Journal of Philosophy*, 92(4), pp. 683–700.

Samuels, R., Margolis, E. and Stich. S. P. (2012). 'Introduction: Philosophy and Cognitive Science'. In R. Samuels, E. Margolis and S. P. Stich (eds.) *The Oxford Handbook of Philosophy of Cognitive Science* (pp. 3–18). New York: Oxford University Press.

Steglich-Petersen, A. (2017). 'Fictional Persuasion and the Nature of Belief'. In E. Sullivan-Bissett, H. Bradley and P. Noordhof (eds.) *Art and Belief* (pp. 174–193). Oxford: Oxford University Press.

Tolhurst, W. (1998). 'Seemings'. *American Philosophical Quarterly*, 35(3), pp. 293–302.

Tucker, C. (2010). 'Why Open-Minded People Should Endorse Dogmatism'. *Philosophical Perspectives,* 24, pp. 529–45.

van Gelder, T. (1998). 'The Roles of Philosophy and Cognitive Science'. *Philosophical Psychology*, 11, pp. 117–136.

Young, G. (2014). 'Amending the Revisionist Model of the Capgras Delusion: A Further Argument for the Role of Patient Experience in the Delusional Belief Formation'. *Avant: Trends in Interdisciplinary Studies* (3), pp. 89–112.

Chapter 2: Is the Capgras Delusion an Endorsement of Experience?[4]

2.0. Abstract

There is evidence indicating that the Capgras delusion is grounded in some kind of anomalous experience. According to the endorsement model, the content of the delusion is already encoded in the Capgras subject's experience, and the delusion is formed simply by endorsing that content as veridical. Elisabeth Pacherie and Sam Wilkinson have in different ways attempted to articulate a comprehensive defence of this strategy, but here I argue that the endorsement account cannot be defended along the lines envisioned by either of them. I then offer a more promising way of spelling out the model, according to which the anomalous experience implicated in Capgras is metaphorical in character.

2.1. Introduction

A prominent line of inquiry in the philosophical literature on delusion relies on the assumption that delusions are beliefs, and tries to work out how such beliefs are formed.[5] The Capgras delusion is often described as the belief that loved ones (most commonly a spouse or family member) have been replaced by identical or near-identical others (these can take on a variety of forms, such as imposters, clones, replicants, cyborgs, etc.). A more accurate characterisation would distinguish between two propositions that the Capgras subject believes (Aimola-Davies and Davies, 2009): the proposition that the perceived person is not (though she looks like) the person whom they love (e.g., 'this man is not my father'), and the proposition that someone else has taken the loved one's place (e.g., 'this man is an imposter'). Call the content of the former proposition MISIDENTIFICATION, and the latter REPLACEMENT.

On one influential view, known as the endorsement account (Bayne and Pacherie, 2004), the content of the delusion is the same or very close to the content of the Capgras subject's experience, and the delusion is formed when the content of experience is taken as veridical and endorsed in belief. This view has been expressed in different ways. Some take

[4] This chapter is an expanded version of an article published in *Mind and Language*, 35, 3, 2020.

[5] For arguments that delusions are beliefs, see Bortolotti, 2010, 2012, and Bayne and Pacherie, 2004. A distinction is commonly made between monothematic delusions, where people exhibit one or more beliefs concerning a single subject, and polythematic delusions, where people exhibit diverse beliefs encompassing a variety of subjects (see Coltheart, 2013, for discussion). By the term 'delusions' I shall henceforth always mean monothematic delusions.

the endorsed content to be limited to MISIDENTIFICATION, for example, 'This is someone who looks just like my close relative but is not really her/him' (Davies et al., 2001, p. 150), or 'That [currently perceived] woman is not that [remembered] woman' (Campbell, 2001, p. 90). Others take it to be directly that of REPLACEMENT (Fine et al., 2005, p. 148, see also Pacherie, 2009, p. 110; Pacherie et al., 2006, p. 2).

Regardless of which description one prefers, there is the problem of whether the Capgras experience can acquire the sort of rich misidentification or replacement contents that the model says it does. Let us call this the *experiential encoding problem* (following Langdon and Bayne, 2010). Two related problems have been pointed out. The first is the problem of specifying the causal connection between the pathologic mechanism underlying the Capgras delusion and the content of the experience that the delusion is supposed to be an endorsement of. I shall call this the *aetiology problem*. The second is the problem of explaining how the experience in question could be had without its content being inherited from the delusional belief. I shall call this the *top-down determination problem*.

Very few have tried to flesh out an endorsement account of the Capgras delusion that specifically addresses the problems above. A first comprehensive attempt has come from Elisabeth Pacherie (2009) who argues that the endorsement account can be defended (and the problems overcome) if the delusion is interpreted as due to a mindreading disorder, and if the processes through which the Capgras experience is generated count as modular in the Fodorian sense (Fodor, 1983). Pacherie builds on William Hirstein's (2005, 2010) 'representational theory' of misidentification, which is based on a distinction between two ways of representing people: (external) representations of a person's physical properties and (internal) representations of their mental properties. According to this theory, in Capgras there is a disconnection between these two forms of representation, with the result that a known person may look the same but appear different on the inside. More recently, Sam Wilkinson (2016) has presented a model of the Capgras delusion that he suggests can help form the basis for an endorsement account, and it can do so by addressing the experiential encoding problem. Wilkinson thinks the problem can be avoided once it is appreciated that the recognition of qualitative similarity (recognition) and the tracking of individuals (identification) correspond to separate cognitive functions. This claim is motivated by the idea that (a) the Capgras experience can be explained in terms of a failed tracking mechanism and that (b) the tracking of individuals does not involve any sort of rich representational content.

In this paper, my aim is twofold. First, I show that the endorsement account cannot be defended along the lines proposed by Pacherie and Wilkinson. Second, I explore a new model of Capgras delusion that addresses the above problems while lending itself to an endorsement interpretation. I shall rely here on a view presented by Christopher Peacocke (2009), which says that we should appreciate that perception sometimes involves both literal perceptual and metaphorical contents. Drawing on this view, I shall propose that the experience implicated in Capgras is made up of two elements: the literal perceptual content 'this [perceived] individual looks just like that [familiar] individual' and the metaphorical content 'this [perceived] individual is a replacer of that [familiar] individual'. In the following, the notations *x* and *S* will be used to refer to the perceived individual and the familiar individual respectively. According to my proposal, Capgras subjects treat the metaphorical content in a literal way, with the result that they come to believe the delusional proposition that *x* is a replacer of *S*. This is conducive to the idea that the replacement content and not the misidentification claim is encoded in experience.

I shall proceed as follows: In Section 2.2. I contrast the endorsement account to the explanationist account, according to which delusions are explanations rather than endorsements of experience. After discussing both options, I focus on the three challenges that confront the endorsement model. In Section 2.3. I present Pacherie's (2009) attempt to overcome these challenges. I demonstrate that Pacherie's proposal is committed to the truth of two claims: (i) that there is a legitimate sense in which perceptual experience can represent a seen person as unfamiliar; (ii) that perceiving a loved person as unfamiliar in this sense is the same experience as looking at an imposter. I show that both (i) and (ii) do not hold. In Section 2.4. I argue that even supposing these claims are true, they are still incongruous with Pacherie's plea for the modularity of Capgras experience. In Section 2.5. I lay out Wilkinson's approach to delusional misidentification, as applied to Capgras. I argue that the approach's appeal to a failed tracking mechanism rather than misrepresented properties makes it unclear exactly how misidentification is to be involved in the experiential content and, consequently, also which content is to be endorsed in the form of belief. In Section 2.6. I briefly sketch what I see as a more promising way of spelling out an endorsement account of Capgras, which conceives of the endorsed content as a metaphorical content of perception.

2.2. Models of Delusion Formation

2.2.1. Top-Down versus Bottom-Up Approaches to Understanding Delusions

Because delusions are often accompanied by anomalous perceptual experiences, one of the main questions arising is that of determining whether, and in what way, such experiences constitute a source of evidence for the content of the delusion. In this respect, we can distinguish between two ways of accounting for delusion formation. On what have been called 'top-down' approaches (which I put aside here), the delusion is a direct product of organic malfunction, and the anomalous experience is a consequence rather than cause of the delusion (Campbell, 2001; Eilan, 2001). In contrast, the so-called 'bottom-up' approaches view the delusion as originating in experience.

2.2.2. The Standard Bottom-Up Account of Capgras Delusion

Bottom-up views are the most prominent in the literature. The standard bottom-up account builds on Ellis and Young's (1990) suggestion that the Capgras delusion arises from a deficit in face processing. It is widely accepted that familiar face recognition correlates with heightened autonomic nervous system activity, which is measurable in terms of increased skin conductance. Ellis and Young (1990) propose that Capgras results because a neuropsychological deficit causes the face recognition system to become disconnected from the autonomic nervous system, such that one fails to discriminate autonomically between familiar and unknown faces (Ellis and Young, 1990). This hypothesis has been confirmed in a number of experiments (Brighetti et al., 2007; Ellis et al., 1997, 2000; Hirstein and Ramachandran, 1997). According to Ellis and Young (1990), it is plausible that the abnormal absence of autonomic response to the sight of a visually familiar face should generate some kind of unusual experience, based on which the person with Capgras forms the belief that the face seen is that of an imposter.

2.2.3. Explanationist versus Endorsement

There still remains uncertainty, however, as to the exact nature of the experience on which the delusion is grounded. On the explanationist option (Ellis and Young, 1990; Stone and Young, 1997; Maher, 2005), the content of the Capgras subject's experience is sparser than the content of the delusion (e.g., '*x* feels unfamiliar'), and the delusion is invoked as a potential explanation for the anomalous experience. On the endorsement option (Bayne and Pacherie, 2004; Fine et al., 2005; Pacherie, 2009), the experience directly represents '*x* is not

S' or '*x* is a replacer of *S'*, and the delusion is acquired simply by endorsing the experience as veridical. Unless otherwise noted, in what follows 'experience' will be used to mean, more strictly, literal perceptual experience.[6] For ease, I will call the explanationist model *ES*, and the endorsement model *EN*.

2.2.4. Strengths and Weaknesses

Both models come with characteristic strengths and weaknesses. One strength that is often claimed for EN is that it makes a plausible case for the degree of subjective certainty with which the delusion is held. If the delusion really is just a matter of endorsing the content of experience, that might explain why the Capgras subject believes his delusion so firmly. On this account, the delusional conviction is onset at the moment the experience is unreflectively adopted into belief. By contrast, ES would seem to have a harder task explaining why the delusion is held with a strong sense of conviction. Some have noted that if the subject was aware of the delusion playing an explanatory role, she would also be aware of its need for justification. But, they argue, this would be inconsistent with the quality of self-evidence with which the delusion is maintained (Langdon and Connaughton, 2013).

Another suggested advantage of EN over ES is that it provides a rationale as to why the delusion has the particular content that it does. The specificity of content might be explained by the fact that the delusion is simply expressing the way things experientially seem to the subject. It follows that there is no (or little) gap to be filled here between the experiential content and the content of the delusion. By contrast, ES incurs the burden of explaining the relation between the coarse-grained content of experience and the fine-grained content of delusion. If the content of the Capgras subject's experience is simply that there is something unfamiliar about a currently perceived person *x*, where *x* looks exactly like a familiar person *S*, what makes them judge that *x* is not *S* or that *S* has been replaced by *x*? And also, why do they not explore a wider range of hypotheses before taking this route (Gold and Gold, 2014; Parrott, 2019)?

One point on which ES enjoys a clear advantage over EN is in specifying the content of the anomalous experience itself. Explanationist theorists are less demanding with respect to what is packed into the experiential content. All that ES requires is the awareness that

[6] Writers about the endorsement model typically individuate the anomalous experience by reference to its propositional representational content, and conceive of it in literal perceptual terms (see, e.g., Bayne and Pacherie, 2004; Davies et al., 2001; Pacherie, 2009). So the suggestion is that Capgras subjects have a perceptual experience with the literal content that '*x* is not *S*' or that '*x* is a replacer of *S*'

there is something unusual or unfamiliar about the perceived person. Things are different when it comes to EN. To begin, if it is to be part of the representational content of a Capgras subject's experience that '*x* is not *S*' or that '*x* is a replacer of *S*', it might seem that the experience would need to have a content into which *x* enters along with the property of being numerically distinct from *S* or the property of being a replacer of *S*, and this raises the problem of whether experiential contents can include such properties (Davies and Egan, 2013). The problem might be understood as EN helping itself to a controversial position on the admissible contents of experience, namely, that high-level properties of objects in our environment (which is to say, properties other than spatial location, colour, shape, motion, etc.) can be experientially represented (for detailed discussion, see Hawley and MacPherson, 2011).[7] Hence a central task for the endorsement theorist is to demonstrate that the Capgras experience can include representations with the kinds of high-level contents that their account would seem to require (Langdon and Bayne, 2010, p. 339).

It is also unclear how the finding suggestive of diminished autonomic response in face processing is supposed to relate to the endorsement framework. As John Campbell has pointed out (Campbell, 2001, p. 96), the mere diminishment of autonomic response to faces does not of itself constitute an experience with any particular content, let alone contents such as MISIDENTIFICATION or REPLACEMENT.

A different sort of objection to EN, also raised by Campbell (2001), is that the Capgras experience could represent high-level contents like MISIDENTIFICATION only as a result of a top-down loading from the delusional belief. If that were true, that would jeopardise the bottom-up commitment of EN, for which the experience is a causal antecedent of the delusional belief. For if the content of the subject's experience is inherited from the delusional belief, then the delusional hypothesis is not derived from experience, it is prior to experience. So, even on the assumption that the Capgras subject's experience can have the content that the account says it does, a further hurdle remains that must be overcome. Endorsement theorists need to explain how a perceptual state can have such a content without inheriting it from a belief with the same content (Bayne and Pacherie, 2004).

[7] In the philosophical literature on delusions, the problem is typically construed as a problem concerning the array of high-level properties that can be included in the literal contents of perception (although see Section 2.5.). I will argue in Section 2.6. that the question about admissible contents of perception is less of a problem for endorsement theorists if the relevant content is thought of in metaphorical terms.

Two things should be noted. First, the significance of the objection is not tied to the viability of a top-down model. Whilst the truth of this model is incompatible with EN, recognition of its falsity is not logically sufficient for showing that the Capgras experience could deliver the content '*x* is not *S*' without inheriting it from the belief that the person before one is not really *S*. It may well be that the top-down model is false, but one still needs some principled basis for saying that experience can have that content without top-down loading from the corresponding belief. The top-down model and EN cannot be right at the same time, but they can both be wrong at the same time.

Second, endorsement theorists need not deny any top-down down influence on the experience. That is to say, they need not assume that the experience is insulated from any other cognitive features than the delusional belief nor from any other beliefs with different contents. There is evidence that experience is the joint product of both top down (or concept-driven) and bottom-up (or data-driven) processing, such that factors like priming, context, and memory-based expectation modulate incoming information from the level of sensory registration (Powers et al., 2017; Corlett, 2019). Therefore, it is possible that top-down and bottom-up processes are combined in various ways to give rise to the high-level content that EN requires. And considering that sensory registrations alone do not provide principles of numerical identity, and that the Capgras subject's sensory systems seem to be functioning properly, this is a likely possibility (but see Section 2.4.).

To summarise, three main lines of objection have been urged against EN:

1. *Experiential Encoding Problem*: the problem of how experience can carry the sorts of contents that EN vindicates.
2. *Aetiology Problem*: the problem of linking the specific breakdown of Capgras to the relevant experience's having those contents.
3. *Top-down determination problem*: the problem of how experiences with those contents can be had without top-down loading from beliefs with the same contents.

In what follows, I consider Pacherie's (2009) endeavour to overcome concerns (1) to (3).

2.3. Pacherie's Proposal for an Endorsement Account of Capgras

2.3.1. Capgras as a Mindreading Disorder

As we said, it is not obvious how a mere failure to exhibit autonomic discrimination between familiar and unfamiliar faces could plausibly be viewed as prompting the contents of experience that need accounting for in EN. Because of that, one might be inclined to think that Ellis and Young's (1990) model is not the best suited to accommodate EN. Pacherie (2009) tries to respond to this concern by surveying some alternative models for the cognitive and neural correlates of face recognition. She claims that at least one of these models makes a strong case for EN over ES. The model in question was originally postulated by James Haxby and colleagues (Haxby et al. 2000) but Pacherie draws on Hirstein's interpretation of it (Hirstein 2005, 2010).

Utilising fMRI, Haxby and colleagues (2000) found that there are two ventral pathways running via the temporal lobe, which they suggest are functionally responsible for different elements of face recognition. The first (*medial temporal*) pathway specialises in representing the invariant traits that convey face identity information, and remains constant across changes in expression, aging, lighting conditions, or whatever. The second (*lateral temporal*) pathway specialises in the computation of the changeable aspects of faces, such as eye gaze, labial speech-movements, emotional expressions, and the like. Whereas the representation of identity involves activation in the area of the fusiform gyrus, the superior temporal sulcus is proposed as the area responsible for encoding the changeable aspects of faces (Haxby et al. 2000, p. 224).

Building on Haxby's model, Hirstein (2005, 2010) presents an account of Capgras as a mindreading disorder (that is, due to a malfunction of our mind-reading systems). According to Hirstein, the two above-mentioned neural pathways play a differential role in social cognition. The medial temporal pathway is thought to generate what he calls 'external representations' (representations of people's facial and bodily features). Conversely, the lateral temporal pathway would subserve the processing of 'internal representations' (representations of that person's mental states). The perceived face's changeable aspects are indeed informative for tracking other people's mental states (eye gaze, for instance, would tell us much about where another individual is attending and what she is currently preoccupied with).

Hirstein claims that the Capgras delusion could result from a dissociation between external and internal representations of a particular person, where the former are intact and

the latter are damaged, inaccessible, or badly replaced (Hirstein, 2005, 2010). Subjects would be able to recognise the seen face of a loved one, but would present a deficit in activating the correct representation of their loved one's mind (e.g., characteristic emotions, moods, reasons for action, and beliefs). A representation would still be in play, but it would be other than the one subjects have used until then. The resulting state would then be the experience of a person retaining the same look but having a different inner world. As Hirstein puts it, 'for the Capgras' subject, the familiar face is present, but the person is not' (Hirstein, 2010, p. 248).

According to Ellis and Young (1990), what is primarily defective is the autonomic processing of familiar faces. By contrast, Hirstein (2005, 2010) thinks that the primary anomaly stems from the fact that the internal representation of a close person is inaccessible and replaced with a new one.[8] This would create the impression of 'looking at an imposter', and the reduced autonomic responsiveness would be an effect rather than a cause: 'The Capgras' subject is looking at someone who visually resembles his father, but who appears to have a different mental life, a different personality, with different dispositions to do different things. This is exactly what an imposter is, and this is exactly the experience one would have looking at an imposter' (Hirstein, 2005, p. 133).

Pacherie (2009) takes Hirstein's proposal to provide support for the plausibility of EN. The reason is that, she argues, in Hirstein's interpretation the subject's experience is not simply that of a coarse-grained feeling of unfamiliarity about a given person, but directly that of a seen person as 'unfamiliar' in the sense of 'different on the inside'. That being so, the delusional belief could be interpreted not as an explanation of the experience, but as an endorsement of it (Pacherie, 2009, p. 116). There are two claims being made here: (i) experience can represent that a seen person is unfamiliar (or familiar); (ii) seeing someone who visually resembles a close person but has a different mind is experientially the same as looking at an imposter. I will argue that both (i) and (ii) are unjustified. Let me begin with (i). We saw earlier that the question of which properties are represented in experience is crucial for the plausibility of EN. Therefore, the proponent of EN could not assume (i) without further argument, as this is exactly what needs explaining. I take it that Pacherie sees (i) as

[8] For Hirstein, the internal representation system would break as a result of damage to the orbitofrontal cortex and the temporal lobe, which are both thought to be implicated in mindreading (Hirstein, 2005, pp. 101–134).

following logically from Hirstein's approach (or at least from one of its possible readings). But, I argue, this interpretation is not correct.

2.3.2. Unfamiliarity

According to Hirstein (2005, 2010), we perform mindreading by simulating the mental states of another and thus by using our mind to put ourselves into their situation. He indeed concedes that in doing so we may take ourselves to directly perceive such mental states, when we are in fact only modelling them in our mind (Hirstein, 2010, p. 245). It might therefore be that people with mindreading disorders 'misperceive' the mental state another is in, for instance by misconstruing their facial expressions of emotions. If the person whose mental states are being misconstrued is a familiar one, patterns of 'misperception' recurring over time may well give rise to an experience of unfamiliarity or to the inference that the person in question has changed.

But from this claim it does not follow, at least not in any obvious way, that the person in question is being perceptually represented as unfamiliar. Neither can one exclude the possibility that the subject is simply having a non-perceptual feeling of unfamiliarity along with the perception of a person who looks just like herself. It is consistent with Hirstein's (2005, 2010) account that familiarity is an associated but distinct state that accompanies the perceptual experience without being part of what determines its content.

Pacherie might respond that a deficit in mindreading would not only affect the capacity to understand other people's mental states, but it could also interfere with the ability to extract the 'dynamic signature' of a face—a term she uses to denote the distinctive movements that a face makes in expressing a particular mental state (e.g., surprise) (Pacherie, 2009, p. 113). She might argue that normally recognising the distinctive dynamics of a person's face is the same thing as perceiving that person literally as familiar. It would follow that failing to recognise the characteristic dynamics of a familiar person's face (e.g., your father's ordinary way of facially expressing surprise) would be the same as perceiving that person as unfamiliar. However, by assimilating familiarity to the dynamic signature of a face, Pacherie would overlook one important pattern that is found in Capgras delusion. As we know, subjects typically insist that the alleged imposter looks just like the replaced person, suggesting that the visual representation might have retained the same content as before the onset of the delusion (see, e.g., Hirstein and Ramachandran, 1997; Ramachandran and Blakeslee, 1998). If the identification of the dynamic signature was disrupted, we would

expect subjects to give more details about how the imposter's way of animating her face would differ from the replaced person's one. This, however, is not the case.[9]

2.3.3. Imposters

The second step of Pacherie's (2009) argument in support of EN is the claim that perceiving a familiar person as unfamiliar would equate to the experience one would have if one were looking at an imposter. Let us assume for the sake of argument that one could indeed perceive a person strongly resembling one's father as unfamiliar. Recall that by 'unfamiliar' we mean here a person who is represented as having different personality traits (namely: different emotional states, moods, motives, desires, intentions, and beliefs). Would one thereby perceive that person as an imposter? This line of reasoning rests on a basic misconception of the distinction between qualitative and numerical identity. One can remain numerically the same individual despite the sometimes dramatic qualitative changes in body, character, and personality one has gone through.

This point is overlooked by Hirstein and Pacherie because they conflate the sight of a person resembling your father but with a *different mind*, and the sight of a person resembling your father but with a *different identity* (see also Wilkinson, 2015, p. 211; 2016, p. 393). The conflation is nicely epitomised by the following passage: '[…] The patient does not merely see someone familiar as unfamiliar, he perceives that person as having a different identity. […] This is because he sees them as no longer having the same mind, the same motives, moods, and emotions' (Hirstein, 2010, p. 244).

Now, one might speculate that if you were to perceive a person who looks like your father and claims to be your father as numerically distinct (i.e., having a different identity) from your father, you would indeed be looking at an imposter. Yet all that Hirstein's account shows is that Capgras subjects may suffer from mindreading-related perceptual failures; it does not show how this should lead them to perceive their familiars as being numerically different from themselves. One might then try to argue that perceiving a person simply as having a different mind from the person you originally knew would be enough for you to perceive that person as an imposter. But this is wrong; an imposter of *x* is not someone that

[9] Capgras subjects might sometimes refer to minor physical discrepancies between the purported imposter and the original person. A woman, for instance, claimed that she could tell her son had been replaced in that her real son 'had different coloured eyes, was not as big and brawny, and […] would not kiss her' (Frazer and Roberts, 1994, p. 557). Such reports, however, can be plausibly interpreted as post-hoc rationalisations for beliefs which are already held rather than proper perceptual reports (Bortolotti, 2010, p. 46).

merely looks like *x* and has a different personality from *x*, but rather someone who is numerically distinct from *x* and intentionally deceives you into thinking that she is *x*.

If the above is correct, then neither of the claims Pacherie (2009) makes in defence of EN are warranted by Hirstein's (2005, 2010) account. It is not clear that such an account justifies the claim that perceptual experience might be able to represent a seen person as unfamiliar; and it is even less clear why allegedly perceiving some familiar person as unfamiliar should give you the experience of seeing that person as an imposter. If so, then Hirstein's mindreading deficit is not better suited than Ellis and Young's affective deficit to generate the sort of experiential content envisaged by endorsement theorists. Neither does Hirstein's account offer any clue as to how such content could represent high-level properties like that of being an imposter. As such, both the aetiology and the experiential encoding problem are left unsolved.

As we will see in the following section, the strategy Pacherie (2009) puts forth to address the top-down determination problem is even less promising. This is to hold that the processing through which the experience of familiarity is generated qualifies as modular in the sense advocated by Jerry Fodor (1983, 1985). For Fodor (1983), the essential criterion of modularity is informational encapsulation (p. 37). I will argue that the experience of unfamiliarity, as Pacherie conceives it, cannot be informationally encapsulated.

2.4. The Modularity of Familiarity

Fodor's (1983) modularity thesis is a claim about mental architecture. The idea is that the mind contains a set of domain-specific processing systems (modules), each of which is independent from one another and devoted to performing specific tasks. Fodor requires modules to be informationally encapsulated, where this means that the processing carried out within a module is insulated from any information stored elsewhere in the cognitive system. High-level cognitive functions such as reasoning or decision-making are not modules, whereas low-level peripheral systems such as the visual system are. One simple way to appreciate the difference is to note that whereas the latter take well-defined inputs and send well-defined outputs (in the case of vision, sensations of colours, shapes, edges, etc.), the former combines, elaborates, and syntheses the outputs from multiple sensory modalities as well as the information present in non-modular systems (e.g., information in memory).

As we saw, Pacherie (2009) conceives of unfamiliarity not as the mere absence of autonomic arousal in the presence of a particular person, but as the way in which that person

is perceived. For Pacherie, the Capgras experience is not just that 'an experience as of a person that looks like one's father but lacks the feeling of familiarity that normally accompanies this visual experience' but rather 'an experience of the visually presented person as unfamiliar' (Pacherie, 2009, p. 116). For this reason, one might worry that the experience could have that content only as a result of a top-down effect from the delusion itself. As such, Pacherie's view is charged with addressing the top-down determination problem for EN. This motivates the appeal to modularity and informational encapsulation in turn. For to say that a system is modular in the sense of being informationally encapsulated is to say that it is impermeable to any top-down influence.

One striking example that Fodor (1983) cites in support of informational encapsulation is the persistence of perceptual illusions such as the Müller-Lyer illusion. In that illusion, two horizontal lines of equal length appear unequal because of the biasing effects of the arrowheads demarcating them. The bottom line still looks longer even if you believe the two lines to be equal in length.

Pacherie (2009) employs this very same example to illustrate her argument that the processes involved in producing the experience of familiarity or unfamiliarity are informationally encapsulated. In the same way as knowing that the two lines are equal in length will not make them look so, being assured that the person you are looking at is someone you know will not lead you to experience her as familiar if your first experience is of unfamiliarity (Pacherie, 2009, p. 117). This, Pacherie argues, may go towards explaining why the Capgras delusion is so steadily adhered to. If the Capgras experience was not informationally encapsulated, the experience of familiarity would be restored on the basis of background beliefs or the testimony of others, but that is exactly not the case (Pacherie, 2009, p. 120).

This line of argument has some weaknesses. Pacherie's (2009) claim about the modularity, and hence about the informational encapsulation, of the processes through which the experience of unfamiliarity is generated applies both to normal subjects, as well as to subjects with Capgras delusion. However, as far as normal subjects are concerned, informational encapsulation is not necessarily a good explanation of why the experience of unfamiliarity is sometimes not eliminated in the face of disconfirming information. An alternative explanation may be that some top-down loading of the experience is already in place, and that the disconfirming information simply is not strong enough to override it.

Consistent with this, we can conceive of scenarios in which an experience of unfamiliarity is not impermeable to disconfirming information.

Suppose someone approaches you on the train asking if you remember them. As hard as you try, you cannot remember and they are still unfamiliar to you. Seeing your perplexity, they offer you a hint. They tell you that you went to the same primary school together. Whilst you still cannot recognise them, they start to feel more familiar. You guess their name but you get it wrong. They give you a second hint: you played together on the same football team until you were 14. Now they have become fully familiar. You guess their name twice until you get it right. Here it seems we have a simple case where top-down cues can help to re-establish an experience of familiarity that was not initially there. If so, this provides a counterexample to Pacherie's (2009) idea that the experience of familiarity results from modular processes equally across normal subjects and subjects with Capgras delusion.

Yet it might be thought that Pacherie could simply give up the claim about the modularity of familiarity in normal subjects, while retaining the view that this claim applies to subjects with Capgras. After all, the example above only shows that top-down influences can restore familiarity in normal subjects. It does not show that top-down influences can restore familiarity in people with Capgras delusion, for whom the absence of familiarity has a neurological origin. It is indeed possible that, where caused by a somatic condition, the experience of familiarity lies outside the control of top-down influences, in accordance with modularity and informational encapsulation.

However, there are still grounds for doubting that the processes responsible for the experience of unfamiliarity involved in the Capgras delusion, as Pacherie (2009) conceives it, can legitimately be understood in modular terms. Recall that Pacherie's conception of the Capgras experience is modelled upon Hirstein's view of mindreading. On this view, experiencing someone as familiar is a result of one's retrieving and employing an internal representation of that person's mental life. An internal representation aggregates a wide range of information that has been accumulated over time about who that person is 'from the inside', which includes distinctive beliefs, desires, attitudes, and intentions. But this is at odds with the principle of encapsulation. The processing through which the experience of familiarity is generated depends heavily on stored memory and past interactions with the relevant person—and so is unlikely to be insulated from the influence of centrally accessible mental states. It follows that the difficulties raised by the top-down determination problem cannot be resolved by appeal to Fodorean modularity.

In the following section, I shall consider an alternative model of Capgras developed by Wilkinson (2016). Wilkinson's approach promises a straightforward way out of the problems associated with an endorsement interpretation of Capgras, but, as we will see, it is not clear that it helps support such an interpretation.

2.5. Wilkinson's Mental Files Approach

2.5.1 Recognition versus Identification

Wilkinson's (2016) proposal rests on two principles: the distinction between *recognition* and *identification* and the notion that talk of 'mental files' can be fruitfully used to understand how identification goes wrong in Capgras. Suppose you were asked to judge whether a currently perceived individual (*x*) is the same as one encountered in the past (*y*). When we speak of *x* and *y* being the same, we might mean one of two things: that *x* is qualitatively identical to *y*, or that *x* and *y* are numerically identical.

Wilkinson (2016) argues that each of these readings reveal logically different tasks. If the former, then the question is whether you recognise that *x* is like *y* (e.g., that the car parked outside your house is *like* the one that caused a crash the day before). If the latter, then the question is whether you identify *x* as *y* (e.g., the car parked outside your house as *one and the same* with the car that caused the crash). For one to recognise that *x* is like *y* is, argues Wilkinson, is for one to judge that *x* has some property in which it resembles *y*. The judgment is of the form *Fx* (where *F* is the property being predicated and *x* is the individual of which *F* is predicated). In contrast, the act of identifying *x* as *y* simply involves making a connection of identity between *x* and *y*, where the judgment predicates no property at all and takes the form $x = y$.

For Wilkinson (2016), this logical distinction between recognition and identification reflects two distinct cognitive functions: the perceiving of qualitative similarity and the tracking or perceiving of numerical identity. Although judgments of identification are often arrived at by personal-level, abductive reasoning based on recognition of similarities or spatiotemporal continuity, this need not be the case. Indeed, for Wilkinson, there are routes to identification that are not evidence-based, and whose processing bypass both matching of qualitative similarity and spatiotemporal considerations (Wilkinson, 2015, p. 212).

For example, if you had two identical-looking dogs, Fido and Scotty, you might be able to keep track of which is which independently of the properties you perceive them as having, without tracing the spatiotemporal path of either dog, and while being unable to

articulate how you do so. Note that Wilkinson (2016) is not denying that properties and spatiotemporal principles are efficacious in the processes that yield judgments about numerical identity. Rather his point is that such properties should not be understood as evidence for those judgments, in that they are not playing any personal-level inferential role in their formation.

2.5.2. Mental Files

In Wilkinson's (2016) approach, the concept of a 'mental file' (Perry, 1980; Recanati, 1993, 2012) serves as a framework within which to understand what happens cognitively when tracking someone's identity. To a first approximation, mental files are conceptions we build up of others based on information gathered about them. Whenever we meet someone for the first time, a mental file is created and filled with the information available. At each new encounter, the same file is retrieved and updated with newly discovered information.[10] Mental files are 'non-descriptive singular mental representations' (Wilkinson, 2016, p. 397). This is not to say that they do not contain any descriptive information about the individuals to which they refer. It is rather that the correct mental file F for an individual x is retrieved regardless of whatever descriptive match there is between the content of F and the qualities possessed by x at the time. Successful retrieval occurs upon encountering the very same individual for which the file was originally created, no matter the extent to which its properties or location have changed since the first encounter (Wilkinson, 2016).[11]

Wilkinson (2016) draws a distinction between 'Demonstrative' and 'Stable' files (p. 398). The former are context-specific files that ensue from the here-now of a perceptual encounter with a given individual and which take something like this form: 'here now this [person]'. The latter are files which contain a 'stable enough conception' of a person and which can be brought out by reflection outside the immediate perceptual context (e.g., Mum).

[10] Relevant information may come in different forms, such as biographical details, physical appearance, and personality characteristics.

[11] Wilkinson's (2016) view is similar but different from Hirstein's (2005, 2010). According to the latter, the Capgras subject fails to activate the correct mental representation when looking at their loved one, and that gives the impression of a person being different on the inside. On this view, misidentification occurs because the wrong mental properties are attributed to the person. In contrast, on Wilkinson's view, it is not the content of a mental file that fixes the reference to the person, as the information stored in the file does not singularly identify—one can, in principle, identify someone as the same person over time despite major qualitative changes. Rather, reference is fixed *causally*, not *descriptively* by matching of qualitative similarity, and this is thought to explain why misidentification can occur independently of attributed properties.

According to Wilkinson (2016), whenever we encounter someone who is familiar to us, we create a new demonstrative file, retrieve the stable file we have for that person, and merge them together. The processes of creating, retrieving, and merging files are not conscious ones. What is conscious is only the outcome of this chain of processes, namely a state with the content 'this person here present is the very same person as my mother' (Wilkinson, 2016, p. 398). Wilkinson suggests that misidentification (at least of the kind involved in Capgras) occurs when files are mismanaged such that the correct stable file fails to be retrieved in the presence of the person for whom the file was first created. This would cause the subject to judge that the person who is perceptually present and looks like the one they have a stable file for is not that person.

2.5.3. Endorsement of What?

Now, what has this got to do with an endorsement interpretation of Capgras? Wilkinson (2016) thinks of the experiential encoding problem specifically as the problem of whether a person's identity can be represented or misrepresented in experience, that is, the problem of whether one 'can perceive (or fail to perceive) a person's identity directly, without inferring it' (p. 399).

However, Wilkinson (2016) argues, the problem arises for EN only if we assume that numerical identity and numerical distinctness are properties that can be given to us in experience. In the case of Capgras, that would involve not only explaining how experiences can convey information about numerical identity, but also how one can experience the numerical distinctiveness of qualitatively indistinguishable individuals. If I show you two indistinguishable faces before and after a certain interval of time, nothing in the properties you perceive will tell you whether you saw the same face twice or two qualitatively indistinguishable but numerically distinct faces. So how can a Capgras subject perceive the numerical distinctiveness of someone who is perceptually indistinguishable from the familiar?

Wilkinson's (2016) strategy is to argue that the identity of an individual is neither worked out from their perceivable properties nor is it itself a perceivable property, but rather something that is tracked prior to, and independent of, any properties the individual may experientially appear to have. Against this background, Wilkinson (2016) writes, the experiential encoding problem no longer exists, since 'there's nothing mysteriously rich about the content of the Capgras experience […]' (p. 400).

We might also think that the aetiological and the top-down determination problems no longer seem quite so onerous for the endorsement theorist. As for the former, the relation between reduced autonomic arousal and conscious experience may be explained by reference to the failed tracking mechanism. That is, the lack of autonomic response may compromise file-retrieval, which in turn may cause a content-specific misidentification experience. As for the latter, because tracking is the condition under which identification is possible, it seems reasonable to exclude that it could be itself contingent on judgments of identification.

While Wilkinson (2016) may well be right that identification does not necessarily pass through the representation of high-level properties, it is still not clear how identity can enter the content of experience, and if so, what form such a content would take. For this reason, although Wilkinson's view has great potential as a self-standing account of Capgras, it seems to me that it does not offer much help in terms of working out an endorsement account. Let me explain.

As I understand it, Wilkinson's (2016) approach is compatible with either of two different interpretations. On option (1), the physical properties of a certain individual (who is *S*) impacts on one's nervous system, leading to a subpersonal tracking mechanism. Due to a tracking deficit, the information that the individual is not *S* is extracted and the subject has an experience as if that (the individual) is not *S*. On option (2), the information that the individual is not *S* is passed straight on to a misidentification belief, without being first parcelled up into experiential content. We have seen that endorsement theorists share a commitment to the idea that the Capgras experience has representational content, and that such a content can be given by a proposition which specifies the way things are represented as being (e.g., '*x* is not *S*'). An implication of this is obviously that features that enter experiential content are features of the way things appear to the conscious subject. So it seems that if one wants to say that experience delivers a content in which '*x* is not *S*' features, then one needs to make reference to the way things appear from the subject's point of view. But Wilkinson rejects this on the grounds that misidentification can be had regardless, and even despite any observable and appearance properties that things may be experienced as having.

The question that then arises is: in what way is misidentification experiential and also, more specifically, how can it be an experience whose content is available as something to be endorsed? Wilkinson's (2016) analysis is couched entirely in negative terms. In saying that the content of Capgras experience is not rich, he does not offer many specifics as to what an

experience with misidentification content is like, nor about what form this content takes. Wilkinson is explicit that people are not consciously aware of their tracking mechanism, and also that people can keep track of individual things regardless of the ways things are for them experientially speaking (e.g., what and where they are). But if so, what constitutes an experiential way of tracking numerically distinct individuals? And more importantly, what makes a state have a content such that it can be endorsed in belief? Sometimes it seems that Wilkinson wants to appeal to a doxastic view of experience, according to which experiences are belief-like states that coerce one into believing rather than presenting one with a content that can be in principle be rejected (p. 402). Yet this at best explains the way in which the content of experience is entertained. It does not show what form it takes. What is the difference between the content of an experience in which a subject misidentifies *x* as ¬ *S* and the content of one in which a subject correctly identifies *x* as *S*? Without such a clarification, it is hard to see how (or even that) the mental files approach does support an endorsement interpretation of Capgras. Indeed, as it stands, option (2) seems to me the only fully intelligible way to understand how misidentification can emerge into consciousness on Wilkinson's account.

2.6. Can You See Someone as an Imposter?

The aim of this section is to develop an alternative way of understanding what constitutes endorsement in Capgras. Endorsement theorists have tended to characterise the content of experience as including at a minimum MISIDENTIFICATION, namely the claim that '*x* is not *S*'. As we have seen, Wilkinson's (2016) strategy does not really address the central question of what kind of experiential content MISIDENTIFICATION is supposed to be, and this makes it unclear why we should suppose that there is any experience at all prior to the formation of the misidentification belief. The fact that conscious experience may play no role in the emergence of the misidentification belief does not threaten Wilkinson's (2016) overall account of delusion misidentification. As he himself concedes, 'perhaps by the time we get to conscious experience the misidentification is already there' (pp. 402–403). But if MISIDENTIFICATION is not an experiential content to be endorsed, this leaves two possibilities for the endorsement theorist: either abandon the idea of endorsement altogether, or suppose that REPLACEMENT alone is endorsed. In what follows, I explore this latter possibility. I first explain why I think REPLACEMENT is not a perceptual content in the literal sense (a content the experience represents literally as such), and then consider in what

sense it might be metaphorical in character (a content the experience represents metaphorically as such). For simplicity, I will use here 'imposter' as a broad term to refer to any form of replacer with the intention to deceive.

2.6.1. Seeing Someone Literally-As an Imposter

One straightforward sense in which we experience *x* as *F* is that in which we perceive something literally as an instance of a kind, as when we see a squirrel as a squirrel; or, when the lighting conditions are not optimal, the squirrel as a rat. Several philosophers have regarded this species of experience as involving the activation of conceptual capacities (e.g., Brewer, 1999; McDowell, 1994). In a recent paper Peter Carruthers (2015b) remarks that we can consciously see something as something of a certain kind only if 'the concept that represents that kind is bound into the object file that nonconceptually represents its other properties' (p. 503).[12] As will become clear shortly, the reason for focusing on this conception is that it offers not only a neuroscientifically informed framework for establishing which properties can be literally represented in perception, but one which allows the property of being an imposter to be among them.

According to Carruthers (2015b), the binding of a concept *F* into your perception of a certain object *x* is what allows you to perceive *x* literally as *F*. This is illustrated by reference to the well-known image of a Dalmatian dog in a black-and-white speckled background (e.g., Marr, 1982). As Carruthers points out, even if you know that the image represents a Dalmatian dog (assuming that you have never encountered the image before), it will take you a little time to recognise the set of black blobs as a Dalmatian. What happens during that time interval is that you are having an experience that non-conceptually represents that there are such-and-such black patches while concurrently entertaining the thought that the image depicts a Dalmatian dog. Carruthers's (2015b) suggestion is that you are able to perceive the blobs as parts of a Dalmatian only when the concept Dalmatian and the nonconceptual representation of the blobs come to be bound together into a single object-file (p. 503).

[12] The idea of object files has been used in vision science to denote location-based indexes whose role is to track the identity of an object across space and time (e.g., Kahneman, Treisman and Gibbs, 1992). Carruthers (2015a) suggests that object files incorporate both conceptual and non-conceptual representations of the object's properties (p. 66). The content of the object-file 'that' produced by, say, the perception of an approaching car might be non-conceptual to some extent (yellow, curvy, moving), but conceptual to another extent (new, fancy, car).

This story is consistent with emerging insights from neuroscience. For instance, a study by Wyatte, Jilk and O'Reilly (2014) presents data illustrating that feedback from higher-order visual areas implicated in semantic representation—specifically in the inferotemporal cortex (IT)—begin having an influence on object recognition-related tasks as early as 100 ms after stimulus onset, which is well before slower top-down attentional processes come into play (at around 200 ms after stimulus presentation). Perhaps the best illustration of this comes from scenarios in which the missing parts of partially occluded objects are filled in by the visual system.

Wyatte and colleagues (2014) introduce, as an example, a case where IT neurons respond to an occluded bicycle stimulus (i.e., with nothing in sight but the wheels). In the circumstances only a small fraction of the neurons will fire—those associated with wheel-shaped features—whereas the remainder population will remain dormant. However, IT neurons will send feedback signals back to earlier visual areas, activating neurons that selectively respond to visual features associated with bicycle wheels (e.g., a bicycle's frame, seat post, and saddle) despite the absence of relevant physical stimuli. The consequence is that IT neurons respond as if the bicycle was fully visible, such that 'there is little-to-no difference between the response to the partially occluded object and the complete object' (Wyatte et al., 2014, p. 4).

Carruthers (2015b) cites cases like the filling-in case above as evidence that conceptual information is used by perceptual systems to test the input signal from lower visual areas, so as to estimate the best conceptual fit for the non-conceptual content of an object-file (p. 502). Now, if Carruthers is right that concepts can be bound into the literal content of perception, then the question arises, what are the limits of abstractedness of the concepts that can be so bound.

According to Carruthers (2015b), the only restriction that applies is that the applicability of the concept in question must be processed quickly enough for binding to take place: 'conceptual information will need to be processed within the window of a few hundred milliseconds that elapses between presentation of a stimulus and its subsequent global broadcast [i.e., its becoming available across a wide range of cognitive systems]' (p. 504). The upshot is that if the applicability of a concept *F* is processed with sufficient speed, then it must be possible for someone to perceive something literally as *F*.

Perhaps the endorsement theorist can draw on such a possibility to argue that the content '*x* is an imposter' can be part of the literal content of the Capgras subject's

perception. Imagine the following situation: as you enter your father's room, you catch a stranger with your father's mask rolled up to his forehead. If we accept Carruthers's account, there is no principled reason why the concept *imposter* in the circumstances could not be processed fast enough to become integrated into the perceptual state itself. However, this is not the scenario Capgras subjects find themselves in, and so no generalisation can be had. It is true that the appearance of the misidentified person's face is not inconsistent with her being an imposter in disguise (a good imposter would presumably try and look as close as possible to the replaced person). But as opposed to the case of the masked man, there is nothing in the behaviour of the misidentified person to indicate that she is an imposter.

Based on this, we could presume that the concept to be bound into the literal content of perception is the one corresponding to the identity of the person in question (say, father), and that the processing of the concept imposter occurs at a later stage (after the perceptual output is broadcast into the impaired affective appraisal stage). If I am right about this, Capgras subjects do not perceive the misidentified person literally as an imposter, and whatever content their perceptual experience may literally represent, it is not REPLACEMENT.

2.6.2. Seeing Someone Metaphorically-As an Imposter

The only necessary condition that must be fulfilled for an endorsement interpretation to be felicitous is that there be an experience whose content is being endorsed in belief. As Susanna Siegel has nicely put it, 'an experience is endorsed when one forms a belief that P on the basis of an experience whose contents include P' (Siegel, 2017, p. 107). This need not entail any commitment as to what sort of mental state experience is. So, although the experience in question is most often regarded by endorsement theorists as being literally perceptual, this is not a requirement of EN per se. That means that while REPLACEMENT is not literally perceptual in character, it may still serve as a content of experience and be available to one as something to be endorsed.

One option would be to pursue the idea that there are perceptual experiences whose contents include *p* but where *p* is not literally represented. This would allow REPLACEMENT to become incorporated into one's perceptual state without itself being literally perceived, which is to say, regardless of whether imposters or the like are represented in the literal content of perception. Do such experiences exist? In what follows, I shall give the outline of an approach which could underpin an affirmative answer.

In a 2009 essay titled *The Perception of Music: Sources of Significance*, Christopher Peacocke sets himself the task of explaining how it is that we are able to experience what we might call expressive qualities of music: 'We can experience music as sad, as exuberant, as sombre. We can experience it as expressing immensity, identification with the rest of humanity, or gratitude' (Peacocke, 2009, p. 257). His main thesis is that 'when a piece of music is heard as expressing some property *F*, some feature of music is heard metaphorically-as *F'* (Peacocke, 2009, p. 257; Peacocke, 2010). Before dealing with the case of auditory perception, Peacocke refers to a painting of pottery jars by Francisco de Zurbarán, which he considers an illustration of his thesis. He argues that the painted pots are seen by many as a group of people, and that this is a matter of seeing them metaphorically-as such. For convenience, I will limit the discussion to this visual case, though the points I am about to make should be projectable mutatis mutandis across all sensory modalities.

There are three things about Peacocke's (2009) view that are worth noting. First, Peacocke thinks that the phenomenon to which we refer when we say that the pots are seen as people is genuinely experiential—rather than, say, just entertaining a thought about people while looking at the painting. For, he argues, it seems phenomenologically that you could look at four items on your desk while imagining that they are four people without having the distinctive experience that you have when you look at Zurbarán's pots.

Second, the experience is acquired through the same mechanism underlying metaphorical cognition. This is a subpersonal detection of isomorphism between two domains, each made up of concepts, properties and relations. When the isomorphism is detected, items in a source domain (say, the domain of emotions) are mapped onto items in another domain, to generate contents such as 'the sky is angry' or 'this music is sad'. In the experience of seeing pots metaphorically-as people, the item *people* is copied from a metaphorically represented domain to a 'special kind of storage', which binds it to any representation involved in perceiving the pots' properties (skinny, elongated, cone-shaped, fragile). Peacocke (2009) recognises that what is designated by the phrase 'special kind of storage' remains open to empirical investigation. But he remarks that this storage has to be special, in the sense that it must be different from the sort of storage that would make one see the pots literally as people (Peacocke, 2009, p. 267).

Third, Peacocke (2009) proposes that the concept *people* enters into the perceptual experience of the pots, only not as part of the literally represented content. What something is seen metaphorically-as, say people, does not contribute to the correctness of the

experience, namely it is partitioned off as anything entitling the subject to judge that the perceived object really is a group of people. So although there is a sense in which the depicted pots are subsumed under the concept *people*, it should not be treated as a predicative sense, for predicative subsumption is typically belief-yielding, whereas this kind of subsumption is not (Boghossian, 2010, p. 73). Rather, Peacocke says that in perceiving the painting we non-predicatively subsume the concept of those pots under the concept people, where by 'non-predicative subsumption' he just means a mapping from one concept to another under the sort of isomorphism described above (Peacocke, 2010, p. 189).

Given all this, we can read Peacocke's (2009) proposal as describing a class of perceptual experiences where an object *x* is literally represented as *Q* while being metaphorically represented as something else, and where both representations contribute their contents to the overall phenomenology of the experience. Applied to Capgras, this could introduce a new perspective on what constitutes an endorsement way of forming the delusion, and one in which the problems facing endorsement theorists may be satisfactorily settled. The central hypothesis would be that a person is in a state in which *x* looks to be *S*, but where *x* is experienced metaphorically-as an imposter. By this way of thinking, the imposter belief could still qualify as an endorsement of content, but that would be the content of a metaphorical rather than literal perceptual representation.[13]

More work would need to be done, but we could view the metaphorical content as a byproduct of misidentification. Rather than thinking that the failed tracking mechanism only generates the belief that someone is not a certain known individual, we might hypothesise that this also produces a perceptual experience with the metaphorical content 'this [perceived] person is an imposter of that [familiar] individual'. Perhaps it happens that when the subject is confronted with a person who looks exactly like *S* but is misidentified as ¬ *S*, there is a mapping from the domain imposter to the domain ¬ *S* which occurs in a manner similar to that required in metaphorical experience. On this sort of view, the subpersonal processing responsible for the failed tracking mechanism has then two conscious outcomes: a misidentification belief that the encountered person is not a certain known individual and a metaphorical experience as of that person being an imposter. The idea would be that while

[13] Of course, this raises the question of how it is that a metaphorical content is endorsed in belief. Although a full treatment of this issue is beyond the scope of this paper, a preliminary indication of how it might be addressed is provided below.

the misidentification belief is acquired through a non-experiential source, the replacement belief is acquired because the metaphorical content is treated as literal and endorsed as such.

This proposal may hold the key for dissolving all the three problems that we have used to assess endorsement proposals, namely the experiential encoding problem, the aetiology problem, and the top-down determination problem. As for the first, there are no strict limits to the range of properties that something can be perceived metaphorically-as possessing, provided that there is a functioning isomorphism between the perceptual and the metaphorical representations. As for the second, one might view the dysfunctional tracking mechanism as the medium by which the absence of autonomic response at the sight of *x* could bring about a metaphorical experience with the content that *x* is an imposter. On this picture, the mapping from imposter to ¬ *S* (which underpins such experience) would be produced in response to a situation in which some *S*-looking person is misidentified as ¬ *S* (due to defective autonomic responsiveness). As for the third problem, it is clear that one need not believe that a perceived object is really *F* in order to perceive it metaphorically-as *F*. I do not have to believe the pots are people in order to perceive them metaphorically-as such. So it seems that Capgras subjects could have a metaphorical experience with REPLACEMENT as content without that needing to be inherited from prior beliefs.

Now, even if it is true that the proposal under consideration avoids all three of the above problems, why suppose that the experiences Capgras subjects are having are metaphorical in nature? Support for this comes from considerations about the various ways in which Capgras subjects are disposed to act with respect their replacement claims. While in some cases they express hostile and even violent attitudes towards the putative imposters (for a gruesome example, see De Pauw and Szulecka, 1988; see also Christodolou, 1977), in other cases they are friendly or indifferent towards them (as with patient Fred and his wife Wilma, see Lucchelli and Spinnler, 2007). The metaphorical framework can make sense of both kinds of subject by offering an explanation of why some fail to act in a way consistent with their assertions. This would be explained by the fact that metaphorical contents of experience are typically make-believed, not believed. That is, it may be that a significant reason why some subjects do not feel hate or version towards the 'imposters' is that they do not actually endorse the metaphorical content as true; rather, perhaps, they use it as a springboard to create stories about the identity of the misidentified people, which result in claims about these being as the metaphorical experience represents them to be.

But then why would some other subject really believe that their loved ones are as they are metaphorically represented to be? Given that typically metaphorical experience does not lead to belief, why would this happen in Capgras? Fully answering this question is a difficult task which cannot be undertaken here, but some speculations can be made. Perhaps one becomes prone to believe that things are as metaphorically represented when one cannot believe anything that counts as good evidence against it. Imagine somebody who perceives *x* as looking like her mother, but who cannot help believing that *x* is not her mother, such that she dismisses any counterevidence to what she believes. Since *x* really is her mother, all of the available evidence is consistent with the way things look and against her belief. In these circumstances, it is arguable that one might be at least somewhat inclined to endorse the metaphorical representation of *x* as an imposter, because this (contrary to the weight of evidence) is consistent not only with *x*'s looking like her mother, but also with her belief that *x* is not really her mother. If so, it may be that the difference between those who go on to endorse the content and those who do not depends on local psychological dispositions that are not uniform across all Capgras subjects, such as perhaps paranoid personality traits like distrust of others and unjustified suspicions of being deceived.

Of course, one might point out that an endorsement account along these lines is far more modest than it is standardly taken to be. Indeed, the misidentification judgement is not formed from endorsement processes, which means that the endorsement hypothesis cannot explain the delusion in its initial formation. However, my proposal locates the explanatory power of the endorsement hypothesis elsewhere. That is, although appeal to endorsement does not explain why Capgras subjects adopt the initial misidentification belief it does explain the trade-off between those who act in a way consistent with their replacement claims, and those who do not. The Capgras delusion is equally importantly characterised by the misidentification of familiar people and their replacement by doubles. Both misidentification and replacement claims acquire prominence in the Capgras subject's mental life and are reported sincerely and with conviction. Where I differ from other writers on the subject is in the idea that the explanatory power of the endorsement hypothesis is relative to the latter feature rather than the former.

2.7. Conclusion

Only two attempts, by Pacherie and Wilkinson, have been made to systematically articulate an endorsement interpretation of the Capgras delusion. The first part of this essay considered

two such attempts and argued that neither is a viable option. Pacherie's strategy does not escape the problems it was intended to, whereas Wilkinson's fails to provide a principled basis for determining what the endorsed content is to be. The second part of the essay introduced a new way of thinking about the experiential endorsement in Capgras, according to which the content endorsed in the delusional belief is metaphorical rather than literal. The claim that metaphorical experience is what is occurring in actual cases of Capgras is very speculative and certainly liable to empirical refutation. Peacocke's model itself requires considerable development before we can decide whether it is empirically adequate. Also, more work would be needed to show how exactly one could form a belief that familiar people are imposters simply on the basis of perceiving them metaphorically-as such. Despite these limitations, postulating a role for metaphorical experience in Capgras is not incompatible with any of the facts we know to be true about this condition, and has the benefit of allowing a novel way in which an endorsement interpretation of Capgras might plausibly proceed.

2.8. References

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

Bayne, T. and Pacherie, E. (2004). 'Bottom-Up or Top-Down? Campbell's Rationalist Account of Monothematic Delusions'. *Philosophy, Psychiatry, and Psychology*, 11, pp. 1–11.

Boghossian, P. (2010). 'The Perception of Music: Comments on Peacocke'. *British Journal of Aesthetics*, 50(1), pp. 71–76.

Bongiorno, F. (2020). 'Is the Capgras Delusion an Endorsement of Experience?' *Mind and Language*, 35(3), pp. 293–312.

Bortolotti, L. (2012). 'In Defence of Modest Doxasticism About Delusions'. *Neuroethics*, 5(1), pp. 39–53.

Bortolotti, L. (2010). *Delusions and Other Irrational Beliefs*. Oxford: Oxford University Press.

Brewer, B. (1999). *Perception and Reason*. Oxford: Oxford University Press.

Brighetti, G., Bonifacci, P., Borlimi, R. and Ottaviani, C. (2007). "Far from the Heart Far from the Eye': Evidence from the Capgras delusion'. *Cognitive Neuropsychiatry*, 12, pp. 189–197.

Campbell, J. (2001). 'Rationality, Meaning, and the Analysis of Delusion'. *Philosophy, Psychiatry, and Psychology*, 8(2/3), pp. 89–100.

Carruthers, P. (2015a). *The Centered Mind: What the Science of Working Memory Shows us About the Nature of Human Thought*. Oxford: Oxford University Press.

Carruthers, P. (2015b). 'Perceiving Mental States'. *Consciousness and Cognition*, 36, pp. 498–507.

Christodolou, G. N. (1977). 'The Syndrome of Capgras'. *British Journal of Psychiatry*, 130, pp. 556–564.

Coltheart, M. (2013). 'On the Distinction between Monothematic and Polythematic Delusions'. *Mind and Language*, 28 (1), pp. 2013–2112.

Corlett, P. R. (2019). 'Factor One, Familiarity and Frontal Cortex: A Challenge to the Two-Factor Theory of Delusions'. *Cognitive Neuropsychiatry*, 24, pp. 165–177.

Davies, M., Coltheart, M., Langdon, R. and Breen, N. (2001). 'Monothematic Delusions: Towards a Two-Factor Account'. *Philosophy, Psychiatry, and Psychology*, 8(2/3), pp. 133–158.

Davies, M. and Egan, A. (2013). 'Delusion: Cognitive Approaches—Bayesian Inference and Compartmentalization'. In K. W. M. Fulford, M. Davies, R. G. T. Gipps, G. Graham, J. Z.

Sadler, G. Stanghellini and T. Thornton (eds.) *The Oxford handbook of philosophy and psychiatry* (pp. 688–727). Oxford: Oxford University Press.

De Pauw, K. W. and Szulecka, T. K. (1988). 'Dangerous Delusions: Violence and the Misidentification Syndromes'. *British Journal of Psychiatry*, 152, pp. 91–96.

Eilan, N. (2001). 'On Understanding Schizophrenia'. In D. Zahavi (ed.) *Exploring the Self* (pp. 97–113). Amsterdam: John Benjamins.

Ellis, H. D. and Young, A. W. (1990). 'Accounting for Delusional Misidentifications'. *British Journal of Psychiatry*, 157, pp. 239–248.

Ellis, H.D., Young, A.W., Quayle, A., and de Pauw, K. W. (1997). 'Reduced Autonomic Responses to Faces in Capgras Delusion'. *Proceedings of the Royal Society: Biological Sciences*, B264, pp. 1085–1092.

Ellis, H. D., Lewis, M. B., Moselhy, H. F. and Young, A. W. (2000). 'Automatic without Autonomic Responses to Familiar Faces: Differential Components of Covert Face Recognition in a Case of Capgras Delusion'. *Cognitive Neuropsychiatry*, 5(4), pp. 255–269.

Fine, C., Craigie, J. and Gold, I. (2005). 'Damned if you Do, Damned if you Don't: the Impasse in Cognitive Accounts of the Capgras delusion'. *Philosophy, Psychiatry, and Psychology*, 12, pp. 143–151.

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

Fodor, J. A. (1985). 'Precis of the Modularity of Mind'. *Behavioural and Brain Sciences*, 8, pp. 1–42.

Frazer, S. J. and Roberts, J. M. (1994). 'Three Cases of Capgras' Syndrome'. *British Journal of Psychiatry*, 164, pp. 557–559.

Gold, J. and Gold, I. (2014). *Suspicious Minds: How Culture Shapes Madness*. New York: Free Press.

Hawley, K. and MacPherson, F. (eds.) (2011). *The Admissible Contents of Experience*. Malden: Wiley-Blackwell.

Haxby, J. V., Hoffman, E. A. and Gobbini, M. I. (2000). 'The Distributed Human Neural System for Face Perception'. *Trends in Cognitive Sciences*, 4(6), pp. 223–233.

Hirstein, W. (2005). *Brain Fiction*. Cambridge, MA: MIT Press.

Hirstein, W. (2010). 'The Misidentification Syndromes Mindreading Disorders'. *Cognitive Neuropsychiatry*, 15(1), pp. 233–260.

Hirstein, W. and Ramachandran, V. S. (1997). 'Capgras Syndrome: A Novel Probe for Understanding the Neural Representation of the Identity and Familiarity of Persons'. *Proceedings of the Royal Society of London B: Biological Sciences*, 264, pp. 437–444.

Kahneman, D., Treisman, A. and Gibbs, B. J. (1992). 'The Reviewing of Object Files: Object-Specific Integration of Information'. *Cognitive Psychology*, 24, pp. 174–219.

Langdon, R. and Bayne, T. (2010). 'Delusion and Confabulation: Mistakes of Perceiving, Remembering and Believing'. *Cognitive Neuropsychiatry*, 15(1/2/3), pp. 319–345.

Langdon, R. and Connaughton, R. (2013). 'The Neuropsychology of Belief Formation'. In F. Kreuger and J. Grafman (eds.) *The Neural Basis of Human Belief Systems* (pp. 19–42). New York: Taylor and Francis.

Lucchelli, F. and Spinnler, H. (2007). 'The Case of Lost Wilma: A Clinical Report of Capgras Delusion'. *Neurological Science*, 28(4), pp. 188–195.

Maher, B. (2005). 'Delusional Thinking and Cognitive Disorder'. *Integrative Physiological and Behavioural Science*, 40(3), pp. 136–146.

Marr, D. (1982). *Vision. A Computational Investigation into Human Representation and Processing of Visual Information*. San Francisco, CA: Freeman.

McDowell, J. (1994). *Mind and World*. Cambridge, MA: Harvard University Press.

Pacherie, E. (2009). 'Perception, motions and Delusions: Revisiting the Capgras Delusion'. In T. Bayne and J. Fernandez (eds.) *Delusions and Self-Deception* (pp. 107–126). Hove: Psychology Press.

Pacherie, E., Green, M. and Bayne, T. (2006). 'Phenomenology and Delusions: Who Put the 'Alien' in alien control?'. *Consciousness and Cognition*, 15(3), pp. 566–577.

Parrott, M. (2019). 'Delusional Predictions and Explanations'. *The British Journal for the Philosophy of Science*, , 0, pp. 1–32.

Peacocke, C. (2009). 'The Perception of Music: Sources of Significance'. *The British Journal of Aesthetics*, 49(3), pp. 293–297.

Peacocke, C. (2010). 'Music and Experiencing Metaphorically-As: Further Delineation'. *The British Journal of Aesthetics*, 50(2), pp. 189–191.

Perry, J. (1980). 'A Problem about Continued Belief'. *Pacific Philosophical Quarterly*, 61, pp. 317–332.

Powers, A. R., Mathys, C. and Corlett, P. R. (2017). 'Pavlovian Conditioning-Induced Hallucinations Result from Overweighting of Perceptual Priors'. *Science*, 357(6351), pp. 596–600.

Ramachandran, V. S. and Blakeslee, S. (1998). *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. New York: William Morrow.

Recanati, F. (1993). *Direct Reference: From Language to Thought*. Oxford: Blackwell.

Recanati, F. (2012). *Mental Files*. Oxford: Oxford University Press.

Siegel, S. (2017). *The Rationality of Perception*. New York: Oxford University Press.

Stone, T. and Young, A. W. (1997). 'Delusions and Brain Injury: The Philosophy and Psychology of Belief'. *Mind and Language*, 12, pp. 327–364.

Wilkinson, S. (2015). 'Delusions, Dreams, and the Nature of Identification'. *Philosophical Psychology*, 28(2), pp. 203–226.

Wilkinson, S. (2016). 'A mental files approach to delusional misidentification'. *Review of Philosophy and Psychology*, 7, pp. 389–404

Wyatte, D., Jilk, D. and O'Reilly, R. (2014). 'Early Recurrent Feedback Facilitates Visual Object Recognition under Challenging Conditions'. *Frontiers in Psychology*, 5(674), pp. 16–25.

Chapter 3: The Role of Unconscious Inference in Models of Delusion Formation[14]

3.0. Abstract

Delusional inference has been increasingly understood in terms of Bayesian updating. In this paper I critically evaluate an influential Bayesian model of delusional inference put forward by Coltheart, Menzies, and Sutton (2010), which I will call, for simplicity, 'the Coltheart model'. Specifically, I consider how well this model fits with two accounts of delusion: the 'explanationist account' (or simply 'explanationism'), according to which the delusional belief is grounded in an explanatory inference from experience, and the 'endorsement account', according to which the delusional belief arises non-inferentially when the experience is taken as veridical. Because it understands delusion formation as explanatory in nature, the Coltheart model has been regarded as a version of explanationism. But, as I show, it is no more or less compatible with the endorsement account.

3.1. Introduction

Delusions are a commonly observed symptom in people with a wide range of psychiatric disorders, including schizophrenia, dementia, schizoaffective disorder, bipolar disorder, and major depression. In the latest version of the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (DSM-V), delusion is defined as, 'a false belief based on incorrect inference about external reality that is firmly held despite what almost everyone else believes and despite what constitutes incontrovertible an obvious proof of evidence of the contrary' (APA, 2013, p. 819).

As delusions are understood in DSM-V, they are based on inference. In this paper, I focus on the role that inference plays in delusion formation. Brendan Maher was the first to suggest that the formation of delusions involves an inferential transition—although he denies that the inference from which delusions arise is faulty (Maher, 1992; Maher, 1999). Maher defends a view known as 'explanationism' (Maher, 1974; Stone and Young, 1997), according to which delusions are hypotheses adopted to explain anomalous perceptual experiences and arrived at by inferential reasoning that is neither abnormally biased nor otherwise deficient (Maher, 1974, p. 180). In essence, delusions for Maher are the product of normal reasoning

[14] This chapter is largely identical to one (co-authored with Lisa Bortolotti) which appears in a book edited by Anders Nes and Timothy Chan, 2020, *Inference and Consciousness*, published by Routledge.

processes brought to bear on some experiential aberration. This means that the pathological nature of the delusion does not lie in the person's inferential reasoning, but only in the experience that generates it. It also means that no additional abnormality is needed to explain delusional belief formation and maintenance, which is why Maher's view has become known as a one-factor theory.

An alternative to explanationism is the endorsement theory, according to which the delusional belief is an acknowledgement that the anomalous experience is veridical and no inference from experience to belief is required (Pacherie et al., 2006; Bayne and Pacherie, 2004a).

Over a number of years, Max Coltheart and colleagues (e.g., Coltheart, 2005; Coltheart, 2007; Davies and Coltheart, 2000; Coltheart et al., 2007, 2011) advocated a two-factor theory of delusions. The theory is primarily in the business of explaining monothematic delusions (i.e., delusions whose content is restricted to a single theme) of neuropsychological origin. The guiding idea is that there are two contributing clinical factors to delusion formation. The first is an impairment that effects the production of abnormal *data*, the explanation of which is supposed to furnish the content of the delusion.[15] The second is an impairment in the mechanism responsible for belief evaluation, which is supposed to explain why the person with the delusion is willing to maintain implausible explanations for the abnormal data.

Coltheart and colleagues postulate a second factor because they deny that the presence of abnormal data is the only abnormality involved in the formation and maintenance of delusions. It is natural to think that the notion of a second factor (construed as an impairment in belief evaluation) is inconsistent with Maher's contention that delusions are formed and maintained as the result of a single experiential anomaly. Indeed, it is precisely for this reason that the two-factor theory has been interpreted in the literature as an alternative to Maher's brand of explanationism. However, Coltheart, Menzies, and Sutton (2010) have offered a Bayesian account of abductive inference (henceforth 'the Coltheart model') that vindicates at least Maher's basic contention that the inference leading to the adoption of the delusional belief is not faulty.

[15] As we shall see in Section 3.2., using the term 'data' instead of 'experience' is meant to avoid commitment to the idea that the abnormalities which prompt delusional hypotheses are always consciously accessible.

The Coltheart model has been developed with specific reference to the Capgras delusion, the belief that a person or persons dear to the deluded individual have been replaced by identical or nearly identical imposters (Capgras and Reboul-Lachaux, 1923). Specifically, the proposal is that in Capgras unconscious abductive reasoning is used to infer from the abnormal data to the delusional hypothesis—which is the hypothesis that best explains the data. In the Coltheart model the inference involved in the formation of the delusion is Bayesian rational and does not involve a reasoning impairment, though a reasoning impairment is postulated to explain the maintenance of the delusional belief in the face of counterevidence.

In the last decade, several theorists have pointed to the use of Bayesian framework for modelling delusional inference, with the debate revolving around the number of factors necessary for delusion formation, and the similarities and differences between the available models (e.g., Coltheart et al., 2010; McKay, 2012; Bortolotti and Miyazono, 2015; Miyazono et al., 2015; Miyazono and McKay, 2019). In this paper, I focus specifically on the question of compatibility between, on the one hand, the Coltheart model, and on the other hand, explanationist and endorsement accounts of delusion formation. To the extent that it conceptualises delusions as hypotheses explaining abnormal data, the Coltheart model has been interpreted as a version of explanationism (Parrott, 2019; Young, 2014). However, I argue that the centrepiece of the Coltheart model—the notion that the delusion is arrived at through unconscious Bayesian-style abductive inference—can be equally well captured in either explanationist or endorsement terms. If that is correct, the presence of such an inference in the process of delusion formation is not sufficient to discriminate between explanationist and endorsement accounts.

Here is the plan. In Section 3.2. I pay special attention to the claim that the formation of the Capgras delusion occurs as a consequence of an unconscious Bayesian-style abductive inference and that the data the delusional hypothesis is invoked to explain are not consciously experienced. In Section 3.3. I elaborate on the differences between explanationism and the endorsement theory and discuss their advantages and disadvantages. In Section 3.4. I describe two ways of incorporating Bayesian-style abductive inference into the explanationist framework. In Section 3.5. I show that Bayesian-style abductive inference can also be incorporated into the endorsement framework. In Section 3.6. I argue that explanationism is no more compatible with the Coltheart model than is the endorsement account. In Section 3.7. I present some concluding remarks.

3.2. The Coltheart Model

3.2.1. Preliminaries

One way to illustrate the Maher approach to delusion is by reference to a model of Capgras advanced by Hayden Ellis and Andrew Young in the nineties (Ellis and Young, 1990). It is widely agreed that familiar face recognition is correlated with an increased activity in the autonomic nervous system that can be measured by changes in skin conductance. The proposal is that in the Capgras delusion a neuropsychological anomaly disrupts the connection between a person's face recognition system and the autonomic nervous system, such that the person with Capgras fails to show differential autonomic responses to familiar compared to unfamiliar faces.

Findings of reduced skin conductance in people with Capgras have confirmed this proposal, making it likely that the formation of the delusions at least partly explained by a lack of responsiveness in the autonomic nervous system (Hirstein and Ramachandran, 1997; Ellis et al., 1997; Ellis et al., 2000; Brighetti et al., 2007). Conceivably, such a lack of autonomic responsiveness in the presence of a visually familiar face could give rise to anomalous experience, such as an experience of a certain person being unfamiliar or different in some way.

In Maher's view, the hypothesis that one's wife has been replaced by an imposter is a *normal* response to the anomalous experience that is devised to explain (Maher, 1988, 1999, 2005).[16] That is, one normal inference to make from the experience of a person who looks like one's mother but feels unfamiliar is that she is not really one's mother, but an imposter. As we will see in more detail in the sections that follow, not only does the Coltheart model vindicate Maher's central idea that the delusional hypothesis arises via normal inferential processes. It goes even further, to argue that the inference to the delusional hypothesis is a *rational* inference to make given the circumstances. In two important respects, however, the Coltheart model differs from Maher's view.

The first difference concerns the role of conscious, person-level processes in the formation of the delusion. Unlike Maher, Coltheart and colleagues (2010) maintain that the 'abnormal data' (as they call them) which initially prompt the imposter idea are not mental events of which a person is conscious (i.e., experiences). Warrant for this claim is grounded in the idea that since we are not conscious of the processes going on in the autonomic

[16] Maher uses the term 'normal' to mean that any irrationality implicated in the inference leading to the delusional hypothesis is not clinically significant.

nervous system, a person would not be conscious of a lack of responsiveness in this system. Note, however, that Coltheart and colleagues (2010) do not deny that people with Capgras have abnormal conscious experiences; they simply claim that these experiences follow on the adoption of the delusion, rather than causing the delusion. In the Coltheart model, everything along the chain of cognitive processes that precedes delusion takes place without conscious awareness. The first content that enters consciousness— 'the first delusion-relevant event of which the patient is aware' (Coltheart et al., 2010, p. 264) —is the delusional hypothesis, 'this person looks like my loved one but is an imposter'.

Second, Maher thinks that discrepant familiarity data (i.e., discrepancies between visual and autonomic recognition of familiar faces) are the only abnormal factor necessary for the delusion to be formed and maintained, whereas Coltheart and colleagues (2010) think that an additional abnormality is needed. Briefly, their reasoning runs as follows. Suppose we grant that a failure of autonomic response to a loved one's face accounts for why the imposter hypothesis is initially adopted as a belief. We would expect an otherwise healthy person to subsequently reject the belief in the face of new data that disconfirm it (such as the testimony of a doctor or the assurance of family and friends). People with Capgras, however, continue to believe that their loved ones have been replaced by imposters. Why is this so? According to Coltheart and colleagues (2010), there has to be an impairment in belief evaluation resulting from right hemisphere damage that prevents the person from rejecting the newly formed belief despite evidence against it. Possible support for the additional factor is provided by the comparison with patients with ventromedial frontal damage (VMF). VMF patients also fail to show autonomic discrimination between familiar and unknown faces, but do not develop delusional beliefs about replacement look-alikes and imposters (Tranel et al., 1995). This has been taken to be evidence that Capgras subjects, unlike VMF patients, must suffer a second impairment.

For current purposes, what matters is the role that the Coltheart model assigns to inferential transitions in the aetiology of Capgras delusion. Coltheart and colleagues (2010) argue that the delusion initially arises from normal unconscious inferential responses to an abnormal input that is due to the absence of autonomic response to familiar faces. In the following subsections I focus on two issues: first, what the nature of the inference is that leads from this absence of autonomic activity to the initial adoption of the delusional belief; and second, to what extent the inference can be said to be rational.

3.2.2. Abductive Inference

There are two basic assumptions in the Coltheart model: that the Capgras delusion is best construed as the conclusion of an abductive inference (also called inference to the best explanation, see Harman, 1965; Lipton, 2004); and that the model that is best suited to explain abductive inference in the context of delusion formation is the Bayesian account.[17] Motivation for the former assumption comes from the fact that the processes leading to the initial onset of the Capgras delusion cannot be successfully described in terms of deductive or inductive inference. The imposter belief cannot be the outcome of a deductive chain of inference, for that would imply that the replacement of a close relative *x* by an imposter is logically entailed by the lack of autonomic response to *x*'s face, which is not. Nor is it plausible to suppose that the imposter belief is obtained by inductive generalization, through the observation that events of type *B* invariably follow upon events of type *A*. Unlike inductive generalization, the inference that *x* is an imposter brings about concepts that go beyond the observable data available to one when one encounters *x*. More plausible is the suggestion that people with Capgras engage in abductive reasoning or inference to the best explanation (henceforth: IBE). Imagine that you are examining a vast array of potential alternative hypothesis for a single body of data. IBE is the process of selecting the hypothesis which, if true, would best explain that data. So, to say that delusional inference is like IBE is to say that it draws on explanatory considerations for an assessment of how likely competing hypotheses are to be true.

3.2.3. Bayesian Abductive Inference

As Coltheart and colleagues note, 'the crucial mark of the model is that it marries a natural probabilistic account of explanation to the standard Bayesian model of rational belief systems' (2010, p. 271). The probabilistic account of explanation provides a measure of the relative explanatory power of the alternative hypotheses H_1, $H_2 \ldots H_n$ in light of actual data. The degree to which a hypothesis *H* explains observations *O* is a function of the probability of *O* given H. It follows that one hypothesis H_1 is a better explanation of *O* than rival hypotheses $H_2 \ldots H_n$ just in case the probability of *O* is higher under H_1 than under $H_2 \ldots$

[17] Davies and Egan (2013) have raised doubts whether the notion of Bayesian inference can be adequately captured in terms of abductive inference (or inference to the best explanation). They argue that all the relevant theoretical considerations made by the Coltheart model could be achieved equally well by replacing its talk of Bayesian abductive inference by the talk of Bayesian inference per se (p. 696). Since my only concern here is to outline the Coltheart model, I do not take stand on this matter.

H_n. For example, in the case of Capgras delusion, the imposter hypothesis explains better the discrepant familiarity data than rival hypotheses just so long as the probability of observing the data is higher under the imposter hypothesis than under the rival hypotheses.

How does the probabilistic account of explanation get combined with the standard Bayesian model of rational belief systems? The Bayesian model describes the rational way to update beliefs as new evidence is gathered (what is called 'Bayesian inference'). An agent's belief system is represented as consisting of subjective probabilities, measures of the degrees of belief that the agent assigns to various hypotheses. A probability function accords to each hypothesis *H* a mean value from the interval [0,1] *p*(*H*), which quantifies the agent's degree of belief in what *H* reports. The agent's subjective probabilities are updated over time through a procedure known as 'conditionalisation'.

Suppose you want to know the probability that some hypothesis, *H,* is true (e.g., 'it is about to rain') given some new evidence, *O*, (e.g., 'the sky is overspread with clouds') you have observed. Conditionalisation requires that you change your degree of belief in *H* so that it is equal to your prior degree of belief in *H* conditional on *O*:

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

The expression can be unpacked by Bayes' theorem in the following way:

$$P(H|O) = P(H) \cdot \frac{P(O|H)}{P(O)}$$

One can read the equation as: 'The conditional probability *P(H|O)* (the posterior probability of *H* given *O*) is proportional to the product of *P(H)* (the prior probability of *H*), *P(O|H)* (the likelihood of *O* given *H*), and *P(O)* (the prior probability of *O*)'.

In this equation, the posterior probability *P(H|O)* is what we are after, the probability that *H* is true given the evidence *O* (e.g., the probability of rain given clouds). The likelihood *P(O|H)* is the probability of attaining the evidence *O* if *H* is true (e.g., the probability of there being clouds given that it is raining), or, as per the probabilistic account of explanation, how well *H* explains the evidence *O* (i.e., the higher the likelihood, the greater *H*'s explanatory power). The prior probability *P(H)* represents the probability attributed to *H* before the evidence *O* is in (e.g., the probability of rain in general). A hypothesis may well have a high likelihood, but assuming that the prior probability is extremely low, the posterior probability will also be low. The hypothesis that aliens have produced clouds around their spacecraft to

hide themselves (H_a) may explain the fact that there are clouds overhead, and therefore have a high likelihood. But since the prior probability is very low, it would be wrong to assign a high posterior probability to H_a (i.e., conclude that H_a is probably true) given that there are clouds overhead. The remaining term *P(O)* refers to the prior probability of *O* before the relevant observation begins (e.g., the probability of there being clouds in general).

Coltheart and colleagues (2010) dismiss *P(O)* as unimportant, for they are less interested in whether evidence supports a hypothesis absolutely than in which of two or more candidate hypotheses the evidence best supports (p. 273). When beliefs about the world can be reduced to a set of two competing hypotheses (H_1, H_2), Bayes' theorem can be used as a normative criterion for selecting the one with the highest posterior probability:

$$\frac{P(H1/O)}{P(H2/O)} = \frac{P(H1) \cdot P(O/H1)}{P(H2) \cdot P(O|H2)}$$

Put into words, the two-hypothesis formulation of Bayes' theorem states that the ratio of the posterior probabilities of H_1 and H_2 (the posterior odds) is equal to the ratio of their prior probabilities (the prior odds, the second term on the right of the previous equation), multiplied by the ratio of the respective likelihoods of *O* given H_1 and H_2 (the likelihood ratio, the first term on the right of the equation previous equation). The term *P(O)* no longer figures in the equation because it was cancelled out during multiplication.

We are now in a position to see how the Bayesian model can accommodate an account of IBE. One such account must answer the question of what makes it rational to accept a hypothesis, or favour one hypothesis over others, in light of certain observations, the evidence. The Bayesian model (via Bayes' theorem) tells us that it is rational to select the hypothesis which best fits the evidence but weighted by its plausibility independent of (or apart from) the evidence (Hohwy, 2013). That is, the hypothesis with a higher posterior than rival alternatives, or, to put it another way, the hypothesis that is better supported by the evidence than rival alternatives are. The central claim is that an abductive inference from the available evidence *O* to the hypothesis *H* is rational or justified only insofar as *O* supports *H* more strongly than it does any other hypotheses *H'* (Coltheart et al., 2010, p. 274). I will now consider the implications of this for understanding the Capgras delusion.

3.2.4. Bayesian Inference and Delusion Formation

As we said at the onset, the Capgras delusion is formed in the presence of discrepant familiarity data resulting from a disconnection between the facial recognition system and the autonomic nervous system. Imagine a man whose face recognition system has become disconnected from the nervous system as the result of a brain injury when he hit his head during a car accident. Let *O* stand for the incongruous data presented to him upon encountering his wife for the first time in the hospital after the accident:

O: lack of autonomic familiarity data to the presence of visual familiarity data.

Suppose our man is considering a set of two mutually exclusive and exhaustive hypotheses to account for *O*: the *wife* hypothesis, denoted by H_w, and the *imposter* hypothesis, denoted by H_i:

H_w: This woman who visibly resembles my wife and claims to be my wife really is my wife.

H_i: This woman who visibly resembles my wife and claims to be my wife is an imposter.[18]

After observing *O*, which hypothesis fares better? From a Bayesian perspective, this question is equivalent to asking whose posterior probability is greater. Bayes' rule dictates how we should update some prior probability distribution over a set of hypotheses to provide the posterior probabilities of the hypotheses given the data. Specifically, the formula tells us that the posterior probability of each of our two hypotheses is obtained by multiplying together the likelihood (the degree of probability of seeing *O* if H_w or H_i were true) and the prior probability (the degree of probability assigned to H_w or H_i before observing *O*).

Accordingly, the first step towards an answer to the question of which hypothesis is more plausible is to examine whether *O* is more likely under the wife hypothesis H_w or under the imposter hypothesis H_i (i.e., whether *O* is better explained by one or the other). In the model, *O* is 'much more likely' under H_i than H_w, since it would be highly unlikely for our

[18] Coltheart and colleagues (2010) use the term 'stranger' instead of 'imposter' to refer to the person who gets misidentified. However, what they mean to designate by 'stranger' is in fact an imposter, namely a person who both looks like, and poses as, another. To avoid confusion, I prefer to speak of 'imposter' and 'imposter hypothesis'.

man to observe discrepant familiarity data if the person was in fact his wife (Coltheart et al., 2010, p. 277). So, it is argued that the likelihood ratio (1)

$$\frac{P(O|Hi)}{P(O|Hw)}$$

will be very high. However, as we saw above, in assessing the posterior odds for any pair of hypotheses, we need to ask not only how well they explain the evidence, but also how probable they were (believed to be) in view of the knowledge available prior to taking the evidence into account. Most of us would say that the prior probability of the imposter hypothesis is very low compared with that of the wife hypothesis, since the former is generally plausible, whereas the latter is implausible. If this is right, the prior odds of the imposter hypothesis (2) will be very low:

$$\frac{P(Hi)}{P(Hw)}$$

Nevertheless, a fundamental component in the Bayesian comparison of hypotheses is that the ratio of posterior probabilities (3)

$$\frac{P(Hi|O)}{P(Hw|O)}$$

can still favour H_i over H_w even when the prior odds of H_i (2) are relatively low, provided that the likelihood ratio (1), i.e., the relative explanatory power of H_i, is sufficient to overwhelm the prior odds against it. According to the model (Coltheart et al., 2010), this is what happens in Capgras. The imposter hypothesis accounts much better for the discrepant familiarity data than the wife hypothesis, and this offsets its relative low prior probability in calculating the ratio of posterior probabilities.

So, a Bayesian-rational agent (i.e., one who updates their beliefs in accordance with Bayes' rule) will infer the belief that the woman who came to visit him in the hospital is an imposter posing as his wife. To this extent, the Coltheart model is consistent with Maher's approach. The adoption of the imposter belief in the wake of discrepant familiarity data is not only a normal, but a 'perfectly rational' response (Coltheart et al., 2010, p. 281; cf. McKay, 2012).

3.2.5. Challenges to the Coltheart Model

Ryan McKay (2012) argues that the Coltheart model presupposes an unrealistic estimation of prior probabilities. In the model, prior probabilities are expressed as $P(H_w) = 0.99$ and $P(H_i) = 0.01$. According to McKay (2012), the hypothesis of one's wife being replaced by a physically identical imposter 'represents an exceedingly unlikely—almost miraculous—state of affairs' (McKay, 2012, p. 340). For this reason, he thinks the value chosen in the Coltheart model for the prior probability of the imposter hypothesis, $P(H_i) = 0{,}01$, is far too high. A more realistic distribution of prior probabilities would so strongly favour the wife hypothesis that it could not be outweighed by the likelihood ratio. McKay suggests assigning a prior probability of 0.00027 to $P(H_i)$ and a prior probability of 0.99973 to $P(H_w)$. If we accept McKay's proposed prior probabilities, the posterior probability for the imposter hypothesis H_i would be much lower than (only 0.27 of that for) H_w, such that it would irrational to choose H_i over H_w.

The question then arises as to why the person with Capgras adopts the imposter hypothesis as a belief. Why is not the belief rejected? McKay (2012) proposes that this is because belief updating is heavily biased towards explanatory adequacy at the expense of conservatism (cf. Stone and Young, 1997). Unbiased updating would strike a fair balance between two complementary demands: forming beliefs that explain the data well (explanatory adequacy) and forming beliefs that require as little as possible readjustment to the subject's prior beliefs (conservatism). On McKay's view, people with Capgras discount the prior implausibility of the imposter scenario on account of its relative explanatory power, and that is why H_i gets a higher posterior than H_w. In this case, then, the second factor is a bias of deference to the abnormal data, and its role is to explain not only the maintenance but also the adoption of the delusional belief (McKay, 2012).

Another challenge to the Coltheart model is that it denies any role to anomalous experience in the aetiology of the delusion. As we have already seen, in the Coltheart model the process from the abnormal data to the onset of the delusion is entirely unconscious. In a recent paper, Garry Young challenges this, suggesting that the imposter hypothesis (i.e., 'that must be an imposter masquerading as my wife') co-occurs with an anomalous experience (i.e., a salient sense of unfamiliarity) before being accepted as a belief (Young, 2014; Young, 2008). Young agrees that the imposter hypothesis emerges as a product of an entirely unconscious inference process. However, he thinks that people with Capgras come to believe this hypothesis because it makes probable an experience that is otherwise

unexpected: 'the experiential state gives credence to the freshly emerged imposter thought such that the corresponding (delusional) belief is formed to explain the experience' (Young, 2014, p. 94).

On a similar note, Davies and Egan note that even if we are not consciously aware of the activities in the autonomic nervous system (e.g., the failure of autonomic response to a spouse's face), it does not follow that there is no conscious experience prior to the formation of the delusional belief (Davies and Egan, 2013). This is because it remains possible that the lack of autonomic response to the sight a spouse's face should generate some kind of anomalous experience, which would in turn serve as a basis for the delusional belief. Specifically, Davies and Egan (2013) claim that the delusional belief is adopted as a result of computations being carried out by a perceptual module, where these are understood as processes of Bayesian inference. On their account, the winning hypothesis (i.e., the conclusion of Bayesian inference) gives the experience the content it has (e.g., 'that's not actually my wife'), and the initial adoption of the delusional belief is a prepotent doxastic response to that content (Davies and Egan, 2013, p. 722). Interestingly, the module's probabilities may be incongruous with the actual state of the world (i.e., unrealistic) because modules can draw only on a restricted class of inputs (i.e., are domain-specific). In addition, the module's probabilities may be at odds with what the person knows or believes because modules are informationally encapsulated (i.e., they have little or no access to information at higher levels of processing); and, to this extent, they may be naturally biased against pre-existing beliefs and in favour of likelihoods (Davies and Egan, 2013, p. 714; Fodor, 1983).

3.3. Explanationist versus Endorsement Accounts

In this section I describe two distinct varieties of 'empiricist' (i.e., experience-based) approaches to delusions: the explanationist and the endorsement accounts.[19] If experience is to figure in the causal chain leading to delusion formation, then the representational content of the experience may be more or less close to the content of the delusion itself. Proponents of the explanationist account (henceforth *ES*) claim that the content of the Capgras experience is sparser than the content of the delusion (e.g., 'this woman feels unfamiliar'), and that the delusion arises as a means to explain the anomalous experience (Ellis and Young,

[19] The distinction between 'explanationist' versus 'endorsement' options has been so phrased by Bayne and Pacherie (2004a). Others have drawn the same distinction using different terminology. See Davies and Coltheart (2000), Fine et al., (2005), and Turner and Coltheart (2010).

1990; Maher, 2005; Coltheart, 2005). On the other hand, those who advocate an endorsement account (henceforth *EN*) hold that (much of) the content of the delusion is already encoded in the content of experience, and that the delusion simply reports the content of that experience—'seeing is believing', as is sometimes said (Bayne and Pacherie, 2004a; Pacherie et al., 2006; Pacherie, 2009).

Endorsement theorists typically take the anomalous experience to be a perceptual state with propositional content, where the content amounts to misidentification. EN breaks down into three components: (1) the person has an experience with the misidentification content 'this woman is not my wife'; (2) the experience is endorsed as veridical, such that the person believes that this woman, who looks like his wife, is not really her; (3) this belief is later developed into the belief that the person's wife has been replaced by an imposter. The content available to be endorsed is the misidentification content, while the imposter belief is just an explanatory hypothesis for the fact that, as the person believes, the woman looks like his wife but is not really her (Davies and Davies, 2009, p. 302; Langdon and Bayne, 2010). Another possibility is that the person's experience already includes the imposter idea as part of its content (see Pacherie, 2009 and Chapter 2 for explorations of this possibility).

ES and EN have complementary strengths and weaknesses. One strength of EN is that it offers a plausible framework for understanding why delusions are held with such high conviction that they outweigh the evidence of testimony. If the conscious perception of the person with Capgras, on looking at his wife, is of seeing someone as not being his wife, testimony may not suffice to persuade him that the woman he sees is in fact his wife (Langdon and Connaughton, 2013, p. 29). On this account, the delusional conviction flows directly from perceptual experience. In contrast, explanationist theorists have a harder time explaining where the conviction comes from. Imagine Matt experiences a coarse-grained feeling of unfamiliarity on looking at his wife's face. If Matt's delusional belief is a result of his attempt to explain this, we would expect him to have some awareness of the explanatory reasons for believing it, and hence of what renders it justified. But this clashes with the quality of self-evident truth with which the delusion is usually maintained (Langdon and Bayne, 2015, p. 332).

Another alleged advantage of EN over ES is that it is better suited to explain the content of the delusion itself. The reason is that delusions that have been formed via endorsement processes preserve, or are very strongly constrained by, fine-grained contents of experience. In contrast, the process of content acquisition is more difficult to explain for

proponents of ES, since on this account there is no tight connection between experiential contents and belief contents. If all that Matt experiences is a coarse-grained feeling of unfamiliarity towards his wife's face, it is not obvious why he should come to believe that the face seen is not his wife's. Nor, for that matter, is it clear why he fails to seek out alternative explanations. Indeed, arguably there are more plausible ways for Matt to make sense of why his wife seems subtly different: she is about to break bad news; she is playing a prank on him; she is unhappy in their marriage; he is unhappy in their marriage; he is falling out of love; he has acquired brain damage, and so on (Gold and Gold, 2014; Langdon and Coltheart, 2000).

What ES is better suited to account for is how the anomalous experience gets its content. Because the content is so sparse, explanationists simply say that when visual familiarity data occur in the absence of autonomic familiarity data, the discrepancy is reported to consciousness, such that one becomes aware that there is something odd about the perceived person (Coltheart, 2005). Things are more complicated when we come to EN, however. For one thing, if the anomalous experience is to have a rich misidentification content like, 'this woman is not my wife', then there is the question of what properties perception can represent. Specifically, the worry is that endorsement theorists might need to accept a controversial claim in the philosophy of perception, the claim that properties such as being numerically identical with, or distinct from, a certain person are properties that one can directly perceive (Davies and Egan, 2013, p. 715).[20] Furthermore, it remains to be seen how EN is to be squared with the finding suggestive of a lack of response to familiar faces in people with Capgras. For arguably, reduced autonomic arousal in response to a one's wife face does not yet imply (at least, not obviously) that one perceives the face as being not one's wife.

3.4. Inference in Explanationism

We have now seen two different ways of developing an account of delusional belief formation, one (ES) according to which the belief arises from an attempt to explain the experience, another (EN) according to which the belief results from an endorsement of the experience. Perhaps the clearest formulation of ES is by Maher: 'a delusion is a hypothesis designed to explain unusual perceptual phenomena and developed though the operation of

[20] This is the same problem I discussed in Chapter 2 under the heading of 'experiential encoding problem'.

normal cognitive processes' (Maher, 1974, p. 103). The language Maher uses elsewhere (Maher, 1988, 2005) suggests that he conceives of the person's attempt at explaining the anomalous experiences a personal-level phenomenon. He compares the mechanisms of delusion formation to scientific theory-building (Maher, 2005, p. 142), and he notes that the content of the delusional explanation reflects from the person's scientific, religious and political background, what he calls 'general explanatory systems' (Maher, 1974, p. 103). If that is right, then, according to Maher, the stimulus input of the inference (i.e., anomalous experience), the end product of the process (i.e., delusional hypothesis), and the inferential process itself are all available to consciousness.

But could the explanatory-inferential route to the delusional belief in ES remain unconscious? There appear to be just two options for envisaging how it could. One is to argue that the inference begins from a conscious experience—e.g., the fact that one's wife feels unfamiliar—but is carried out unconsciously. Although the psychological processes leading to the belief are opaque to the person, the belief is susceptible to a personal-level explanation. Indeed, on this story, the person can appeal to the fact that his wife feels unfamiliar to explain why he believes that she is not really his wife. Alternatively, one can argue that the unconscious inference starts from an unconscious input—that is, there is no role for anomalous experience, as in the Coltheart model. On this account, there is no fact at the personal level that grounds the delusional belief.

3.5. Inference in the Endorsement Account

3.5.1. Can Perception Be Inferential?

The fundamental idea of EN is that the person with Capgras experiences the woman in front of him (who is his wife) as not being his wife, or directly as an imposter. It is tempting to assume that if you are consciously perceiving something, you are not making any inference. However, this is true only if one rules out the possibility that perception itself is inferential. Let us now then consider this possibility, and let us look at the general possibility of unconscious inference, which I have thus far taken for granted. The idea of perceptual inference can be traced back at least as far as to Hermann von Helmholtz (1867) and has been later developed by the work of a number of scholars (see e.g., Rock, 1983; Barlow, 1990; Gregory, 1997; Rescorla, 2015; Siegel, 2017). von Helmholtz (1867) proposed that some of our visual perceptions result from unconscious inferences in which sensory clues are interpreted to form hypotheses about the distal environment. The visual system would

make such inferences in compliance with what has nowadays been known as the 'likelihood principle', according to which perception represents the most likely environmental situation given the pattern of sensory stimuli. The rationale for this proposal derives from the attempt to overcome what Tyler Burge has labelled the 'underdetermination problem' (Burge, 2010, p. 91). At the core of the problem lies the fact that any given encoding of proximal stimulation underdetermines its possible distal causes. This is to say that the same proximal stimulations are compatible with a variety of different entities in the environment, and therefore with numerous possible perceptual representations. A good example of underdetermination comes from visual illusions. For example, consider the hollow-face illusion (Gregory, 1973). In this illusion, a concave facial mask lighted from below is misperceived as being convex. Here the same encoding of proximal stimulation could have been produced by a convex mask with overhead illumination. Then we would have had a veridical perceptual representation of a different distal cause. The same pattern of sensory registration is consistent with either environmental scenario, but only one perceptual state is formed.

Helmholtz's theory of perceptual inference can seemingly account for this and similar cases of underdetermination by relying on the likelihood principle. As we saw, the principle says that the visual system would check sensory cues against implicit assumptions so as to infer the most likely interpretation of those sensory cues. As such, it would seem, it may happen that reliance on implicit assumptions generates biases towards false interpretations engendering misperceptions. In the specific case of the hollow-mask illusion, the source of bias may consist of a two-fold assumption which has been thought to underlie inferences from shading to shape: (1) that there is only one light source; and that (2) the light source is positioned overhead (Ramachandran, 1998). There is, of course, a wider range of perceptual phenomena (e.g., perceptual constancies) that can be purportedly explained by appeal to unconscious inference, but I will not address them here. In the next subsections I will turn to the question of whether perception as unconscious inference can be integrated into EN. I will do so by considering the notion of Bayesian inference as a model for perceptual inference.

3.5.2. Unconscious Perceptual Inference

Michael Rescorla has identified three questions that any plausible theory of unconscious perceptual inference must satisfactorily answer (Rescorla, 2015, p. 696). These questions are:

(1) In what sense does the perceptual system carry out inferential tasks? (2) In what sense do implicit assumptions count as premises from which to draw inferential conclusions? (3) In what sense does perceptual inference select the best hypothesis among the available candidates? Currently, the most prominent approach to answering such questions is to treat the perceptual system as executing Bayesian inferences (e.g., Knill and Richards, 1996; Bülthoff and Yuille, 1991; Kersten et al., 2004).

We saw when discussing the Coltheart model how Bayes' theorem can serve as a framework for characterising inferences that are unconscious, and whose conclusions are belief contents rather than experiential contents. A problem that arises in applying Bayes' theorem to conscious perception is that of explaining how Bayesian inference can select one hypothesis to be the content of perceptual experience, or in other words, how a Bayesian inference to the effect that *p* can lead to a perceptual experience with content *p*. For instance, how does the inference of shape from shading result in the experience of an object's shape? A further difficulty lurks in the question of what kind of content perceptual experiences have. Advocates of a conservative view (e.g., Tye, 1995; Price, 2009; Brogaard, 2013) claim that perceptual experience is restricted to representing basic properties that are straightforwardly available to sensory transducers, such as spatial and chromatic ones. Other scholars (e.g., Siegel, 2006, 2010; Bayne, 2009) adopt a more liberal view, which grants that perceptual experience represents a wide array of high-level properties, including natural kinds but also perhaps causal interactions and highly specific subordinate-level categories.

Bayesian formulations of perceptual inference differ depending on which of these views one takes. If one takes the conservative view, then perceptual inferences process only low-level information, that is, information about colour, size, shape, motion, and so on. In such a context, what needs explaining is how perceptual unconscious inferences determine low-level perceptual states like seeing yellow. To illustrate, consider the case where an achromatic grey picture of a banana looks yellow (Hansen et al., 2006). From a Bayesian perspective, this misperception can be interpreted as caused by an overreliance on one's prior expectations as to what bananas look like. Indeed, the likelihood that there is a grey banana in front of one may be overridden by one's belief that bananas are yellow, yielding the inference that the banana is yellow, and thus the experience as of there being a yellow banana.

On the other hand, if one accepts the view that high-level and categorical content features in perceptual phenomenology, presumably one will also take unconscious inference to have high-level effect on phenomenal content, namely, an influence on the perceptual

categorisation of an object as an object of a certain kind. Bayesian theories of this second kind will thus have the additional task of describing how inferences underwrite one's ability to perceptually represent objects in respect of high-level categories. Consider the following case of a high-level property, where the property represented is that of *being a snake*. Imagine walking down a busy street in London and you seem to see a snake wriggling across the ground. What you are really seeing is a coil of rope and yet it looks to you as if a snake is there. A Bayesian interpretation of this example would look somewhat as follows. The prior probability of there being a snake in central London would seem comparatively too low for the likelihood ratio in favour of the snake hypothesis to override it. Yet your prior belief that the zoo is around the corner paired with your fear of snakes may lead you to favour the snake hypothesis over the rope one, with the result that the snake hypothesis is selected to be the content of experience.

3.5.3. Inference and Endorsement

Let us now return to our original question, namely, whether EN can be characterised in terms of unconscious perceptual inference. Recall that EN is based on the intuition that an anomalous experience plays a prominent role in the aetiology of the delusion. Also, recall that on the endorsement account the anomalous experience has the same, or nearly the same content as the delusion. That is to say, for the delusional belief to be that one's wife has been replaced by an imposter, one must at a minimum have an experience with the content that the woman one is looking at is not one's wife. This view can be developed in a Bayesian, inferential framework only if one assumes that perceptual inferences have a high-level effect on experiential content, in the sense sketched above;[21] and that a Bayesian framework provides a systematic means to accommodate perceptual inferences within the context of delusion formation.

We saw that one alleged advantage of EN over ES is that EN posits no (or little) gap between the content of the experience and the content of the resulting delusion. So, if the endorsement theorist wants to avoid leaving the gap open, they are committed to the claim that the person seen is experienced by the subject with Capgras delusion as being numerically distinct from a loved one. This calls for a fairly liberal conception of perceptual content

[21] This is not to deny that it may be possible for the perceptual system to keep track of individuals regardless of the properties of which they are bearers (see Chapter 2; see also Wilkinson, 2016). It is merely to deny that one can experience a person as being or not being one's wife, let alone as an imposter, without one's experience representing the instantiation of some high-level property.

according to which properties like numerical identity and distinctness are represented in perceptual experience (cf. Wilkinson, 2016 and Chapter 2). It follows that if such a content is brought about by perceptual inferences, those inferences ought to have a high-level influence on perceptual phenomenology. This means, for example, that the hypotheses generated through perceptual inferences should not be thought of being merely about the physical qualities of a person's face (e.g., shape from shading and pigmentation values), but also as being about the identity of the visually presented person, such as whether the person is numerically distinct from, or identical to, one's wife (Davies and Egan, 2013).

Can a Bayesian framework accommodate perceptual inference within the context of delusion formation? There is evidence to suggest that the perceptual system operates through unconscious inferences that conform to Bayes' theorem. We noted above that the raison d'etre behind the inferential account of perception is to explain how the perceptual system draws on proximal stimuli to yield hypotheses as to what in the environment is causing them. We further noted that the registrations of the proximal stimuli underdetermine their possible distal causes. This suggests that the perceptual system must conduct probabilistic assessments to select one distal cause over others. A number of studies have shown that this is often done by computing posterior distributions in accordance with Bayes' theorem (see for instance, Knill and Richards, 1996; Mamassian et al., 2002; Kersten and Yuille, 2003; Kersten et al., 2004). These considerations put together should make it plausible, at least, that if perceptual inference plays a role in Capgras delusion, then it can be accommodated within the Bayesian paradigm.

I have already made reference to Davies and Egan's (2013) suggestion that Bayesian probability distribution over absent autonomic activation in people with Capgras should be understood as the operation of a perceptual module. To my knowledge, this represents the only attempt so far to conceptualise EN in terms of the Bayesian framework of perceptual inference. In a previous work, Davies and colleagues proposed to describe the route from anomalous experience to delusional belief as a prepotent doxastic response to the experience (Davies et al., 2001). After the initial adoption, the delusion would be maintained due to the person's inability to refrain from accepting the experience as true. I take it that Davies and Egan's (2013) suggestion is to be interpreted as a corollary to this earlier proposal. Assuming that the delusion arises as a prepotent doxastic response to some perceptual experience, how does this experience come about? Davies and Egan's approach to answering this question

appeals to Bayesian inference. In this respect, their approach is equivalent to the Coltheart model, the difference being in the product of the inference.

Here too, we have a cluster of hypotheses, all of which reflect a possible identity of the visually presented person. Each hypothesis is assigned a certain subjective probability prior to updating on the new evidence. The prior distribution is then updated to reflect the new evidence provided by the facial appearance of the visually presented person, a person whose face looks identical to the face of the wife. After this first updating, the posterior probability of the person being the wife is higher than the probability of her being an imposter. At this stage, argue Davies and Egan (2013), the perceptual module formulates predictions as to the level of autonomic activation that would be present if either hypothesis was true. The prediction derived from the wife hypothesis is that there will be a high level of activation, whereas the prediction from the imposter hypothesis is that there will be little or no autonomic activity. If those predictions are obtained by Bayesian conditionalisation, then it would seem that the likelihood ratio between the two hypotheses conditional to the absence of autonomic activity will favour the imposter hypothesis over the wife hypothesis. The preferred hypothesis, in turn, will be appointed to be the content of experience, such that the person who in fact is the wife will be experienced as another individual (Davies and Egan, 2013, pp. 713–714).

On these grounds, Davies and Egan arrive at the conclusion that the explanationist and endorsement options are both compatible with the possibility of Bayesian inference being involved in the onset of the Capgras delusion (Davies and Egan, 2013, p. 712). The above discussion has offered further reasons to support this conclusion. We saw that inferences that are unconsciously performed can produce conscious perception and determine its content. We also saw that Bayes' theorem offers a perspicuous framework for characterising perceptual inference. All the endorsement theorist needs to argue is that the delusional belief is an endorsement of the anomalous experience. I take it that an experience is endorsed if one forms a belief that *p* based on an experience that has *p* as its content (Siegel, 2017, p. 107). So, as long as there can be inferred experiences (viz. experiences whose content *p* is determined by the outcome of an inference that *p*), there is no principled basis to deny compatibility between EN and the idea that delusions arise from inferences about the world. All of that being said, the notion of EN being implemented in terms of perceptual inference is far from unproblematic. Although it is widely accepted that unconscious processes of Bayesian inference determine how things are experienced to be, it is more controversial

whether an individual can be experienced as numerically identical or distinct from another, let alone as an imposter. In other words, it is one thing to say that unconscious inferences may be a factor behind experiencing a grey banana as yellow; it is another thing to say that they can prompt one to experience one's wife as being not really her (see Chapter 2).

3.6. Lessons from the Coltheart Model

So, does the Coltheart model embody a version of explanationism? Three features of the model are most relevant to answering the question. First of all, Coltheart and colleagues (2010) think that the content 'this woman here present is an imposter' enters into consciousness through a process of Bayesian inference on the basis of abnormal unconscious data. Second, this process is itself assumed to happen at an unconscious level. Finally, by implication, there is no causal antecedent of the delusion that can be consciously experienced.

It may seem obvious that the Coltheart model is not an endorsement theory, since it disavows any role for conscious experience prior to delusion formation. Indeed, several theorists are confident that the Coltheart model yields a modern version of explanationism. For example, Young (2014, p. 92) assures us that the Coltheart model 'is explanationist because the process of Bayesian-style abductive inference is used to select the best hypothesis from those available to explain abnormal data *O*' (see also Parrott, 2019). However, I suspect that this confidence is misplaced. As I have claimed, the Coltheart model and EN are alike with respect to the role of inference in the processes of hypothesis selection. In both cases, inference can be invoked as a means of selecting from competing hypotheses the one that best explains discrepant familiarity data. That said, if one interprets EN and ES in the strict (empiricist) sense to rely on there being a conscious experience which is causally antecedent to the adoption of the delusional belief, the Coltheart model does not fit neatly into either. For according to the model, conscious experience plays no role in delusional belief formation.

Another consideration favouring this conclusion is that the Coltheart model shares some elements with 'rationalist' accounts, such as John Campbell's (2001), that are neither explanationist nor endorsement accounts. The thrust of Campbell's account is that delusion formation 'is a matter of top-down disturbance in some fundamental beliefs of the subject which may subsequently affect experiences and actions' (Campbell, 2001, p. 89). What does this mean? First, delusions do not originate in experience, but are a direct result of

neurobiological alterations in the brain. Second, any experience that the subject reports are a consequence rather than cause of the delusion. The Coltheart model shares the latter feature, as the delusional belief is the first delusion-related event of which the person is conscious.

One must be careful not to overstate what these considerations show. They do not show that the Coltheart model is consistent with Campbell's rationalism. If the cause of a delusion is crudely organic, as Campbell suggests, then the delusion cannot be seen as the product of an inferential process, let alone a Bayesian rational one. What they do show, however, is that some of the implications of the Coltheart model tally with rationalism better than any version of empiricism. This causes further doubt about the Coltheart model being an instance of explanationism.

3.7. Conclusion

The idea that delusions are acquired through inferential processes is not new. But there has been little attempt to explain how exactly this may work and what the role of conscious and unconscious inference is in influential models of delusion formation. One notable exception is the Coltheart model, which is a detailed proposal about how delusions arise via a process of probabilistic Bayesian inference. Although the Coltheart model itself does not seem to share the typical features of either explanationism or the endorsement accounts, I have argued that the role of Bayesian inference in delusion formation is compatible with both endorsement and explanationist accounts of delusions. If the rival accounts are alike in this respect, that is, they allow for Bayesian inference to be involved in the process of delusion formation, how should they be distinguished? Those who interpret the Coltheart model in explanationist terms might respond as follows: in ES, the direction of Bayesian inference is from unconscious data to beliefs; in EN, it is from unconscious data to perceptual experiences. The problem with this distinction is that Bayesian inference plays the same role in ES that it plays in EN, i.e., to explain unconscious data, thereby blurring the difference between the two. The problem does not arise if we take both ES and EN to presuppose a grounding relationship between conscious experience and the resulting delusion. In this case, one can reasonably consider EN and ES to differ in terms of where the inferential transitions responsible for the delusion are positioned in relation to the experience (Langdon and Bayne, 2010). In EN, the key inferential transitions yield perceptual experiences, whereas in ES they start from perceptual experiences.

3.8. References

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of mental Disorders: Fifth Edition*. Washington: American Psychiatric Press.

Barlow, H. B. (1990). 'Conditions for Versatile Learning. Helmholtz's Unconscious Inference, and the Task of Perception'. *Vision Research*, 30, pp. 1561–1571.

Bayne, T. (2009). 'Perception and the Reach of Phenomenal Content'. *Philosophical Quarterly*, 59(236), pp. 385–404.

Bayne, T. and Pacherie, E. (2004a). 'Bottom-Up or Top-Down? Campbell's Rationalist Account of Monothematic Delusions'. *Philosophy, Psychiatry, and Psychology*, 11, pp. 1–11.

Bayne, T. and Pacherie, E. (2004b). 'Experience, Belief, and the Interpretive Fold'. *Philosophy, Psychiatry, and Psychology*, 11, pp. 81–86.

Bortolotti, L., and Miyazono, K. (2015). 'Recent Work on the Nature and the Development of Delusions'. *Philosophy Compass*, 10(9), pp. 636–645.

Brighetti, G., Bonifacci, P., Borlimi, R. and Ottaviani, C. (2007). "Far from the Heart Far from the Eye': Evidence from the Capgras Delusion'. *Cognitive Neuropsychiatry*, 12, pp. 189–197.

Brogaard, B. (2013). 'Do we Perceive Natural Kind Properties?' *Philosophical Studies*, 162(1), pp. 35–42.

Bülthoff, H. H. and Yuille, A. L. (1991). 'Bayesian Models for Seeing Shapes and Depth'. *Comments of Theoretical Biology*, 2(4), pp. 283–314.

Burge, T. (2010). *Origins of Objectivity*. Oxford: Oxford University Press.

Campbell, J. (2001). 'Rationality, Meaning, and the Analysis of Delusion'. *Philosophy, Psychiatry, and Psychology*, 8(2–3), pp. 89–100.

Capgras, J. and Reboul-Lachaux, J. (1923). 'L'illusion des 'Sosies'. Dans un Délire Systématisé Chronique'. *Bulletin de la Société. Clinique de Médicine Mentale*, 11, pp. 6–16.

Coltheart, M. (2005). 'Conscious Experience and Delusional Belief.' *Philosophy, Psychiatry, and Psychology*, 12(2), pp. 153–157.

Coltheart, M. (2007). 'The 33rd Sir Frederick Barlett Lecture: Cognitive Neuropsychiatry and Delusional Belief'. *The Quarterly Journal of Experimental Psychology*, 60(8), pp. 1041–1062.

Coltheart, M., Langdon, R. and McKay, R. (2011). 'Delusional Belief'. *Annual Review of Psychology*, 62, pp. 271–298.

Coltheart, M., Menzies, P. and Sutton, J. (2010). 'Abductive Inference and Delusional Belief.' *Cognitive Neuropsychiatry*, 15(1–2–3), pp. 261–287.

Davies, M. and Coltheart, M. (2000). 'Introduction: Pathologies of Belief'. In M. Coltheart and M. Davies (eds.) *Pathologies of Belief* (pp. 1–46). Oxford: Blackwell.

Davies, M., Coltheart, M., Langdon, R. and Breen, N. (2001). 'Monothematic delusions: towards a two-factor account'. *Philosophy, Psychiatry, and Psychology*, 8(2/3), pp. 133–158.

Davies, M. and Egan, A. (2013). 'Delusion: Cognitive Approaches—Bayesian Inference and Compartmentalization'. In K. W. M. Fulford, M. Davies, R.G. T. Gipps, G. Graham, J. Z. Sadler, G. Stanghellini and T. Thornton (eds.) *The Oxford Handbook of Philosophy and Psychiatry* (pp. 689–727). Oxford: Oxford University Press.

Ellis, H. D., Lewis, M. B., Moselhy, H. F. and Young, A. W. (2000). 'Automatic without Autonomic Responses to Familiar Faces: Differential Components of Covert Face Recognition in a Case of Capgras Delusion'. *Cognitive Neuropsychiatry*, 5, pp. 255–269.

Ellis, H. D. and Young, A. W. (1990). 'Accounting for Delusional Misidentifications'. *British Journal of Psychiatry*, 157, pp. 239–248.

Ellis, H. D. Young, A. W., Quayle, A. H. and De Pauw, K. W. (1997). 'Reduced Autonomic Responses to Face in Capgras Delusion'. *Proceedings of the Royal Society, London B: Biological Sciences*, 264, pp. 1085–1092.

Fine, C., Craigie, J. and Gold, I. (2005). 'Damned If You Do, Damned If You Don't: The Impasse in Cognitive Accounts of the Capgras Delusion'. *Philosophy, Psychiatry, and Psychology*, 12, pp. 143–151.

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

Gold, J. and Gold, I. (2014). *Suspicious Minds: How Culture Shapes Madness*. New York: Free Press.

Gregory, R. L. (1973). 'The Confounded Eye'. In R. L. Gregory and E. H. Gombrich (eds.) *Illusion in Nature and Art* (pp. 49–95). London: Duckworth.

Gregory, R. L. (1997). *Knowledge in Perception and Illusion. Philosophical Transactions of the Royal Society of London B*, 352, pp. 1121–1127.

Hansen, T., Gegenfurtner, K., Olkkonen, M. and Walter, S. (2006). 'Memory Modulates Colour Experience'. *Nature Neuroscience*, 9(11), pp. 1367–1368.

Harman, G. (1965). 'The Inference to the Best Explanation'. *Philosophical Review*, 74, pp. 88–95.

Helmholtz, H. von. (1867). *Handbuch der Physiologischen Optik*. Leipzig: Voss.

Hirstein, W. and Ramachandran, V. S. (1997). 'Capgras Syndrome: A Novel Probe for Understanding the Neural Representation of the Identity and Familiarity of Persons'. *Proceedings of the Royal Society of London B: Biological Sciences*, 264, pp. 437–444.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Kersten, D., Mamassian, P. and Yuille, A. (2004). 'Object Perception as Bayesian Inference'. *Annual Review of Psychology*, 55, pp. 271–304.

Kersten, D., and Yuille, A. L. (2003). 'Bayesian Models of Object Perception'. *Current Opinion in Neurobiology*, 13(2), pp. 150–158.

Knill, D., and Richards, W. (eds.) (1996). *Perception as Bayesian Inference*. Cambridge: Cambridge University Press.

Langdon, R. and Bayne, T. (2010). 'Delusion and Confabulation: Mistakes of Perceiving, Remembering and Believing'. *Cognitive Neuropsychiatry*, 15, pp. 319–345.

Langdon, R. and Coltheart, M. (2000). 'The Cognitive Neuropsychology of Delusions'. *Mind and Language*, 15, pp. 184–218.

Langdon, R. and Connaughton, R. (2013). 'The Neuropsychology of Belief Formation'. In F. Kreuger and J. Grafman (eds.) *The Neural Basis of Human Belief Systems* (pp. 19–42). New York: Taylor and Francis.

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

Maher, B. A. (1974). 'Delusional Thinking and Perceptual Disorder'. *Journal of Individual Psychology*, 30, pp. 98–113.

Maher, B. A. (1988). 'Anomalous Experience and Delusional Thinking: The Logic of Explanations'. In T. F. Oltmanns and B. A. Maher (eds.) *Delusional Beliefs* (pp. 15–33). Chichester: John Wiley and Sons.

Maher, B. A. (1992). 'Delusions: Contemporary Etiological Hypotheses'. *Psychiatric Annals*, 22(5), pp. 260-268.

Maher, B. A. (1999). 'Anomalous Experience in Everyday Life: Its Significance for Psychopathology'. *The Monist*, 82(4), pp. 547–570.

Maher, B. A. (2005). 'Delusional Thinking and Cognitive Disorder'. *Integrative Physiological and Behavioural Science*, 40, pp. 136–146.

Mamassian, P., Landy, M. and Maloney, L. (2002). 'Bayesian Modelling of Visual Perception.' In P. N. Rao, B. A. Olshausen and M. S. Lewicki (eds.) *Probabilistic Models of the Brain* (pp. 13–60). Cambridge, MA: MIT Press.

McKay, R. (2012). 'Delusional Inference'. *Mind and Language*, 27, pp. 330–355.

Miyazono, K., Bortolotti, L. and Broome, M. (2015). 'Prediction-Error and Two-Factor Theories of Delusion Formation: Competitors or Allies?' In N. Galbraith (ed.) *Aberrant Beliefs and Reasoning*. London: Psychology Press.

Pacherie, E. (2009). 'Perception, Emotions and Delusions: Revisiting the Capgras Delusion'. In T. Bayne and J. Fernandez (eds.) *Delusions and Self-Deception* (pp. 107–126). Hove: Psychology Press.

Pacherie, E., Green, M. and Bayne, T. (2006). 'Phenomenology and Delusions: Who Put the 'Alien' in Alien Control?' *Consciousness and Cognition*, 15(3), pp. 566–577.

Parrott, M. (2019). 'Delusional Predictions and Explanations'. *British Journal for the Philosophy of Science*, 0, pp. 1-32.

Price, R. (2009). 'Aspect-Switching and Visual Phenomenal Character'. *Philosophical Quarterly*, 59(236), pp. 508–518.

Ramachandran, V. S. and Blakeslee, S. (1998). *Phantoms in the Brain: Human Nature and the Architecture of the Mind*. London: Fourth Estate.

Rescorla, M. (2015). 'Bayesian Perceptual Psychology'. In M. Matthen (ed.) *The Oxford Handbook of the Philosophy of Perception* (pp. 694–716). Oxford: Oxford University Press.

Rock, I. (1983). *The Logic of Perception*. Cambridge, MA: MIT Press.

Siegel, S. (2006). 'Which Properties are Represented in Perception?' In T. Gendler Szabo and J. Hawthorne (eds.) *Perceptual Experience* (pp. 481–503). Oxford: Oxford University Press.

Siegel, S. (2010). *The Contents of Visual Experience. New York: Oxford University Press.*

Siegel, S. (2017). *The Rationality of Perception*. New York: Oxford University Press.

Stone, T. and Young, A. W. (1997). 'Delusions and Brain Injury: The Philosophy and Psychology of Belief'. *Mind and Language*, 12, pp. 327–364.

Tranel, D., Damasio, H. and Damasio, A. R. (1995). 'Double Dissociation Between Overt and Covert Recognition'. *Journal of Cognitive Neuroscience*, 7, pp. 425–432.

Turner, M., and Coltheart, M. (2010). 'Confabulation and Delusion: A Common Monitoring Framework'. *Cognitive Neuropsychiatry*, 15, pp. 346–376.

Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.

Wilkinson, S. (2016). 'A Mental File Approach to Delusional Misidentification'. *Review of Philosophy and Psychology*, 7, pp. 389–404.

Young, G. (2008). 'Capgras Delusion: An Interactionist Model'. *Consciousness and Cognition*, 17, pp. 863–876.

Young, G. (2014). 'Amending the Revisionist Model of the Capgras Delusion: A Further Argument for the Role of Patient Experience in the Delusional Belief Formation'. *Avant: Trends in Interdisciplinary Studies* (3), pp. 89–112.

Chapter 4: Delusions, Explanations, and the Predictive Mind[22]

4.0. Abstract

A growing number of studies in both the scientific and the philosophical literature have drawn on a Bayesian predictive processing framework to account for the formation of delusions. The key here is that delusions form because of disrupted prediction error signalling. A recent critique argues that the framework is incomplete in two respects: it leaves unclear why delusional hypotheses are selected over none at all or over more plausible alternatives; it leaves unclear how exactly it is that delusional hypotheses are generated in the first place. In this paper, I take up each of these points in turn. As for the first, I argue that (a) subjects with delusions may end up believing something (even delusional) rather than nothing, because only in this way can they minimise residual prediction error; and (b), that alternative beliefs may be rendered unavailable for selection, due to aberrant prediction errors. As for the second concern, I distinguish three ways in which delusional hypotheses may be initially acquired, all of which are based on experience.

4.1. Introduction

Recent years have seen a surge of interest in hierarchical Bayesian models of cognition (Clark, 2013, 2016; Friston, 2005, 2008; Hohwy, 2013, 2017). A key feature in the formulation of these models is the *predictive processing* theory of cognition, which understands the mind as a prediction machine whose objective function is to minimise *prediction error* (the discrepancy between the signals the brain 'expects', based on its model of the world, and the signals it actually receives).

Those interested in how this scheme might be instantiated in the brain often use the brain's hierarchical structural organization (particularly in sensory cortices) to suggest that the model is likewise *hierarchically* organised in interacting layers whose prior knowledge (beliefs/hypotheses) is conveyed top-down, via backward connections, to predict inputs from the layer below. Only unpredicted aspects of the input (errors) are propagated forward to adjust subsequent expectations. The upshot is a unique story about belief updating, understood as the mechanism by which organisms adjust higher level priors so as to minimise ascending prediction errors from the lower levels of the hierarchy. This process represents a biologically plausible scheme for implementing Bayesian inference in the brain, because a

[22] This chapter was written jointly with Philip Corlett.

system that minimises prediction error in the long run will necessarily approximate Bayesian inference (Hohwy, 2017).

Several projects pursued in the emerging field of computational psychiatry have sought to explain the emergence of delusional ideas by appeal to an abnormal integration of predictions with error signals, where this is alleged to misguide the revision of internal models (Adams et al., 2015; Corlett et al., 2010; Fletcher and Frith, 2009). In a recent article, Matthew Parrott (2019) argues that the predictive processing framework cannot entirely explain the adoption of delusions. Specifically, he argues that the framework suffers two limitations. First, it only partially explains why an agent believes something delusional as opposed to suspending judgment or believing something else. Second, it does not clarify how delusional hypotheses are generated in the first place. Here I attempt to show that these limitations can be met from within the predictive processing framework, or at least within bounds that are compatible with said framework.

With regard to the first limitation, I argue on the one hand that believing something, even if delusional, explains away prediction error (i.e., resolves uncertainty), and as such it is preferable to believing nothing. On the other hand, I suggest that alternative hypotheses may be in some sense unavailable to the person at the time due to aberrant prediction error signals. As for the second concern, I briefly sketch two ways to understand hypotheses spaces so as to explain how hypotheses get off the ground. I then turn specifically to the matter of how delusional hypotheses get generated. On the hierarchical Bayesian picture, the brain constructs a hierarchical model of the world that it uses to predict future sensory inputs, and to infer the causes of current inputs. Upper levels provide constraints (i.e., generate hypothesis spaces) for levels below. These constraints are called *empirical priors* (Friston, 2005) because they are learned from experience. Parrott (2019) is sceptical whether a system could generate delusional hypotheses in any way that is empirically informed. Although more research is needed, I speculate that there are at least three plausible ways in which delusional hypotheses can be understood as empirical. Ultimately, I conclude that the concerns raised by Parrott do not dilute the explanatory power of the predictive processing framework.

4.2. Predictive Processing

A key question for cognitive science is how the brain is to select from among several possible causes those which are really responsible for the observed effect on our senses (Tenenbaum et al., 2011). Many have sought to answer this question by appeal to the so-called 'Bayesian

brain hypothesis' (henceforth, 'the Bayesian brain'), according to which the brain infers the most likely causes of its sensory input in compliance with Bayes's rule (Knill and Pouget, 2004). This is an account of belief updating under conditions of uncertainty. Using Bayes's rule to update beliefs is termed *Bayesian inference*. Beliefs are represented mathematically as subjective probabilities. A probability function p accords to each hypothesis h a mean value from the interval [0,1] $p(h)$, which quantifies the agent's degree of belief in what h reports. Bayes's rule simply states that the probability of a hypothesis h after observing some new evidence e, the *posterior* $p(h|e)$, is proportional to the product of two other probabilities: the probability of observing e under h—the *likelihood* $p(e|h)$—and the probability of h independent of the observed evidence e—the prior $p(h)$.

$$p(h|e) \propto p(e|h) \cdot p(h)$$

Bayes's rule dictates how you should change your credence in h (i.e., what posterior probability you should assign to h) in light of new evidence. From the Bayesian brain standpoint, by inferring the most probable cause underlying sensory input, the brain is in fact making an inference to the hypothesis with the highest posterior (e.g., Hohwy, 2013).

Exact Bayesian inference is computationally expensive and often intractable (i.e., the posterior cannot be computed analytically). So, one challenge is to devise the algorithms by means of which the brain performs approximate Bayesian inference. A further challenge is how the brain's neural circuits might actually implement such algorithms (Friston et al., 2016). These challenges have been addressed primarily from within the predictive processing framework, as we shall now explore. Henceforth, I shall refer to predictive processing simply as *PP*.

At the core of PP lies the claim that in cognition, perception, and action the brain follows a single computational principle which essentially involves the following four steps.

(1) The brain constantly extracts regularities from the environment so as to build and maintain an internal model of the world.
(2) The brain uses the model to generate top-down predictions about incoming sensory inputs.
(3) When predictions are wrong (i.e., when a mismatch between predicted and actual input occurs), error signals are generated which can result in updates of the model.

(4) These updates serve to minimise the ensuing prediction error in such a way as to approximate optimal Bayesian inference in the long run.

Having described the general principle behind PP, let us now turn to the more specific issues of how Bayesian inference can be formalised through the PP framework and implemented in the brain through prediction error minimisation (call it *PEM*).

Assuming Gaussian (normal probability) distributions, Bayesian inference can be operationalised in the following terms. One's prior (what one already believes) works as a prediction about (a hypothesis *h* that predicts) what the next sensory input should be (Hohwy and Michael, 2018). The prediction error is the difference between such a prior-based prediction and the actual sensory input (the evidence *e* received through the senses). As such, prediction error is captured by the likelihood, for it is a measure of how well *h* predicts *e*: the closer *h* and evidence *e* are, the smaller the prediction error will be. The amount by which *h* should change to minimise prediction error, and improve the fit with *e*, provides the posterior in Bayesian inference (Hohwy, 2017). Here, the posterior is arrived at by updating one's belief (prediction or prior mean) in response to a prediction error signal. Now one crucial feature of PP is that the process of updating is not regulated just by what the prediction error reports (e.g., how far *h* is from *e*), but also by its *precision*. Bayesian inference is implemented as a process of belief updating wherein prediction error is weighted in proportion to its precision.

Precision is an index for measuring whether a prediction error signal carries information that is reliable and can be trusted. Prediction error is variable, it varies under varying conditions in the world and the organism. Ideally, the brain would accomplish PEM in conditions where noise is constant, and the variability of all error signals is the same. However, in real-world contexts the levels of variability are themselves variable due to varying levels of noise in the signals (Hohwy, 2013). A prediction error which varies as time passes due to noise in the signal does not update one's predictions, since it is more likely the reflection of random noise rather than wrong predictions. PEM must therefore take variability into account to assess whether prediction error is 'genuine' and 'worth minimising' (Hohwy, 2013, p. 65).

Precision is but the inverse of variability. A signal is imprecise if it is variable, whereas it is precise if it is invariant. In terms of the PP framework, error signals that are deemed imprecise are normally attenuated (suppressed, ignored); if the evidence is noisy, there is little to learn from it. In this case, the brain will assign more weight to its prior predictions in the

process of updating, and the posterior will be strongly influenced by the prior. So, the *learning rate* in Bayesian inference—the extent to which priors are updated in response to sensory errors—will be low. In contrast, error signals that are deemed precise are heavily weighted in process of updating, which means that the posterior will be influenced more by the data than by prior predictions. Here, then, the learning rate will be high, since if the quality of evidence can be trusted, there is more to be learnt from it (Hohwy, 2012)

According to the picture sketched so far, the learning rate will gradually decline as new evidence is observed and the prior updated to become more precise. This is, however, much too simplistic a model to capture how learning takes place in the real world. The sensory inputs to which our brain responds are generated by a volatile, ever changing environment, where multiple causes interact across different time scales. The suggestion then is that an inferential system must flexibly regulate the learning rate to accommodate contextual variables and environmental changes. For example, the learning rate in object detection should be regulated up and down depending on lighting conditions, since visual precision decreases as day turns into night (Hohwy, 2017).

This involves appeal to hierarchical Bayesian inference. The brain must infer the causes of impinging signals without having any direct access to the causes themselves. All that the brain has to go by are the statistical regularities in the signals it receives. The question is then, how can the brain minimise prediction error under such conditions? The interactions between regularities can be arranged hierarchically on several levels, with longer-term regularities occupying each higher level, and fast changing regularities occupying each lower level. Given that the input to the brain is a function of this hierarchical structure, the function must be inverted to deconvolve the hidden causes from the input itself (Friston, 2008). In its simplest form, the idea is that the brain builds a hierarchy of expectations tracking causal regularities across increasingly larger timescales, where expectations at any level serve as priors for the next level down and evidence for the next level up. The PEM mechanism is propagated upwards through the hierarchy in a way that allows a reciprocal message passing between levels: expectations encoded at higher levels send top-down predictions that can drive inferences lower down, and these in turn send bottom-up messages that inform expectations higher up.

The most biologically plausible explanation for how hierarchical message-passing is implemented in the neural circuitry is *predictive coding* (Rao and Ballard, 1999; Clark, 2013, 2016). Predictive coding is a compression procedure for reducing data redundancy in signal

processing. In this scheme, backward connections from higher cortical areas test their predictions (correlated with activity in deep pyramidal cells) against activity at lower levels to generate prediction errors (correlated with activity in the superficial pyramidal cells). What is important is that only the mismatch between the actual current signal and the predicted one––the residual error signal—is passed back up the hierarchy to refine models of the world and optimise top-down predictions (Friston, 2012). The strength of the ascending prediction error message (i.e., its impact on higher levels of the hierarchy) is a function of its precision where precision is calculated relative to the postsynaptic gain or excitability of superficial pyramidal cells (Feldman and Friston, 2010).

Neural message passing in predictive coding formulations exemplifies the way in which prediction errors are computed as quantities that the brain can access and minimise. PP's suggestion is that the brain's ongoing attempt to minimise such quantities on average and in the long run will approximate exact Bayesian inference (Hohwy, 2017; Kiefer and Hohwy, 2018).

There is much more to say about PP but the foregoing will suffice for the purpose of exploring the prospects of PP's account of delusions, to which I turn next.

4.3. Delusions and Prediction Error

Increasingly, writers about PP have proposed to understand delusions as departures from the normal functioning of prediction error mechanisms. Whilst they can disagree on how best to map the mind to the brain, they largely concur that people with delusions suffer from a disturbance in their capacity for error-dependent updating of beliefs and inferences about the world (Fletcher and Frith, 2009, p. 48).

PP theories are often presented in contrast with theories of delusion in which a sharp distinction is made between perception and cognition. These latter theories explain delusions by appeal to perceptual anomalies, either alone (in which case there is only one clinically significant causal contribution to delusion formation) or in combination with pathologically abnormal reasoning (in which case a second factor is needed). An early and now-canonical formulation of the one-factor theory came from Brendan Maher's so-called 'explanationist' claim that 'a delusion is a hypothesis designed to explain unusual perceptual phenomena and developed through the operation of normal cognitive processes' (Maher, 1974, p. 103). This means that the difference between delusional subjects and the general population does not lie in their reasoning, but in a perceptual dysfunction, to which the delusion is a perfectly

normal response. One-factor theorists do not deny that reasoning biases (e.g., biased belief-forming or maintaining mechanisms) might be involved in the processes leading from abnormal data to the onset of the delusion; they simply deny that such biases are outside the normal range of human psychology (e.g., Sullivan-Bissett et al., 2017). According to the two-factor theory, two factors are necessary for delusions: (i) a perceptual anomaly that effects the abnormal data which the delusion is developed to explain, and (ii) an abnormal deficit or bias (i.e., one which falls outside the normal range) in the reasoning responsible for the adoption and/or maintenance of the delusion (Davies et al., 2001). The most popular version of this theory holds that the second factor is a failure on the part of the agent to evaluate beliefs, where such a failure is thought to be due to frontal right hemisphere damage (Coltheart, 2005, 2007; Coltheart et al., 2007).

The PP account of delusions can be regarded as a one-factor account to the extent that it appeals to a single deficit in explaining their formation—namely, an aberrant prediction error-driven belief updating. There is, however, a fundamental difference between the PP account and the Maher-style one-factor account described just above. Unlike the latter, the former is not wedded to a neat distinction between perception and reasoning. The single deficit with which PP theorists are primarily concerned is not one that affects either the agent's perceptual systems (factor 1) or reasoning capacities (factor 2), but one that disrupts the appropriate coding of prediction errors.[23] From this perspective, the distinction between perception and reasoning is harder to pin down, since perception even at the lowest levels is shaped by predictions (Bortolotti, 2016). In other words, although it is still possible to describe predictions at higher-level cortical stages as reasoning and predictions at lower-level cortical stages as perceptions, there is a uniform kind of processing occurring across all levels of the hierarchy (Corlett et al., 2016; Corlett and Fletcher, 2015).

But there is also a point of agreement between the two accounts. This is the idea that delusions are adopted as explanatory responses to abnormal events, which in PP's case are the effects of alterations in predictive coding mechanisms. Parrott (2019) helpfully offers a fourfold schematisation of the ways that predictive coding can be altered in subjects with delusions, a typology I shall follow closely in my exposition.

[23] To be sure, attempts have been made to develop the PP account of delusions in terms of two factors, with the first factor being an abnormal prediction error and the second being the overestimation of the precision of this abnormal prediction error (Miyazono and McKay, 2019; cf. McKay, 2012). Seen in this light, the PP account for Capgras delusion outlined below involves precisely the aforementioned combination of factors. I am not convinced that prediction errors and precision-weighting are really dissociable in the envisaged manner. I need not adjudicate this matter here, though, for even if they were, nothing much is affected in the present discussion.

The most influential proposal is that the fault lies with the precision-estimation of prediction errors. There are two strands among PP theorists in this regard. According to the first, defended by, among others, Jakob Hohwy (2013, p. 158), the precision of error signals is constantly underestimated relative to sensory feedback, such that one expects all incoming external and sensible data to be noisier (and, by extension, less reliable) than they actually are. An agent with an excessive expectation for imprecise prediction errors would show heightened reliance on her internal model of the world, while continually discounting the incoming sensory feedback as unreliable. At the extreme, such an agent's model of the world would be one whose predictions are poorly compared against actual sensory signals, and where PEM takes the route of idiosyncratic interpretations of prediction error. Hohwy thinks this could lead to the sorts of entrenched delusional systems present across individuals with psychosis (Hohwy, 2015, p. 307).

An alternative possibility is that the precision of error signals is overestimated relative to prior beliefs (Clark, 2013, 2016; Corlett et al., 2016). Suppose the process of precision-weighting is disturbed in such a way that the brain constantly expects the unfolding sensory stream to be more precise than it really is (thus failing to attenuate the impact of the prediction error messages being passed up the hierarchy to higher level predictions). This would result in recurrent, highly weighted prediction errors calling for deep revisions in the agent's model of the world.

To understand how an abnormally inflated confidence in the reliability of sensory evidence could drive the formation of delusional beliefs, Parrott (2019) suggests considering the case of the Capgras delusion. This is the belief that a loved one—often a spouse or relative, for example, let us say your wife—has been replaced by an imposter who looks identical to (or at least very like) your wife. In the Capgras delusion, the connection between the face recognition system and the autonomic nervous system is damaged, so that the sight of a loved one's face does not elicit the appropriate level of autonomic response, as it does in healthy subjects (Ellis et al.,1997; Hirstein and Ramachandran, 1997; Brighetti et al., 2007; Ellis et al., 2000). The ensuing mismatch between the expected response and the actual lack of response would give rise to a misleading prediction error that is not minimised until the brain finds an explanation that can predict the unexpected signal well enough. The delusional hypothesis 'that's not actually my wife' constitutes one such explanation, for it can explain why the person's face elicits reduced or no response, rather than the expected (normal) response (Coltheart et al., 2010; Corlett, 2018). A problem is that there are people (patients

with ventromedial frontal damage; Tranel et al., 1995) who, like Capgras subjects, fail to show differential autonomic response between familiar and unfamiliar faces, but who, contrary to them, are not delusional. It thus seems that the generation of a misleading error signal (due to a failure of autonomic responsivity) alone is not enough to bring about Capgras delusion. But then what else explains delusion formation?

One possibility is that Capgras results because the misleading prediction error is imbued with excessive precision, and by extension, allotted undue influence on model revision. The claim then would be that a prediction error signal which ought to be disregarded is passed up to higher levels of the hierarchy, forcing the agent to revise an old and commonplace belief that does not have to be revised ('this person who looks like my wife really is my wife'). If Capgras subjects, unlike ventromedial patients, overestimate precision, this would explain why the former develop the delusion, whereas the latter do not (Miyazono, 2018).

Parrott (2019) mentions two other ways in which delusions can be modelled as resulting from abnormal prediction error signalling. Perhaps some delusions arise due to a system's failure to generate an error signal in the event of a mismatch between the predicted information and the actual sensory input. According to Parrott (2019), anosognosia with hemiplegia (*AHP*) might be one such case. AHP refers to a condition in which left- and right-brain-damaged patients with contralateral hemiplegia (i.e., paralysis on the opposite side of the body) deny being paralysed despite obvious evidence to the contrary (Davies et al., 2005; Fotopoulou et al., 2008).[24] For example, an AHP patient with right brain damage might insist that she can clap (or indeed is clapping) even though her left hand is motionless and no sound of clap is audible (Berti et al., 1998, pp. 29-30).

It has been established that patients with AHP remain able to generate motor commands and predict the sensory consequences of movement (e,g., Fotopoulou, 2012a; Garbarini et al., 2012). Normally, these predictions will be compared against sensory evidence to enable belief updating on the basis of error (e.g., update to the belief that the intended movement is not performed). Yet patients with AHP fail to update their premorbid beliefs about their motor abilities. One plausible explanation is that, while they form accurate representations of the predicted movement of their limb, they are unaware of the mismatch between their prediction and the impending sensory feedback (Berti et al., 2007; cf. Frith et

[24] Note, however, that AHP is far more common after right than after left hemisphere damage.

al., 2000). On this view, patients' awareness is monopolised by motor predictions, to such an extent that it overlooks sensory evidence about lack of movement. Some take the further step of arguing that, because no mismatch is registered, there is no error signal available to inform motor awareness as to the absence of the anticipated movement (Davies et al., forthcoming). This would explain why, even when observations do not match predictions, patients remain unaware that the intended movement has not actually been executed. To be sure, though, the hypothesis that no prediction error is generated is only one possible explanation. Prediction errors need not be absent in order for motor predictions to go unchecked; they could just be weak or not sufficiently precise to cause revision (Fotopoulou, 2015). If the latter, AHP too may be dependent on a problem with precision estimation (more on this below).

Finally, another way that predictive coding could be disrupted is by generating 'false' prediction errors (Fletcher and Frith, 2009), i.e., errors in response to events that should not be treated as a predictive failure. The problem here is not that a system fails to generate or adequately register prediction errors when the input mismatches what is predicted, but rather that prediction errors are generated when there is no real mismatch between expectancy and outcome (Corlett et al., 2010; Corlett et al., 2016; Corlett, 2018). According to this proposal, the occurrence of false prediction errors cause the brain to adjust one's existing model of the world in order to fit the unpredicted aspects of the incoming sensory signal. Yet, since the errors are false, no amount of adjustment can ever accommodate them. Consequently, these persistent unresolved errors are propagated upwards to ever-higher levels of the cortical hierarchy, where they update the world model with new and inappropriate learning, culminating in delusions.

As Parrott (2019) rightly notes, the idea that links delusion formation to falsely generated error signals is by no means inconsistent with the one which links it to a problem of precision estimation (p. 12). Indeed, the claim is often made that delusions are formed as a way to explain prediction errors that arise inappropriately and with an abnormally high-level precision (Corlett et al., 2016, p. 1146; Corlett, 2018, p. 54; Corlett and Fletcher, 2015, p. 98). One example where this might happen is thought insertion (Parrott, 2017; Parrott, 2019). Thought insertion is a condition in which one reports that thoughts that are not one's own are being inserted into one's mind. A possible explanation for why one might form such a belief is that false prediction error signals are generated when one is engaged in spontaneous thought. Thoughts which are co-incident with false error signals (and ones that perhaps are

afforded excessive precision) would be coded as irregular or unpredicted. As a result, one would experience one's own thoughts as oddly salient or anomalous without being able to understand why. Confronted with these experiences, one would plausibly search one's mind for an explanation, and the hypothesis that the thoughts are not one's own but are being placed into one's mind might provide such an explanation.

To sum up: there are at least four ways in which the PP framework can be developed to account for the emergence of delusions. Each of these ways corresponds to a distinct alteration in predictive coding. This tells us that the PP framework may be flexible enough to accommodate several different kinds of delusions in terms of a failure (whatever, exactly, it might be) in predictive coding. However, Parrott (2019) claims that important open problems remain to be faced if PP accounts are to genuinely illuminate how delusions emerge under the influence of aberrant prediction errors. These problems reflect concerns that accounts like those just described might be lacking in explanatory power. In Section 4.4. I explore these concerns in some detail and in Sections 4.5., 4.6., and 4.7. I sketch how the PP framework might plausibly address them.

4.4. Contrastive Why-Questions and Explanatory Nonstarters

As we have seen, the PP framework shares the explanationist assumption that the formation of a delusional belief is the end product of an explanatory process. One objection that is often heard is that explanationism is unable to account for why individuals fail to suspend judgment or choose alternative explanations for the experiences on which their delusions are based. Return to the case of the Capgras delusion. According to explanationism, the hypothesis that the person observed is not in fact one's wife is adopted to explain some irregular experience elicited by the unexpected lack of autonomic response to their faces. But it seems that this does not tell us why that hypothesis is chosen instead of none at all, or instead of some other one. Indeed, one could ask why someone believes that the person observed is not one's wife rather than suspending judgment, or rather than believing that he is falling out of love, or that he has acquired brain damage, etc. This objection can be reinstated against PP theories of delusion formation. The search for the causal explanation of some phenomenon is a search for the reason why it happened, and thus any such search can be interpreted as a way of answering a why-question. In the case of delusion formation, the relevant question is why delusional hypotheses are adopted. Let us call this the *adoption* question.

Parrott (2019) argues that PP accounts do not provide a full answer to that question, and this because they only seek causes to explain why an agent fails to believe the most plausible hypothesis but do not address the contrastive question of why the delusional hypothesis is adopted instead. The starting point is the intuition that whenever we ask a why-question, we assume a contrast class, such that the question 'Why this?' has implicitly built into it the phrase 'rather than that' (Lipton, 2004). In cases where the explanandum is the adoption of a delusional hypothesis, Parrott says that the contrastive foil is what we would expect the average person to believe (the 'obvious belief', as he calls it). For example, we would expect someone who sees her mother's face to believe that the face seen is that of her mother. Or, similarly, we would expect someone who reports herself as thinking thoughts to believe these thoughts are exclusively her own. If this is correct, then we can rephrase the adoption question as that of why delusional hypotheses are selected, rather than obvious beliefs. Appeals to alterations in predictive coding provide a good foundation to understand why people with delusions fail to hold the obvious belief. Consider, for instance, the idea that the occurrence of erroneous prediction errors would lead to inappropriate updating of one's current internal model of the world. Presuming that such a model includes the obvious belief, we can see how aberrant predictive coding could cause one to abandon the obvious belief that we would expect them to have. However, Parrott (2019) complains that this still does not fully explain the adoption of the delusional hypothesis. As he notes, in most cases of delusion the range of options from which an agent can select is not restricted to the delusional hypothesis and the obvious belief. One could simply suspend judgment until further evidence presents itself. So, although the lack of an obvious belief can be explained, that alone does not tell us why the delusional belief is adopted, rather than none at all. But, argues Parrott, even together with the assumption that one ought to believe something, no specific impairment in predictive coding can tell us why the delusional hypothesis h is preferred over any of the more plausible alternatives $h_1, h_2 \ldots h_n$ (h's multiple foils) besides the obvious belief.

A second standard objection to explanationism, pressed forcefully by Cordelia Fine and colleagues, is that it fails to account for the fact that delusions are often absurd, fantastic, or implausible accounts of the events they purport to explain (Fine et al., 2005, p. 160). Indeed, argue Fine and colleagues (2005), delusions are explanatory 'nonstarters', which is to say that they are not even candidates to serve as potential explanations. The objection can be put equally forcefully for the PP framework. Let the 'candidate set' (used interchangeably

in this essay with 'hypothesis space') represent the set of potentially explanatory hypotheses for some observed phenomenon. According to this objection, what PP theories need to explain is not just how the brain constructs a prior probability distribution over the candidate set, but also, and more importantly, how the brain generates the members of that set (Parrott, 2019). From a PP perspective, to ask how a hypothesis is generated is to ask how the brain selects a hypothesis space with that hypothesis as a member. Since the brain would normally select the space that better minimises prediction error, Parrott (2019) speculates, if a space includes an implausible hypothesis, the most likely explanation is that some kind of disruption to predictive processing could lead a system to select an abnormal hypothesis space. However, Parrott thinks this just pushes things back one step. If the number of possible hypotheses were infinite but countable, PP theorists could make simplifying assumptions to reduce complexity (e.g., they might sample from complex distributions). But these simplifying assumptions do not hold in realistic contexts where the number of possible hypotheses is indefinitely large, making it unclear as to how a hypothesis space including a nonstarter could become available for selection by a system. The worry, then, is that even if impaired predictive coding can account for aberrant learning over a hypothesis space, that does not explain how such a space could include a nonstarter as a candidate for belief.

Parrott (2019) also notes that this question cannot be answered by referring to priors as 'empirical', i.e., priors that are informed by empirical evidence (Friston, 2005). For a simple example, consider the 'ventriloquist effect' (e.g., Bertelson, 1998) where a visual cue (the mouth movements of the dummy) will trick the brain to misjudge the location of the sound source (the ventriloquist's voice). Plausibly, what happens is that the brain is biased by empirical priors (i.e., in this case, priors from past experiences of synchrony for audio-visual speech stimuli). One worry is that even if the appeal to empirical priors could be important in understanding aspects of ordinary cognition, it is useless when it comes to understanding how priors are generated in cases of delusion. This is because, the worry goes, since delusional cognition is characteristically insensitive to empirical evidence, delusion-related priors like 'my thoughts are under the control of alien forces' cannot be estimated from sensory data. In the remainder of this paper, I propose some ways in which the PP framework might address the above concerns.

4.5. Why Believe a Delusional Proposition Rather than Nothing?

First limitation: PP accounts are good explanations of why an agent fails to believe the obvious proposition that we would expect her to believe, but they explain neither (Q1) why the agent believes a delusional proposition rather than suspending judgment, nor (Q2) why she believes a delusional proposition rather than some other proposition.

One way out of (Q1) is simply to claim that the prediction error responsible for discarding an obvious belief calls for a new belief to minimise it. The error indicates that a person's internal model of the world is defective and must be supplemented in a way that will fill in the void left behind by the obvious belief. Let me make a bit more tangible how this might work. Consider the context in which delusion-prone individuals initially find themselves. In this context, one or more obvious beliefs have failed to make an accurate prediction. As I have repeatedly stressed, a system's overarching goal according to PP is to reduce surprise and resolve uncertainty. For this to be done properly, more is required than the mere discarding of the current hypothesis: there has to be some new hypothesis under which prediction error is minimised. Delusions may then be instrumental to this end. A system which is strongly motivated to re-establish predictability may prefer believing something (even if false, even if delusional) to believing nothing at all, because only in this way can it be kept out of states that are surprising for it (Hohwy, 2015). Borrowing a term from social psychology, one can frame this point in terms of a 'need for closure' (e.g., Kruglanski, 1989), which signifies a yearning for certainty and an aversion toward uncertainty.[25] To oversimplify the point to a slogan, believing something delusional brings to the agent 'closure' that could not be attained merely by suspending judgment.

It is noteworthy to add, though, that in some cases discarding an obvious belief may be tantamount to accepting a delusional proposition. A perfectly natural way to discard a belief is through accepting its negation or denial. Suppose I stop believing that it is raining because it stops raining. Is the appropriate attitude here to simply replace the attitude of belief with that of suspension? Presumably not. It seems fairly clear that I should disbelieve that it is raining. Now consider again the case of the Capgras delusion. The kind of belief in question, e.g., 'that is Ada', results from empirical learning and is only obvious with respect to certain contexts, ones in which Ada is present. When the PP system registers a mismatch between the expected and actual response to Ada's face, this suggests that the system is in a

[25] For evidence of self-reported need for closure in delusional and delusion-prone individuals, see McKay et al. (2007) and Colbert and Peters, 2002.

context where 'that is Ada' does not hold true. Parrott (2019) seems to be assuming that suspension is the appropriate attitude the system should have toward 'that is Ada'. But it is not obvious that the system should cease to believe 'that is Ada' without believing its negation. It is equally possible, and perhaps more probable, that the belief would be discarded as false. In that case, and assuming (as most people do) that a *p*-disbelief is nothing more than a ¬ *p*-belief, the system is left with little choice but to accept the delusional proposition 'that is not Ada'.

So, what is the upshot? Parrott (2019) is right that one way to discard beliefs is by suspending judgment on their contents. But that is not always the case. Sometimes—indeed often or typically—we discard beliefs by deeming their contents false or, which is the same, by believing their negations. When the belief so discarded is an obvious belief in *p*, the resulting belief in the negation of *p* may very well wind up delusional. If this is correct, then (pace Parrott) what we know about why someone discards an obvious belief does (at least in some cases) help explain why she believes something delusional rather than nothing at all.

A stronger case can be made that even if one suspends judgment about *p* at time t_1 one might still be led to believe *p* or some importantly related content at time t_2. Philosophers have sometimes described suspending judgment as simply lacking belief on some matter (e.g., Chisholm, 1976). Many now argue that this is too a restrictive understanding of what it means to suspend judgment (e.g., Friedman, 2013). Instead they understand suspension as an attitude of its own, what Scott Sturgeon (2010) has called 'committed neutrality'. Attitude accounts differ with respect to (i) whether suspension is a sui generis attitude or reducible to ordinary attitudes like beliefs; (ii) whether suspension is a metacognitive or first-order attitude; and (iii) whether suspension is a propositional or question-directed attitude. For present purposes we need only consider a consequence of the propositionalist view and we can set aside the issue of which combinations of the above features can be found in the literature. In suspending about *p* one has an attitude at least towards *p* and some importantly related content (e.g., ¬ *p*). As such, being in a suspended judgment at *t* involves one having grasped or considered *p* and some importantly related content at *t*. This means that if one suspends about an obviously true proposition (e.g., 'that woman who looks like my wife and claims to be my wife is in fact my wife'), then one is likely to have grasped or considered its (possibly delusional) negation (e.g., 'that woman who looks like my wife and claims to be my wife is in fact not my wife').

Interestingly, there is evidence that merely grasping a proposition tends to increase credence in it (e.g., Gilbert, 1991; Gilbert et al., 1990). This has been reported even in cases where the person knows the proposition is false. For example, Daniel Wegner, Gary Coulton, and Richard Wenzlaff (Wegner et al., 1985) devised an experiment in which participants were warned that they would be getting false feedback about how they had performed on a task. They were then exposed to such feedback while under time pressure. When later asked to estimate their performance accuracy, they nevertheless formed judgments that mirrored the feedback. A plausible account of this is that time pressure depletes participants' cognitive resources, degrading their ability to override false information upon comprehension (Mandelbaum, 2014). It could well be that a similar mechanism is at play in delusion formation, where there is an urgent need to come to terms with a state of unexplained uncertainty. Consistent with this, longitudinal studies have reported that the formation of delusional beliefs is often not a rapid, one-step process, but one in which a specific proposition, at first merely grasped through cognition, gradually gains credence and evolves into a fully delusional state (Moritz and Woodward, 2004).

A critic may respond by claiming that suspended judgment is an attitude with a question (rather than a proposition) as its content. If what is suspended on is a question, suspension per se need not involve one having grasped or considered any proposition (Friedman, 2013). For instance, one would suspend about who a person is rather than the propositions that she is or is not one's wife. However, for the reasons already discussed, it is psychologically implausible that an alarming state of uncertainty about the world could consist only of an interrogative attitude in which no hypotheses are contemplated.

There is, moreover, a further problem that I have yet to mention. As I said above, I take it that suspension about p likely involves having considered both p and $\neg p$. Take now p to be an obviously true proposition and $\neg p$ to be its delusional negation. If considering $\neg p$ under time pressure is enough to induce belief in $\neg p$, then the same is true for p. But that means one would end up believing contradictory propositions, for instance, that the same person is and is not one's wife. I will explore the specifics of this issue in more detail below (Chapter 5). For now, suffice it to say that the possibility of contradictions being believed is a genuine possibility.

To sum up: there are at least three possible responses to (Q1). One response is that adopting delusional beliefs fits with the broader aim to keep an organism in states that best minimise surprise (hence respond to prediction errors). The failure to respond to prediction

errors prompts the formation of new beliefs. Belief, however false may it be, better minimises surprise than no belief at all. A second response is to point out that often when we cease to believe something previously believed, we believe its negation. In the case where the content of the negated proposition is obviously true, the resulting belief may well be called a delusion. The upshot is that, at least sometimes, the obvious belief and the delusional belief are exhaustive, in the sense that discarding the former already means adopting the latter. A third response is to say that suspending judgment about *p* may nonetheless drive an unconscious passive acceptance of *p* or some related content. This gives us an alternative possible account of why an agent believes something delusional rather than suspending judgment: perhaps the agent begins by suspending and ends up believing one of the propositions which are being suspended on (by virtue of having merely grasped it).

4.6. Why Believe a Delusional Proposition rather than Something Else?

As soon as we pose (Q1) another question (Q2) immediately arises. Even granted that the agent has a 'preference' for certainty (believing something) over uncertainty (believing nothing), why is it that a delusional hypothesis is chosen over more plausible (non-delusional) alternatives?

A possible answer is that epistemically better explanations are simply unavailable to the agent (i.e., she has no better choice). Daniel Freeman and colleagues (Freeman et al., 2004) asked a sample of 100 individuals with delusions at the acute stages of psychosis whether there were alternative explanations for the experiences on which their delusions were based (even should they disagree with these alternatives). They found that three-quarters of the sample could not report alternative explanations to the delusion.

Ema Sullivan-Bissett (2015) provides a useful taxonomy for classifying different types of unavailability, each of which can explain the failure to adopt less epistemically costly hypotheses. The taxonomy covers strict unavailability, motivational unavailability, and explanatory unavailability. a) Alternatives are *strictly* unavailable if they depend on information to which the agent has no access whatsoever (e.g., destruction of memory due to synapse failure). b) Alternatives are *motivationally* unavailable if the information is not completely gone but has been rendered inaccessible by motivational factors. c) Alternatives are *explanatorily* unavailable in cases where they appear poor enough to be disqualified as

potential explanatory models (Sullivan-Bissett, 2018, pp. 925–926).[26] I will now argue that the PP framework is consistent with at least the last two senses of unavailability. (I wish to remain neutral on whether there are, in fact, delusions in which alternatives are strictly unavailable).

Let us start with motivational unavailability. As a case in point, consider an AHP patient who denies that her left arm is paralysed.[27] We have seen that PP theorists can explain this in terms of absent or unreliable prediction errors about the actual state of the paralysed limb, which result in premorbid, habitual predictions exerting undue influence on motor awareness. The lack of a prediction-error-based updating means that motor planning dominates motor awareness, with the effect that the patient considers she has moved her arm when actually she has not (Fotopoulou et al., 2008).

For this approach to work, however, there would have to be an explanation as to how patients with AHP could possess implicit knowledge of their paralysis. Indeed, empirical evidence has found that the presence of a deficit is recorded at some deeper level of cognitive process which is usually (but not always) inaccessible to conscious awareness. One suggestive finding is that unawareness can be made to temporarily remit using vestibular stimulation, which consists of irrigating the contralesional ear with cold water (Bisiach et al., 1991; Cappa et al., 1987). For example, Ramachandran (1995) demonstrated that after just a few seconds from the start of the stimulation, a previously anosognosic patient was able to report that her left side was paralysed and had been for several weeks. As the effects of the procedure had worn off, she again reverted to denial and could not recall her earlier admission of paralysis (though she did remember that her ear had been irrigated).

The fact that some patients can transiently gain insight after vestibular stimulation shows that the presence of a deficit had been recorded in the cognitive system all along, only to be overridden by higher-order cognitive demands (Ramachandran and Blakeslee, 1998, p. 105). One plausible hypothesis is that it is defence mechanisms which keep such tacit knowledge of the deficit from reaching awareness (see Turnbull et al. (2014) for a recent development). Rather than adjusting to a profoundly aversive situation, as is a breach of bodily integrity, the patient diverts the unwanted information into the unconscious. This

[26] Here I adopt a slightly different definition in comparison with that proposed in Sullivan-Bissett (2018), whereby alternatives are deemed explanatorily unavailable only if they strike the subject as implausible (p. 941).
[27] The reader will notice that AHP is somewhat a special case; one in which an obvious belief is itself turned into a delusion owing to a change of circumstances. However, it is important to realise that the basic question remains the same: why does an agent believe something delusional rather than something else?

would make for a credible explanation of why non-delusional alternatives are unavailable to the patient—namely, they are in conflict with a strong motivation to maintain bodily integrity.

But can we make sense of motivational unavailability in terms of PP? To answer this question, we need to ask how motivational factors can preclude prediction errors from influencing prior beliefs. My conjecture is that they do so by interfering with precision estimation, which is indispensable for reality testing.

To begin with, recall that to minimise prediction errors, the brain needs to estimate their precision, or equivalently, to predict how reliable these errors are. Precision estimates modulate the weighting on prediction errors accordingly, and in so doing adjust the relative influence of sensory evidence and prior beliefs along the processing stream (Clark, 2016, p. 57). Higher expected precision means a higher weight on the relevant prediction error, and thus a higher influence on hypothesis revision. Conversely, signals expected to have low precision, since they cannot be trusted to recruit reliable information, are dampened down, where this means that they carry less weight, and are thus less apt to drive update (Hohwy, 2013, p. 66).

A growing body of evidence about so-called 'optimistic update bias' suggests that the valence (positive-negative) of incoming information about the future determines the extent to which humans update beliefs: we readily assimilate desirable information into our existing beliefs, but we are reluctant to accommodate our beliefs with undesirable information (Sharot and Garrett, 2016; Kuzmanovic and Rigoux, 2017; Sharot et al., 2012). For example, we are less likely to update our beliefs when we receive information that average possibility of experiencing a negative event, such as being robbed, is higher than they had expected (Sharot et al., 2012).

Sharot, Korn, and Dolan (2011) found that selective updating was correlated with a diminished neural coding of estimation errors that called for negative update, suggesting their signalling is underweighted in overall processing. The PP framework provides a natural explanation for this in terms valence-dependent precision estimation. A motivation to maintain positive self-regard induces an expectation for low precision on negative prediction errors, thereby decreasing the influence these errors have on updating.

This tendency to discount negative information may be especially reinforced by at least two conditions: first, if motivational modulation further sabotages already suboptimal precision parameters; second, if positive priors are very strong.

If negative signals are already expected to be imprecise, motivational modulation will decrease their already low-precision estimate even lower. As a result, such signals will be too underweighted to have even the slightest influence on updating. On the other hand, if positive priors are very strong, only large and sustained predictions errors will make it possible to update them; those upon which precision has been lowered by motivation will stand no chance.

Two things should be noted here. First is that each condition gives us a useful grip on the notion of motivational unavailability; since motivation precludes updating, there is no need to consider alternatives that differ from initial beliefs. So, in this sense, alternatives are motivationally unavailable. Second, these conditions are not mutually exclusive. Both may obtain simultaneously—and they likely do in individuals with AHP. For one thing, precision is thought to be encoded by dopamine in terms of postsynaptic gain. This means that anything that affects dopamine release also affects precision estimation. Crucially, there is evidence of fronto-striatal dopamine depletion in AHP (Fotopoulou et al., 2010). For another thing, patients with AHP may premorbidly present with very strong priors about their body, e.g., a positive body image, which are particularly difficult to relinquish (Fotopoulou, 2014).

The upshot is that motivational unavailability can be understood in PP terms and can help explain why in AHP alternatives are not adopted in place of the delusional belief. To sum up, the idea is that the estimates of the precision associated with negative predictions errors (i.e., ones signalling unpleasant facts, and in this case, the presence of a deficit) are biased down by motivation, such the delusional belief is not updated, nor are alternatives considered.[28]

I see no reason why, in principle, this approach could not be applied to other delusions in which defence mechanisms are to play a role (think, for instance, of the Reverse Othello syndrome, in which a person retains belief in the fidelity of his or her partner despite strong evidence to the contrary, see e.g., Butler, 2000). More in general, I think it may be applied in any case where delusional beliefs are favoured over alternatives involving profoundly aversive realisations (e.g., 'I am going mad') or disability-related thoughts (e.g., 'I have suffered from brain damage').

[28] It is a hard question, one not addressed here, why vestibular stimulation transiently improves anosognosia. Like McKay et al. (2013), I speculate that it selectively augments belief updating in response to negative prediction errors, perhaps by increasing the extent to which patients expect precision in them. At any rate, mine remains only a speculation, which awaits future research.

What about cases, though, where alternatives are not unsettling but instead quite mundane? For example, why does the Capgras subject believe that the person in front of them is not their wife, rather than something is off about her today, or that her attire makes her look subtly different? Here the concept of explanatory unavailability comes in handy.

Perhaps the most in depth treatment of this concept in the context of PP is to be found in the work of Jakob Hohwy (Hohwy, 2013; cf. Hohwy and Rosenberg, 2005), although he does not himself make use the term 'explanatory unavailability'. On Hohwy's view, mundane hypotheses like the ones mentioned above are not selected because they are unable to explain away aberrant prediction errors at the right spatiotemporal fineness of grain (Hohwy, 2013, p. 161).

It is a peculiar feature of PP that levels in the hierarchy pertain to different spatiotemporal scales (Clark, 2012). Low levels predict causal regularities over short spatiotemporal scales, which means they encode rapidly changing information from multiple (exteroceptive, interoceptive, and proprioceptive) sources. Presumably, autonomic prediction errors of the sort involved in Capgras arise at low levels of the hierarchy, since arousal responses vary greatly across different perceptual conditions. Now, compared to the delusion in its nascent state, which amounts to misidentification, either of the above alternatives has a reasonably high prior probability of being true. After all, the prior probability that a person who looks/sounds like your wife and claims to be your wife is not your wife should be extremely small. Even so, alternatives may fail to suppress prediction error efficiently lower down in the hierarchy, which is simply to say that they cannot make it go away. For example, top-down predictions like 'something is off about her today', or even relatively more aversive ones like 'something is wrong with my brain', may be too general, too abstract, or just unfit for the level of spatiotemporal fineness of grain at which the prediction error arises.

This gives us a sense in which alternatives may be explanatorily unavailable; they may fail to apply at the spatiotemporal scales relevant for minimising prediction error. The delusional hypothesis may be favoured, then, because it succeeds where competing hypotheses fails—namely, in explaining away prediction error at a fine-grained enough spatiotemporal scale.

There is another sense of explanatory unavailability, which can be understood in terms of a bias toward observational (or explanatory) adequacy ('OA bias' for simplicity), a tendency to update beliefs so as to accommodate the evidence, while disregarding the prior

probabilities of relative candidate hypotheses (Stone and Young, 1997). Some writers have suggested that the OA bias is at the heart delusion formation, as evidenced by the fact that people with delusions tend to favour hypotheses with more explanatory power over ones with more general plausibility (McKay, 2012; Stone and Young, 1997). On this view, for example, subjects with Capgras believe what is most likely to explain the lack of autonomic responsiveness to a familiar face, namely that the face is in fact not familiar (cf. Coltheart et al., 2010). Miyazono and Mckay (2019) plausibly suggest that the OA bias is underpinned by an overestimation of the precision of aberrant prediction errors (see Section 4.3. above).[29] A system in which aberrant prediction errors are misestimated to be overly precise will give them undue weight in the process of belief updating (Frith and Friston, 2013). The tendency in such a system will be to privilege hypotheses that best fit highly weighted data, no matter how improbable these hypotheses are before observing the data. And since data are abnormal, the likely result will be delusions. In this sense, mundane alternatives may be explanatorily unavailable simply because they do not fit the data equally well—namely, they have lower likelihood. If so, the answer to (Q2) is that delusions do a better job explaining away ultra-precise prediction errors.

Before moving on, it is worthwhile to note that the two notions of 'explanatory unavailability' we have identified are not mutually exclusive as more realistic alternatives may be explanatorily unavailable in both senses at the same time. That is, alternatives may be less likely given the aberrant data, while also failing to match the spatiotemporal fineness of grain of the relevant prediction errors. To be sure, this is not always the case (even if it is usually the case). For instance, in Capgras, the hypothesis 'this is my wife' is parsed at the same spatiotemporal scale as the hypothesis 'this is not my wife', and yet it provides a less likely explanation for the lack of autonomic response.

4.7. Where Did the Delusional Hypothesis Come from?

Second limitation: PP accounts fail to explain (Q3) how implausible candidate hypotheses are generated. Section 4.4. introduced this problem in detail, but I need briefly to restate it here so that my response is clear. Typical Bayesian systems presuppose knowledge of a hypothesis space and a prior probability distribution over that hypothesis space. To specify the

[29] According to Miyazono and McKay (2019), the OA bias, so understood, is the second factor casually responsible for delusion formation, alongside a misleading prediction error—i.e., one that occurs when there is no real mismatch between predicted and actual input. Whether this interpretation is plausible is an issue that I will not take up here (though see fn. 23 above).

hypothesis space is to specify a set of hypotheses about the causal structures by which the observed phenomenon could have been generated. Learning involves searching through candidate hypotheses to find one that is most likely to be correct. To this extent, Bayesian systems can only test hypotheses within an already-specified hypothesis space. Assuming that delusional hypotheses are not already members of an agent's hypothesis space, there is the problem of explaining how they become candidate explanations.

The problem can be broken down into two questions, namely, 1) 'How do we conceptualise a system that has access to a broader range of hypotheses than the ones it has already entertained?' and 2) 'What causes a system to select highly irregular hypothesis spaces, which include implausible candidates?' One way to address (1) is to distinguish between *parametric* and *nonparametric* Bayesian systems (Friston, personal communication). A full explanation requires technical resources which I do not have space to discuss (here, I will simplify matters; a more detailed treatment can be found in e.g., Austerweil et al., 2015)

Parametric and nonparametric systems employ two types of hypothesis spaces. Parametric systems take the set of possible hypotheses to be settled a priori. Consider the problem of inferring hidden structures from observed data. The key characteristic of parametric systems is that they commit to a fixed number of structures which does not change no matter how much data are pooled. This means that the structure inferred by these systems will not be the hidden structure generating the sensory information unless the hidden structure is already part of the system's hypothesis space (Austerweil et al., 2015).

By contrast, nonparametric systems do not impose any built-in constraints on the number of structures to be expressed in the observed data. They allow for a potentially infinite number of structures, only a subset of which is represented at a time. When these systems encounter sensory signals which they cannot currently explain (as may happen for sensations unexperienced before), hypotheses spaces are extended on the fly to accommodate the evidence. This flexibility enables nonparametric systems to infer rich latent structures from sensory data, without having to specify them in advance (Austerweil and Griffiths, 2009).). Neurobiologically, this may correspond to the use of latent synapses and connectivity that is called upon in the appropriate circumstances (Friston, personal communication).

Another way to address (1) is to make a distinction between *latent* and *explicit* hypothesis spaces (Perfors, 2012). A system's latent hypothesis space consists in the system's representational capacities (i.e., primitives). It denotes the space of all hypotheses that the

system is capable of representing. The explicit hypothesis space denotes the hypotheses that are currently being represented by the system—the hypotheses that the system has explicitly identified for inference and testing. Under this interpretation, hypothesis generation is the process by which hypotheses are moved from the latent to the explicit hypothesis space. Of course, this should not be taken too literally. Explicit hypotheses are not simply read off from the latent space. Rather, they are constructed by manipulating psychologically primitive resources that exist within the latent space (Perfors, 2012).

The foregoing considerations deal with hypothesis generation in the abstract, and as such they tell us hardly anything about how hypothesis generation plays out in concrete instances, let alone delusional ones. This is the question of (2) above, to which we now turn.

I consider three speculative, open-ended possibilities here, all in need of further empirical exploration. First, it might be thought that unexplainable sensations (e.g., persistently high surprise) cause the current hypothesis space to be discarded, leaving room for completely novel hypotheses to develop (something like this is envisaged by Parrott, 2019, p. 24). Hierarchical Bayesian models allow for multiple hypothesis spaces ranging over different levels of abstraction. Hypotheses at the higher levels of the hierarchy (e.g., Level 2) constrain or generate hypotheses at Level 1 and are in turn constrained or generated by still others at Level 3, higher up in the hierarchy. This means that hypotheses at each higher level can be updated to reflect changes at the levels below, similar to the way in which hypotheses at Level 1 are updated after the receipt of new data.

Kemp and colleagues (2007) helpfully capture this feature of hierarchical Bayesian modelling with the notion of *overhypotheses*, where overhypotheses refer to 'any form of abstract knowledge that sets up a hypothesis space at a less abstract level' (p. 308). To illustrate, consider an example taken from Nelson Goodman (1955). Suppose you are presented with a stack (*S*) in which are contained several bags of marbles. After opening many different bags, you discover that each bag contains either all black or all white marbles. Suppose now you open a new bag (*n*) and draw a single black marble out of it. Though a single draw is not decisive, prior experience may lead you to entertain the hypothesis that *n* contains all black marbles. If I then ask you what justified choosing that hypothesis, you may appeal to the following overhypothesis:

O: all bags in *S* have uniform color distributions.

The reason that *O* is an overhypothesis is that it constrains the possible hypotheses about each newly encountered bag – e.g., 'uniformly black', 'uniformly white', 'uniformly red', and similar other such hypotheses (Kemp et al., 2007, p. 308). To the extent that they can be learned from experience, overhypotheses are empirical priors.

Consider how this applies to the case of delusions. Let us take thought insertion as an example. It is generally agreed that thought insertion involves an experience of one's thoughts as somehow not under one's control (or at least a failure to experience one's thoughts as under one's control). Normally, we experience thoughts as nobody else's but ours, and we encode this knowledge in the form of an overhypothesis, at the level that describes thoughts in general. We might think of it as something like 'I am the unique thinker of my own thoughts', though of course other interpretations are possible, for instance, 'all the thoughts in the domain *D* belong to me'. Overhypotheses in hierarchical Bayesian systems are tested against data, and the one with the best fit to the data sets up a space of hypotheses at the next level down. If none of them can explain away the data, a prediction error is fed forward to the next higher level to update the overhypothesis. Now imagine that a dysfunctional mechanism generates aberrant error signals, leading to your thoughts being persistently coded as unpredicted or surprising. This situation of impoverished accuracy may license discarding of your initial overhypothesis by means of negating its content, 'I am not the unique thinker of my thoughts'. (Note that the same considerations raised in Section 4.5. apply here again: discarding a belief can, but need not, and often does not, effect suspended judgment). This in turn would set up a new hypothesis space, one which allows delusional candidates as members ('somebody else is the thinker of my thoughts', 'thoughts are being inserted into my mind', etc.).

Another explanation I suggest for why a system produces delusional candidate hypotheses appeals to premorbid developmental abnormalities. The thought is that delusional candidates arise because of the way overhypotheses (i.e., higher-level priors) are acquired developmentally. Overhypotheses are learned inter alia through sensorimotor interactions with the environment, which are abnormal across development in preschizophrenia children (Walker et al., 1994). This can culminate in a willingness to accept as plausible hypotheses which most of us agree are implausible (Moritz and Woodward, 2004).

Evidence of sensorimotor abnormalities in children with preschizophrenia traits has been documented in an ingenious study by Elaine Walker and Richard Lewine (1990). They

recruited parents of schizophrenia patients to share home movies featuring their children during childhood. Then they asked child development specialists to rate the videotapes for neurological soft signs (NSS), which comprise subtle but observable deficits in sensory integration, motor coordination, and sequencing of complex movements. The raters had no previous exposure to the videotapes and were unaware of the subjects' diagnosis. Group comparisons revealed a higher rate of NSSs in preschizophrenia children when compared to healthy siblings, ranging from poor balance and clumsiness to impaired sequential motor performance.

Sensorimotor dysfunctions during childhood could contribute to an anomalous sense of agency, with bodily movements being experienced as non-volitional, uncoordinated, and fragmented (McGhie and Chapman, 1961). A set of overhypotheses that accommodates such imprecision from early in life would necessarily rate a broader range of candidate hypotheses as more likely than do neurotypical subjects. Also, presumably, the disorientation one would feel would garner hypotheses that would not otherwise arise. Imagine a person who grew up experiencing a lack of control over her body as the norm. It is not hard to see how such a person could form the overhypothesis that she is not the only source of her bodily movements. From there it is but a very small step to the delusional hypothesis of an external agency being the cause of movement.

There is a third possibility I wish to consider: implausible hypotheses are learned through cultural social interactions before the onset of any neurocognitive impairment. An example will illustrate this. Alien-abduction stories involve 'subjectively real memories' of being kidnapped by aliens and subjected to physical and psychological experimentation. The commonality of the experience across abductees (i.e., gray beings with large black eyes) is taken by some believers as evidence of the veracity of abductees' claims (Clancy, 2007). One scientific explanation for abduction beliefs involves an aberrant experience; specifically, sleep paralysis accompanied by hypnopompic hallucinations (McNally and Clancy, 2005). That ineffable and terrifying experience is explained with an abduction narrative. But why the commonality in the accounts? It is widely recognised that memory recall involves reconstructive processes which are strongly influenced by one's current beliefs and expectations (Bower, 1990; Loftus, 1979). The earliest publicised abduction story was told by a married couple, Betty and Barney Hill, in 1964. Twelve days before the 'abduction' an episode of the TV show *The Outer Limits* aired, depicting aliens that looked just like those the

Hills and countless others have described.[30] It is plausible that Betty and Barney had seen the show, and that many of the other claimants to abduction had too. I wonder whether at least some delusional hypotheses may be learned similarly, by assimilating culturally available ideas (Gold and Gold, 2014). For instance, the movie *Face/Off* visually conveys the possibility of two individuals switching faces via plastic surgery and still be convincing as one another. It may be that merely being exposed to such a possibility could raise an agent's credence for that possibility (perhaps slightly but enough for it to enter the agent's hypothesis space).

To recap briefly, we have considered three ways in which delusional hypotheses might arise in the first place: (1) they could take over where current hypotheses leave puzzling experiences unexplained; (2) they could be the end products of developmental processes gone awry; (3) they could be acquired through ordinary processes of sociocultural learning. Of course, there may be other possible explanations for hypothesis generation in cases of delusions. My survey is meant to be merely suggestive, not exhaustive. What is clear, though, is that empirical priors do help us understand the origin of delusional hypothesis. Indeed, for each case considered above, those hypotheses are learned empirically (on the basis of introspective or interactive experiences).

4.8. Conclusion

Matthew Parrott (2019) has raised two concerns about the PP framework for explaining delusions: first, it fails to explain why someone believes something delusional rather than suspend judgment or believe something different, and second, it does not capture how delusional hypotheses are generated to begin with. My aim in this paper has been to address these concerns. I drew attention to factors that might prevent or invalidate suspension. I highlighted that epistemically better alternatives than delusional hypotheses may be unavailable to the subject in some important respects. Finally, I considered three ways in which delusional hypotheses may be generated which are in keeping with the PP framework and explainable in terms of empirical priors. If my analysis is correct, then the Parrott's concerns do not convincingly undermine the explanatory power of the PP framework.

[30] Betty's narrative also has a lot in common with the storyline of *Invaders from Mars* (1953).

4.9. References

Adams, R. A., Brown, H. R. and Friston, K. J. (2015). 'Bayesian Inference, Predictive Coding and Delusions'. *Avant*, 5, pp. 51–88.

Austerweil, J. L. and Griffiths, T. L. (2009). 'Analyzing Human Feature Learning as Nonparametric Bayesian Inference'. In D. Koller, Y. Bengio, D. Shuurmans and L. Bottou (eds.) *Advances in Neural Information Processing Systems,* Vol. 21. (pp. 97–104). Cambridge, MA: MIT Press.

Austerweil, Joseph L., Gershman, S. J. and Griffiths, T. L. (2015). 'Structure and Flexibility in Bayesian Models of Cognition'. In J.R. Busemeyer, Z. Wang, J.T. Townsend and A. Eidels (eds.) *The Oxford Handbook of Computational and Mathematical Psychology* (pp. 187–208). New York: Oxford University Press.

Bertelson, P. (1998). 'Starting from the Ventriloquist: The Perception of Multimodal Events'. In M. Sabourin, F. Craik, and M. Robert (eds.) *Advances in Psychological Science, Vol. 2: Biological and cognitive aspects* (pp. 419–439). Hove, England: Psychology Press.

Berti, A., Làdavas, E., Stracciari, A., Giannarelli, C. and Ossola, A. (1998). 'Anosognosia for Motor Impairment and Dissociations with Patients Evaluation of the Disorder: Theoretical Considerations'. *Cognitive Neuropsychiatry*, 3, pp. 21–44.

Bisiach, E., Rusconi, M. L. and Vallar, G. (1991). 'Remission of Somatoparaphrenic Delusion through estibular Stimulation'. *Neuropsychologia*, 29, pp. 1029–1031.

Bortolotti, L. (2016). Epistemic Benefits of Elaborated and Systematized Delusions in Schizophrenia. *British Journal for the Philosophy of Science*, 67, pp. 879–900.

Bower, G. H. (1990). 'Awareness, the Unconscious, and Repression: An Experimental Psychologist's perspective'. In J. A. Singer (ed.) *Defence Mechanism and Personality Style* (pp. 209–231). Chicago: University of Chicago Press.

Brighetti, G., Bonifacci, P., Borlimi, R. and Ottaviani, C. (2007). "Far from the Heart Far from the Eye': Evidence from the Capgras Delusion'. *Cognitive Neuropsychiatry*, 12, pp. 189–197.

Butler, P. V. (2000). 'Reverse Othello Syndrome Subsequent to Traumatic Brain Injury'. *Psychiatry*, 63(1), pp. 85–92.

Cappa, S., Sterzi, R., Vallar, G. and Bisiach, E. (1987). 'Remission of Hemineglect and Anosognosia during Vestibular Stimulation'. *Neuropsychologia*, 25, pp. 775–782.

Chisholm, R. M. (1976). *Person and Object*. La Salle: Open Court.

Clancy, S. A. (2007). *Abducted: How People Come To Believe They Were Kidnapped By Aliens*. Cambridge, MA: Harvard University Press.

Clark, A. (2012). 'Dreaming the Whole Cat: Generative Models, Predictive Processing, and the Enactivist Conception of Perceptual Experience'. *Mind*, 121(483), pp. 753–771.

Clark, A. (2013). 'Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science'. *Behavioral and Brain Sciences*, 36, pp. 181–204.

Clark, A. (2016). *Surfing Uncertainty*. Oxford: Oxford University Press.

Colbert, S. M. and Peters, E. R. (2002). 'Need for Closure and Jumping-to-Conclusions in Delusion-Prone Individuals'. *Journal of Nervous and Mental Disease,* 190, pp. 27–31

Coltheart, M., Langdon, R. and McKay, R. (2007). 'Schizophrenia and Monothematic Delusions'. *Schizophrenia Bulletin*, 33, pp. 642–647,.

Coltheart, M. (2005). 'Delusional Belief'. *Australian Journal of Psychology*, 57, pp. 72–76.

Coltheart, M. (2007). 'The 33rd Sir Frederick Bartlett Lecture Cognitive Neuropsychiatry and Delusional Belief'. *Quarterly Journal of Experimental Psychology*, 60, pp. 1041–1062.

Coltheart, M., Menzies, P. and Sutton, J. (2010). 'Abductive Inference and Delusional Belief'. *Cognitive Neuropsychiatry*, 15, pp. 261–287.

Corlett, P. R. (2018). 'Delusions and Prediction Error'. In L. Bortolotti (ed.) *Delusions in Context*. (pp. 35–66). Cham: Springer International Publishing.

Corlett, P. R., Taylor, J. R., Wang, X.-J., Fletcher, P. C. and Krystal, J. H. (2010). 'Toward a Neurobiology of Delusions'. *Progress in Neurobiology*, 92, pp. 345–369.

Corlett, P. R., Honey, G. D. and Fletcher, P. C. (2016). 'Prediction Error, Ketamine and Psychosis: An Updated Model'. *Journal of Psychopharmacology*, 30, pp. 1145–1155.

Corlett, P. R. and Fletcher, P. C. (2015). 'Delusions and Prediction Error: Clarifying the Roles of ehavioural and Brain Responses'. *Cognitive Neuropsychiatry*, 20, pp. 95–105.

Davies, M., Coltheart, M., Langdon, R. and Breen, N. (2001) 'Monothematic Delusions: Towards a Two-Factor Account'. *Philosophy, Psychiatry, and Psychology*, 8, pp. 133–158.

Davies, M., Aimola Davies, A. and Coltheart, M. (2005). 'Anosognosia and the Two-factor Theory of Delusions'. *Mind and Language*, 20, pp. 209–236.

Davies, M., McGill, C. and Aimola Davies, A. (forthcoming). 'Anosognosia for Motor Impairments as a Delusion: Anomalies of Experience and Belief Evaluation'. In A. Mishara, P. Corlett, P. Fletcher, A. Kranjec and M. Schwartz (eds.) *Phenomenological Neuropsychiatry: How Patient Experience Bridges Clinic with Clinical Neuroscience*. New York: Springer.

Ellis, H. D., Lewis, M. B., Moselhy, H. F. and Young, A. W. (2000). 'Automatic without Autonomic Responses to Familiar Faces: Differential Components of Covert Face Recognition in a Case of Capgras Delusion'. *Cognitive Neuropsychiatry*, 5, pp. 255–269.

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. Series B: Biological Sciences*, 264, pp. 1085–1092.

Feldman, H. and Friston, K. J. (2010). 'Attention, Uncertainty, and Free-Energy'. *Frontiers in Human Neuroscience*, 4, pp. 1–23.

Fine, C., Gold, I. and Craigie, J. (2005). 'Damned if You Do; Damned if You Don't: The Impasse in Cognitive Accounts of the Capgras Delusion'. *Philosophy, Psychiatry, and Psychology*, 12, pp. 143–151.

Fletcher, P. C. and Frith, C. D. (2009). 'Perceiving is Believing: A Bayesian Approach to Explaining the Positive Symptoms of Schizophrenia'. *Nature Reviews Neuroscience*, 10, pp. 48–58.

Fotopoulou, A. (2012a). 'Illusions and Delusions in Anosognosia for Hemiplegia: From Motor Predictions to Prior Beliefs'. *Brain*, 135, pp. 1344–1346.

Fotopoulou, A. (2012b). 'Towards a Psychodynamic Neuroscience'. In A. Fotopoulou, D. Pfaff and M. A. Conway (eds.) *From the Couch to the Lab: Trends in Psychodynamic Neuroscience* (pp. 25–46). Oxford: Oxford University Press.

Fotopoulou, A. (2014). 'Time to Get Rid of the 'Modular' in Neuropsychology: A Unified Theory of Anosognosia as Aberrant Predictive Coding'. *Journal of Neuropsychology*, 8, pp. 1–19.

Fotopoulou, A., Tsakiris, M., Haggard, P., Vagopoulou, A., Rudd, A. and Kopelman, M. (2010). 'Implicit Awareness in Anosognosia for emiplegia: Unconscious Interference Without Conscious Re-Representation'. *Brain*, 133, pp. 3564–3577.

Fotopoulou, A., Tsakiris, M., Haggard, P., Vagopoulou, A., Rudd, A. and Kopelman, M. (2008). 'The Role of Motor Intention in Motor Awareness: An Experimental Study on Anosognosia for Hemiplegia'. *Brain*, 131, pp. 3432–3442.

Freeman, D., Garety, P. A., Kuipers, E., Bebbington, P. E., Fowler, D. and Dunn, G. (2004). 'Why Do People With Delusions Fail to Choose More Realistic Explanations for their Experiences? An Empirical Investigation'. *Journal of Consulting and Clinical Psychology*, 72, pp. 671–680.

Friedman, J. (2013). 'Suspended Judgment'. *Philosophical Studies*, 162, pp. 165–181.

Friston, K. J. (2005).'A Theory of Cortical Responses'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360, pp. 815–836.

Friston, K. J. (2008). 'Hierarchical Models in the Brain'. *PLoS Computational Biology*, 4, pp. 1–24.

Friston, K. J. (2012). 'Prediction, Perception and Agency'. *International Journal of Psychophysiology*, 83, pp. 248–252.

Friston, K. J., Litvak, V., Oswal, A., Razi, A., Stephan, K. E., Van Wijk, B. C. M. and Zeidman, P. (2016). 'Bayesian Model Reduction and Empirical Bayes for Group (DCM) Studies'. *NeuroImage*, 128, pp. 413–431.

Frith, C. D., Blakemore, S. J. and Wolpert, D. M. (2000). 'Abnormalities in the Awareness and Control of Action'. *Philosophical Transactions of the Royal Society of London, Series B. Biological Sciences*, 355, pp. 1771–1788.

Frith, C. and Friston, K. (2013). 'False Perceptions and False Beliefs: Understanding Schizophrenia'. *Neuroscience and the Human Person: New Perspectives on Human Activities*, pp. 1–15.

Garbarini, F., Rabuffetti, M., Piedimonte, A., Pia, L., Ferrarin, M., Frassinetti, F. and Berti, A. (2012). 'Moving a Paralysed Hand: Bimanual Coupling Effect in Patients with Anosognosia for Hemiplegia'. *Brain*, 135, pp. 1486–1497.

Gilbert, D. T. (1991). 'How Mental Systems Believe'. *American Psychologist*, 46, pp. 107–119.

Gilbert, D. T., Krull, D. S. and Malone, P. S. (1990). 'Unbelieving the Unbelievable: Some Problems in the Rejection of False Information'. *Journal of Personality and Social Psychology*, 59, pp. 601–613.

Gold, J. and Gold, I. (2014). *Suspicious Minds: How Culture Shapes Madness*. New York: Free Press.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Hirstein, W. and Ramachandran, V. S. (1997). 'Capgras Syndrome: a Novel Probe for Understanding the Neural Representation of the Identity and Familiarity of Persons', *Proceedings of the Royal Society of London, Series B: Biological Sciences*, 264, pp. 437–444.

Hohwy, J. (2012). 'Attention and Conscious Perception in the Hypothesis Testing Brain'. *Frontiers in Psychology*, 3, pp. 1–14.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Hohwy, J. (2015). 'Prediction Error Minimization, Mental and Developmental Disorder, and Statistical Theories of Consciousness'. In R. Gennaro (ed.) *Disturbed Consciousness: New Essays on Psychopathology and Theories of Consciousness* (pp. 29–54). Cambridge, MA: MIT Press.

Hohwy, J. (2017). 'Priors in Perception: Top-Down Modulation, Bayesian Perceptual Learning Rate, and Prediction Error Minimization'. *Consciousness and Cognition*, 47, pp. 75–85.

Hohwy, J. and Michael, J. (2018). 'Why Should Any Body Have a Self?'. In F. de Vignemont and A. J. T. Alsmith (eds.) *The Subject's Matter*: Self-Consciousness and the Body (pp. 363–391). Cambridge, MA: MIT Press.

Hohwy, J. and Rosenberg, R. (2005). 'Unusual Experiences, Reality Testing and Delusions of Alien Control'. *Mind and Language, 20*(2), pp. 141–162.

Kemp, C., Perfors, A. and Tenenbaum, J. B. (2007). 'Learning Overhypotheses with Hierarchical Bayesian Models'. *Developmental Science*, 10, pp. 307–321.

Kiefer, A. and Hohwy, J. (2018). 'Content and Misrepresentation in Hierarchical Generative Models'. *Synthese*, 195, pp. 2387–2415.

Knill, D. C. and Pouget, A. (2004). 'The Bayesian Brain: the Role of Uncertainty in Neural Coding and Computation'. *TRENDS in Neuroscience*, 27, pp. 712–719.

Kruglanski, A. W. (1989). *Lay Epistemics and Human Knowledge*. Boston: Springer US.

Kuzmanovic, B. and Rigoux, L. (2017). 'Valence-Dependent Belief Updating: Computational Validation'. *Frontiers in psychology*, 8, pp. 1–11.

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

Loftus, E. F. (1979). *Eyewitness Testimony*. Cambridge, MA: Harvard University Press.

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

Mandelbaum, E. (2014). 'Thinking is Believing'. *Inquiry*, 57, pp. 55–96.

McGhie, A. and Chapman, J. (1961). 'Disorders of Attention and Perception in Early Schizophrenia'. *British Journal of Medical Psychology*, 34, pp. 103–116.

McKay, R., Langdon, R. and Coltheart, M. (2007). 'Jumping to Delusions? Paranoia, Probabilistic Reasoning, and Need for Closure'. *Cognitive Neuropsychiatry*, 12(4), pp. 362–376.

McKay, R., Tamagni, C., Palla, A., Krummenacher, P., Hegemann, S. A. A., Straumann, D. and Brugger, P. (2013). 'Vestibular Stimulation Attenuates Unrealistic Optmisim'. *Cortex*, 49(8), pp. 2272–2275.

McNally, R. J. and Clancy, S. A. (2005). 'Sleep Paralysis, Sexual Abuse, and Space Alien Abduction'. *Transcultural Psychiatry*, 42, pp. 113–122

Miyazono, K. (2018). *Delusions and Beliefs A philosophical Inquiry*, London: Routledge.

Miyazono, K. and McKay, R. (2019). 'Explaining Delusional Beliefs: a Hybrid Model'. *Cognitive Neuropsychiatr*y, 24(5), pp. 335–346.

Moritz, S. and Woodward, T. S. (2004). 'Plausibility Judgement in Schizophrenic Patients: Evidence for a Liberal Acceptance Bias'. *German Journal of Psychiatry*, 7, pp. 66–74.

Parrott, M. (2017). 'Subjective Misidentification and Thought Insertion'. *Mind and Language*, 32, pp. 39–64.

Parrott, M. (2019). 'Delusional Predictions and Explanations'. *The British Journal for the Philosophy of Science*, 0, pp. 1–32.

Perfors, A. (2012). 'Bayesian Models of Cognition: What's Built in After All?'. *Philosophy Compass*, 7, pp. 127–138.

Peters, E. R. and Garety, P. A. (1996). 'The Peters et al. Delusions Inventory (PDI): New Norms for the 21-Item Version'. *Schizophrenia Research*, 18, pp. 118–119.

Ramachandran, V. S. (1995). 'Anosognosia in Parietal Lobe Syndrome'. *Consciousness and Cognition*, 4, pp. 22–51.

Ramachandran, V. S. and Blakeslee, S. (1998). *Phantoms in the Brain: Human Nature and the Architecture of the Mind*. New York: William Morrow and Co.

Rao, R. P. N. and Ballard, D. H. (1999). 'Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-Classical Receptive-Field Effects'. *Nature Neuroscience*, 2, pp. 79–87.

Roberts, G. (1992). 'Origins of Delusion'. *British Journal of Psychiatry*, 161, pp. 298–308.

Sharot, T. and Garrett, N. (2016). 'Forming Beliefs: Why Valence Matters'. *Trends in Cognitive Sciences*, 20(1), pp. 25–33.

Sharot, T., Kanai, R., Marston, D., Korn, C. W., Rees, G. and Dolan, R. J. (2012). 'Selectively Altering Belief Formation in the Human Brain'. *Proceedings of the National Academy of Sciences of the United States of America*, 109(42), pp. 17058–17062.

Sharot, T., Korn, C. W., and Dolan, R. J. (2011). 'How Unrealistic Optimism is Maintained in the Face of Reality'. *Nature Neuroscience*, 14(11), pp. 1475–1479.

Sturgeon, S. (2010). 'Confidence and Coarse-Grained Attitudes'. In T. Gendler and J. Hawthorne (eds.) *Oxford Studies in Epistemology*. Oxford: Oxford University Press.

Sullivan-Bissett, E. (2015). 'Implicit Bias, Confabulation, and Epistemic Innocence'. *Consciousness and Cognition*, 33, pp. 548–560.

Sullivan-Bissett, E. (2018). 'Monothematic Delusion: A Case of Innocence from Experience'. *Philosophical Psychology*, 31, pp. 920–947.

Sullivan-Bissett, E., Bortolotti, L., Broome, M. and Mameli, M. (2017). 'Moral and Legal Implications of the Continuity between Delusional and Non-Delusional Beliefs'. In G. Keil, L. Keuck and Rico Hauswald (eds.) *Vagueness in Psychiatry* (pp. 191–210). Oxford: Oxford University Press.

Stone, T. and Young, A. W. (2007). 'Delusions and Brain Injury: The Philosophy and Psychology of Belief'. *Mind and Language*, 12(3–4), pp. 327–364.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. and Goodman, N. D. (2011). 'How to Grow a Mind: Statistics, Structure, and Abstraction'. *Science*, 331, pp. 1279–1285.

Tranel, D., Damasio, H. and Damasio, A. R. (1995). 'Double Dissociation between Overt and Covert Face Recognition'. *Journal of Cognitive Neuroscience*, 7, pp. 425–432.

Turnbull, O. H., Fotopoulou, A. and Solms, M. (2014). 'Anosognosia as Motivated nawareness: The 'Defence' Hypothesis Revisited'. *Cortex*, 61, pp. 18–29.

Walker, E. F., Savole, T. and Davis, D. (1994). 'Neuromotor Precursors of Schizophrenia'. *Schizophrenia Bulletin*, 20, pp. 441–451.

Walker, E. and Lewine, R. J. (1990). 'Prediction of Adult-Onset Schizophrenia from Childhood Home Movies of the Patients'. *American Journal of Psychiatry*, 147, pp. 1052–1056.

Wegner, D. M., Coulton, G. F. and Wenzlaff, R. (1985). 'The Transparency of Denial: Briefing in the Debriefing Paradigm'. *Journal of Personality and Social Psychology*, 49, pp. 338–346.

Chapter 5: Spinozan Doxasticism About Delusions

5.0. Abstract

The Spinozan theory of belief fixation holds that mentally representing truth-apt propositions leads to immediately believing them. This theory has some striking implications for the nature of beliefs: they are acquired arationally, they are introspectively opaque, they reside in different fragments of the mind. My aim in this paper is to explore the theory's significance relative to the doxastic status of delusions (their status as beliefs). Anti-doxastic arguments point to the fact that delusions fail to meet standards of rationality which we expect beliefs to conform to; for example, they are often unresponsive to evidence, badly integrated into the psychology of the agent, and weakly behaviour guiding. A common counterargument is that failures of rationality do not preclude the ascription of beliefs, because we frequently ascribe beliefs which also fail to satisfy standards of rationality. It may, however, be questioned whether the debate on the status of delusions can be settled in this way, relying on ordinary patterns of belief ascription in folk psychology. Indeed, it may be thought that a fully satisfactory defence of doxasticism would have to explain why delusions (qua beliefs) exhibit the failures of rationality they do. I argue that the Spinozan theory is especially well-suited to provide such an explanation.

5.1. Introduction

The notion that people can entertain a proposition without believing it has historically had and continues to have widespread intuitive appeal. It is intuitively natural to think that when presented with a proposition whose truth-values are unknown, we have the ability to *merely* entertain it.[31] There are, however, some dissenters from this view. For some scholars (Gilbert, 1991; Gilbert et al., 1993; Huebner, 2009; Levy and Mandelbaum, 2014; Mandelbaum, 2010, 2014; Mandelbaum and Quilty-Dunn, 2015), we automatically believe every truth-apt proposition which we entertain and only then perhaps take measures to revise our initial belief. I will call this view (following Gilbert, 1991) the *Spinozan* theory of belief fixation, to be contrasted with what I will call the *Cartesian* theory of belief fixation, the view

[31] For current purposes, to 'entertain' a proposition is to have access to it, though not necessarily consciously so. Instances of 'entertaining that p' are 'perceiving that *p*', 'considering that *p*', 'supposing that *p*', 'imagining that *p*', 'accepting that *p* for the sake of argument', etc.

that we first entertain a proposition, then subsequently believe it, disbelieve it, or suspend judgment about it.[32]

According to the Spinozan theory, if one entertains the proposition 'clouds are made of cotton candy', one thus believes that 'clouds are made of cotton candy'. The theory has been defended at length first by Dan Gilbert (1991) and more recently by Eric Mandelbaum (2010, 2014). I should note at the outset that my aim here is not to argue that the theory is true, although I will examine the motivations for employing it. Instead, my question is: assuming that the Spinozan theory is true, what consequences may there be for our understanding of delusions?

The term 'delusion' refers to a clinical symptom observed across a range of psychiatric disorders, including schizophrenia, dementia, schizoaffective disorder, bipolar disorder, and major depression. Delusions can take various forms (Coltheart, 2013). Monothematic delusions concern a single topic. Much philosophical discussion has focused on *Capgras*, which presents as the conviction that a loved one has been replaced by an imposter, and *Cotard*, in which the person is convinced that she is dead, that she is missing internal organs, or even that her body is rotting away. Many individuals with schizophrenia experience polythematic delusions, which are not restricted to a single topic but encompass a wide variety of subjects.

There has been much debate about whether delusions are beliefs. The position according to which delusions are beliefs is known as 'doxasticism about delusions' (Bortolotti, 2010, 2012; Bayne and Pacherie, 2005). The most common versions of anti-doxastic arguments rely on interpretationist assumptions about what is required for belief ascription. According to interpretationism, we ascribe beliefs in order to make sense of what other people do (Davidson, 1985; Dennett, 1987). Those who take this line posit a tight connection between being rational and being ascribed beliefs. At least three standards of rationality are taken to be essential to belief ascription (Bortolotti, 2010). Minimally, beliefs ought to be well supported by evidence (epistemic rationality), in line with other beliefs (procedural rationality), and action guiding in the appropriate circumstances (agential rationality). Suppose I tell you that Paula dislikes wine and my only evidence is that I once saw her drink a glass of punch. Suppose that later in the day I tell you that she is going to a

[32] If you are to take a doxastic attitude toward *p*, three alternatives are possible: believing that *p* is true (believing *p*), believing that *p* is false (disbelieving *p*), or refraining from either believing or disbelieving *p* (suspending judgment about *p*). To keep things simple, I will henceforth speak only in terms of belief and disbelief. For more on suspended judgment, see Friedman (2013).

wine-tasting class and that I have bought three bottles of wine for her dinner party tonight. These deviations from rationality raise questions about whether I really believe that Paula dislikes wine. Delusions are often like my claim about Paula: they are unsupported by evidence, they are badly integrated with other beliefs, and they fail to manifest in action. This has led some philosophers to conclude that delusions are not beliefs, but nondoxastic, or partially doxastic, states.[33]

In her defence of doxasticism, Lisa Bortolotti (2010, 2012) rejects the idea that rationality *per se* is individuative of beliefs. She justifies this by saying that if we deny belief status to delusions on grounds of irrationality, then we must conclude the same of many (apparent) beliefs which also do not live up to standards of rationality. On the positive side, she argues that the irrationality of delusions is no obstacle to their being classified as beliefs. Call this the 'standard' defence of doxasticism. Bortolotti is only concerned with the notion of belief that underlies our ordinary mentalistic ascriptions, and she makes no commitment to what beliefs are outside the folk-psychological discourse. It may be wondered, however, whether the doxastic status of delusions can be genuinely secured within the explanatory framework of folk psychology. One might want more from a defence of doxasticism. One might, for example, want an analysis that helps explain why delusions, qua beliefs, fail to satisfy norms of rationality. This would require, at a minimum, a view of cognitive architecture that settles how an agent fixates her beliefs. If it could be shown that the failures of rationality found in delusions result naturally from belief-fixating processes, then the argument for doxasticism would be stronger.

The paper unfolds as follows. I first outline the Spinozan theory in more detail, with particular attention to the way it characterises belief. I then consider the implications of applying this theory to delusions. This generates a new version of doxasticism, which has two advantages over the standard defence: (i) it puts pressure on the very notion that one can be deluded that *p* without believing that *p*, (ii) it accommodates the deviations from rationality attributed to delusion within a plausible theory of belief fixation. Finally, I address whether delusions really fit the Spinozan characterisation of beliefs, and why this is something all doxasticists should want to accept.[34]

---

[33] Examples are imaginings that one misinterprets as beliefs (Currie and Jureidini, 2001; Currie and Ravenscroft, 2002), perceptual illusions , empty speech acts (Berrios, 1991), in-between states part belief-like, part not belief-like (Schwitzgebel, 2012), 'bimaginings' part belief-like, part imagining-like (Egan, 2008a), acceptances (Frankish, 2013; Dub, 2017), or thoughts reporting the contents of default processes (Gerrans, 2014).

[34] A cautionary note about terminology is due before setting off. The Spinozan theory works with a graded notion of belief, which allows that one can believe things with varying degrees of confidence or credence

5.2. I Think, therefore I Believe

5.2.1. Cartesian versus Spinozan Conception of Belief

In our ordinary thinking about ourselves, it seems obvious to us that we are able to evaluate propositions as candidates for belief, before believing them. This simple notion is the centrepiece of the Cartesian view and has been standard in both philosophy and cognitive science (Quine, 1960; Dennett, 1987; Fodor, 1983; Pylyshyn, 1989). The Cartesian view, when fleshed out more fully, is defined by three following claims (Mandelbaum, 2014). First, agents are able to assess the truth-value of a proposition they entertain before believing it. In this sense belief fixation consists in a sequential process with multiple stages, one involving the entertaining of the proposition, and the other involving going ahead and believing or disbelieving it. Second, believing and rejecting a proposition are alternative outcomes of a single mental process (Gilbert, 1991), and thus should show the same pattern of effects in situations where cognitive resources are depleted. Third, both believing and rejecting a proposition involve effortful mental activity. [35] That is, in order for a proposition to be believed or rejected, a person needs to have sufficient cognitive capacity to evaluate that proposition.

The Spinozan takes an alternative view, arguing for the temporal precedence of belief fixation over rejection, and the unity of entertaining some proposition and believing it (Gilbert, 1991). This view can be stated in three broad claims as follows (Mandelbaum, 2014). First, agents are unable to entertain propositions without believing them. In other words, they have no choice but to believe whatever proposition they happen to entertain. Second, believing a proposition and rejecting it as false are functionally different cognitive operations, which suggests that they should be affected differently when people's cognitive resources are limited or depleted. Third, forming a belief is an automatic and effortless process. By contrast, replacing the attitude of belief with that of disbelief is an effortful mental task,

(Mandelbaum, 2014). One's belief that *p* is assigned a value in the interval [0,1] that represents one's degree of confidence in that proposition–the higher the number, the greater the degree of confidence. This contrasts with the standard notion of belief as a binary, all-or-none state, where one either believes or does not believe that *p*. On the Spinozan theory to say that one believes that *p* is to say that one has high enough credence in *p*. How high is high enough? Mandelbaum (2014) contends that the credence is raised to a level that would guide action and that would allow *p* to serve as a premise in inferences–presumably a level much higher than 0.0001 (as that would be behaviourally inert) and yet it need not be extremely high (one that equals or exceeds 0.9).

[35] 'Effortful' here is used in the narrow sense of 'cognitively demanding', covering any mental task that consumes substantial cognitive resources. In this narrow sense it is possible for a mental task to be effortful and yet occur outside of conscious awareness and without conscious intent (Mandelbaum, 2010, 2014).

which can only be carried out once the belief has been acquired. This is why belief fixation can occur effortlessly under cognitive load, while the task of rejecting previously held beliefs is made more difficult by increased cognitive load.

5.2.2. Motivation for the Spinozan Theory

The Cartesian and Spinozan theories make different predictions about the ways in which the belief-fixating process should break down in the wake of resource depletion (Mandelbaum, 2014; Gilbert, 1991). The Cartesian theory predicts that depleting cognitive resources should disable the system from deciding about the truth of the propositions it entertains. Since cognitive load depletes cognitive resources, propositions should neither be believed nor rejected when individuals are placed under cognitive load. This is because believing and rejecting a proposition are products of one and the same mental process, and so should be affected by cognitive load to an equal extent (Mandelbaum, 2014).

Conversely, if the Spinozan theory were correct, resource depletion should prevent the system from rejecting a proposition, but not from believing a proposition. The Spinozan theory treats believing as a reflex. Because cognitive load interferes only with effortful and deliberative processing, if believing truly is reflexive, it should remain unaffected by concurrent load. However, since rejection is effortful, the prediction is that load should affect rejecting a proposition (Mandelbaum, 2014).

In a series of papers Mandelbaum (2010, 2014) and Gilbert (Gilbert, 1991; Gilbert et al., 1990; Gilbert et al., 1993) cite a number of empirical facts that they claim are best accommodated by a Spinozan theory of belief fixation. I do not have the space to cover these data in their entirety. Instead, I will focus on two strands of evidence suggesting that the Spinozan theory is a more plausible account of belief than is the Cartesian.

One strand highlights a difference between the way people remember truths and the way they remember falsehoods. In an experiment by Gilbert, Krull, and Malone (1990), participants were instructed to learn a sequence of fictitious Hopi Indian word definitions. On a computer screen, they were sequentially presented propositions of the form 'An X is a Y', where 'X' was a nonsense Hopi noun and 'Y' was a common noun in English (e.g., *A twyrin is a doctor*). Right after presenting a proposition, the screen flashed up either the word *true* or the word *false* to indicate whether the statement in question was veridical or not. The participants' cognitive resources were depleted, and the processing of the proposition temporarily dismantled, by having them to press a button whenever a particular sound was

heard. On 4 of the 12 trials participants were interrupted by a tone sounding 750 ms after the appearance of the assessment word. At the end of session, the 4 statements were rephrased as questions (e.g., *Is a twyrin a doctor?*), to which participants were asked to answer by pressing one of four buttons on a response box, labelled respectively *true*, *false*, *no information*, and *never seen*.

If the Cartesian view were right, the interruption should hurt not only the participants' ability to accurately remember the false propositions, but also the correct identification of true propositions. A Cartesian learner, while being able to entertain each proposition, should not have had the time to assess its credibility and, thus, to either believe it or reject it. By contrast, a Spinozan learner should have believed every proposition that was entertained and the failed to reject those which they were told to be false. The prediction here is that the interruption should shut down rejection but leave the initial belief-assenting stage intact, because only the former is an active process that can be blocked by cognitive load. Therefore, if the Spinozan view were right, we would expect that, by and large, participants in interrupted trials would correctly remember true propositions as true, but incorrectly remember false propositions as true. Results supported this prediction: the added cognitive load did not affect the ability of the participants to answer correctly on true propositions (55% of the time when uninterrupted versus 58% when interrupted), but it did affect their performance in identifying false propositions (55% of the time when uninterrupted versus 35% when interrupted).

A second line of evidence for the Spinozan view comes from well-known experiments on belief perseverance that induce participants to retain false beliefs in the face of disconfirming evidence. In one experiment, Lee Ross and his colleagues (1975) gave participants 25 pairs of suicide notes and instructions that one note in each pair was real, the other a fake. Participants were then asked to sort real suicide notes from fakes, and told that they had done either well or poorly. Next, participants were debriefed and shown evidence that the feedback they had received was bogus, it being randomly determined by the experiments regardless of their actual performance. After being debriefed, participants were asked to estimate their own competence at the task (i.e., whether their overall success rate was better or worse than average). Despite knowing that the feedback was bogus, participants who had earlier received positive feedback believed their performance to be above average, and those who had been given negative feedback believed their performance to be below average.

These findings by themselves are not decisive for whether belief fixation is Spinozan. Participants believed information whose fictitiousness was disclosed only after comprehension and continued to believe that information even in the face of debriefing. All this shows is that once formed, the participants' beliefs were very resistant to change. It does not show how participants formed their beliefs in the first place (viz. whether belief was part of comprehension or the outcome of an effortful assessment process). Wegner, Coulton, and Wenzloff (1985) replicated Ross and colleagues' (1975) study, except this time participants received a briefing about the invalidity of the feedback *before* they took part in the experiment. Participants behaved exactly as we would expect a Spinozan system to behave. Even after being forewarned, participants made judgments of performance that conformed to the feedback they were given, acting as though they believed the feedback to be true. They were unable to reject information which they knew did not reflect their actual performance on the task. This result is entirely in accord with the Spinozan position, which states that comprehension invariably includes belief as a part, such that whenever people comprehend information, that information is immediately believed. If participants were Cartesian learners, they should have looked sceptically upon the feedback, forestalling the formation of any belief about their performance. Instead, participants automatically believed the feedback, despite being motivated not to do so.

Additional evidence from a variety of experimental settings appears to be consistent with the Spinozan theory. Studies on persuasion through fiction suggested that people change their real-world beliefs upon engaging with fictional stories (Appel and Richter, 2007; Marsh and Fazio, 2006; Green and Brock, 2002). Other studies found that repeated warnings about false consumer claims fosters the remembering of those claims as true (Skurnik et al., 2005). Still other studies have shown that repetition increases statements' perceived truthfulness, even when those statements contradict well-known facts (Fazio et al., 2019). I note, however, that the status of the evidence for the Spinozan theory has not gone uncontested. Lena Nadarevic and Edgar Erdfelder (2019), for example, performed experiments similar to Gilbert and colleagues' and got very different results (see also Street and Kingstone, 2016).[36]

[36] More studies gave results purportedly inconsistent with the Spinozan theory. Hasson and colleagues (2005) presented their participants with statements that differed in their informativeness. It was found that the cognitive load manipulations selectively impaired their ability to remember statements that were uninformative when false (i.e., this person walks barefoot to work), but had no effect on statements that were informative when false (i.e., this person owns a television). See Mandelbaum (2014, pp. 71–75) for a proposal on how to accommodate these findings within the Spinozan theory. Other authors (Richter et al., 2009) reported that

To repeat, my goal is not to argue for the truth of the Spinozan theory. My goal is more modest and narrow: to evaluate its capacity to improve the prospects of doxasticism about delusions. The evidence surveyed here does not conclusively support the Spinozan theory, but it makes the hypothesis plausible enough that it seems worthwhile to begin considering its implications for whether delusions are beliefs. Before we get to that, however, we need to take stock of the theory's understanding of belief.

5.3. Spinozan Beliefs

5.3.1. Arational Belief Formation

The main thrust of the Spinozan view is that belief fixation operates below the psychological level of explanation, in a way that resists being cast as rational or irrational. Levy and Mandelbaum (2014) make the point through the following example. Imagine a brick falls on your head, causing you to believe that there are only ten planets in the universe. Now of course, the process that led to the belief would be less than epistemically ideal. However, it would be just as strange and misleading to say that your being moved to belief resulted from an irrational inference (or in fact any inference at all). After all, the belief was acquired by means of brute-causal, triggering processes, and not as a result, say, of guessing or wishful thinking. Levy and Mandelbaum (2014) propose to call these types of belief-forming processes 'arational'—neither rational nor irrational—since there are no facts about the agent's prior mental states from which to assess their (ir)rationality (Wedgwood, 2017). Just as it is arational to believe that the universe has ten planets as a result of being hit by a brick, so too it is arational to believe any arbitrary proposition we happen to entertain. In both cases, beliefs are acquired in a brute causal way, such that there is no room even for reasoning to go awry.

cognitive load during feedback processing reduces memory for false statements for which participants had no or weak prior knowledge ('Krypton is a noble gas'), but not for false statements for which they had strong prior knowledge ('Soft soap is edible'). In contrast with this evidence, recently published findings suggest that belief in false statements is increased by repetition independent of prior knowledge, whether it be none, weak or strong (Fazio et al., 2015; Unkelbach and Greifeneder, 2018).

5.3.2. Inconsistency

The Spinozan theory is committed to the view that people's beliefs can be and often are inconsistent.[37] Inconsistent propositions are pervasive in our daily lives, and since we are most of the time under cognitive load, we often lack the capacity to reject all such propositions after believing them (Mandelbaum, 2014). The view that people believe inconsistencies needs some qualifications and some filling in of details. First, the claim does not imply that people who believe inconsistencies will *assert* those inconsistencies (Mandelbaum, 2014, p. 64, fn. 24). Asserting that *p* entails overtly showing one's commitment to *p* and, on this picture, beliefs that *p* carry no such commitment. One can believe *p* and believe ¬ *p* without consciously endorsing *p* and ¬ *p* (for more on this see next subsection). Second, what explains the existence of inconsistent beliefs is the fact our total set of beliefs (henceforth 'belief storage') is 'fragmented' into separate belief systems (Bendana and Mandelbaum, forthcoming). Belief storage is often described by the metaphor of a *web of belief* (Quine and Ullian, 1970). According to this metaphor, all of our beliefs are arranged into a single web-like network, which (i) is consistent and closed under entailment, and which (ii) guides action in all contexts. If our beliefs are connected in a web-like way, altering one belief will have consequences for the entire belief set, such that when an agent changes any of her beliefs, all other beliefs will adjust in the light of that. Some have argued, however, that actual human cognition falls short of this idealization. Rather than being stored in a single coherent database, our beliefs are stored into multiple clusters or fragments which may be inconsistent with one another (Lewis, 1982; Stalnaker, 1984; Egan, 2008b). This means that an agent may believe that *p* in one fragment (f_1) and that ¬ *p* in a separate fragment (f_2) without being prompted to eliminate the conflict between those beliefs. Indeed, if an agent holds inconsistent beliefs in different fragments, it is possible that *p* may be true according to f_1, but false according to f_2. Adjustment would be required only if a belief that *p* and a belief that ¬ *p* were stored in the same fragment, or if they were stored in separate fragments activated at the same time.

A related point is that separate and perhaps inconsistent fragments may be activated in different contexts to produce different patterns of behaviour. This can have various

[37] In a narrow sense, 'inconsistent' does not mean the same as 'contradictory'. If beliefs are inconsistent, then they can't both be true but can both be false (e.g., 'Vicky is my sister', 'Vicky is my daughter'); whereas if beliefs are contradictory, then both can't be true and can't be false (e.g., 'Vicky is my sister', 'Vicky is not my sister'). For the sake of simplicity, however, I will here use 'inconsistent' in a broad sense to also include contradictory beliefs.

consequences. Agents may believe that *p* and that *if p then q*, but fail to act appropriately for *q*, since the beliefs that *p* and that *if p then q* are in separate fragments (Egan, 2008b). It can also happen that certain actions or aspects of behaviour are guided by a fragment that includes *p*, while others are guided by a fragment that includes ¬ *p*. For some Spinozans, this is reflected in cases where there is a trade-off between implicit racism and the antiracist attitudes that one explicitly reports. Consider someone (e.g., a white, liberal law professor) who sincerely professes the equality of all races but whose low-level, nonverbal behaviours are consistently racist (e.g., when a black person gets on the bus, she looks down to avoid eye contact). Assuming fragmentation, one could suggest that this person's racist and antiracists beliefs reside in separate fragments, and that while the non-racist fragment is accessed for overt behaviours (i.e., speech), the racist belief fragment is accessed for spontaneous, unguarded reactions.[38]

5.3.3. Objections and Further Clarifications

Before concluding this subsection, we can look briefly at two prima facie objections that may be pressed against the Spinozan theory (Mandelbaum, 2014).

The first is based on intuition: the intuition is that we can know introspectively what our current attitude is towards a proposition, such as whether we believe it or not. In general, we tend to think of belief as something that is accessible through introspection. If I ask you whether you believe that clouds are made of cotton candy, you would be confident that you do not, and that is because you think you know what you believe. If your intuition is right, then the Spinozan theory is false, for the Spinozan theory excludes that you can entertain a proposition without believing it.

The Spinozan's response to this line of argument is to question whether beliefs are introspectively accessible across the board. According to Mandelbaum (2014), the intuition that they are is a kind of cognitive illusion stemming from the fluency with which information is processed in cases of commonplace or affectively toned belief. In such cases, the belief has been so frequently activated that we need not do any inferential work. If I ask you whether the sky is blue, or whether you love your parents, you can answer immediately because you have close familiarity with the questions. However, argues Mandelbaum (2014),

[38] It goes without saying that this implies a doxastic model of implicit bias, according to which the implicit attitudes driving racist behaviours are unconscious beliefs. This is a minority view among implicit bias theorists (see e.g., Mandelbaum, 2016; Mandelbaum and Quilty-Dunn, 2015; Frankish, 2016; cf. Holroyd, 2016; Levy, 2015; Madva, 2016).

when we turn to beliefs that are not so familiar or strongly value-laden, things look more problematic. If I ask you whether you are prone to altitude sickness, you may answer immediately (perhaps you are a mountaineer), but more likely you will infer what you believe based on recollections of past experiences and behaviours.[39]

Mandelbaum (2014) makes yet another important point, namely that we tend to overstate the extent to which we can introspect beliefs because we too narrowly think of them as mental contents rather than functional states. Beliefs include contents as parts, but they are over and above their contents. They play a functional role in our mental economy. In this role, beliefs operate on contents, but they are not contents themselves (Mandelbaum, 2014, p. 76). For Mandelbaum, functional roles (e.g., a disposition to guide X kinds of behaviours) are not introspectable. Belief contents, on the other hand, are sometimes available to introspection and sometimes not (such as the contents of racist beliefs that cause one to avoid eye contact with black people). Being able to introspect the content of beliefs does not entail being able to introspect the beliefs themselves. Therefore, we do not have introspective access to our beliefs (Mandelbaum, 2014, p. 77).

But what is it to introspect the content of beliefs? Mandelbaum's answer is that we introspect a judgment that *p*, from which we infer that we believe that *p*. So the 'beliefs' that we report having are in fact judgments about beliefs. There are two possible scenarios where judgments become accessible. In the first, we have access to a judgment that accords with the belief content (i.e., we judge that *p* because we introspect the content that *p*). From this we can rightly infer that we believe *p*. In the second, we form a judgment that *p* as a result of deliberation, whereby we examine our past behaviours, simulate what others would do in our situation, register what we have reason to believe, etc. The point is that we go through some reasoning process and come up with a judgment about what we believe (Mandelbaum, 2014).

Either scenario rests on a distinction between judgments or belief reports and beliefs *per se*. The former are personal-level states of a conscious subject. The latter are unconscious propositional attitudes. According to Mandelbaum (2014), we only report beliefs that we consciously *endorse*, these being the ones which we count as justifiable and socially acceptable. What we endorse is determined by host of practical (i.e., not truth-relevant) factors, including avoidance of blame, social anxiety, and shame-avoiding mechanisms. This is why the beliefs

[39] For instance, perhaps you went hiking in the Swiss Alps a month ago, so you will infer that you are not prone to altitude sickness or you would have had troubles breathing in.

that we report may not often match what we actually believe (e.g., it is why the implicit racist only reports her egalitarian beliefs).

A second prima facie objection to the Spinozan view is to ask why we should accept that the states that Spinozans call 'beliefs' are in fact *beliefs* (Mandelbaum, 2014, p. 82). These states lack properties that beliefs are standardly assumed to have (e.g., they are acquired arationally, they are opaque to introspection, we can harbour inconsistent ones, etc.). Mandelbaum's response comes in two steps: the first step is to identify a complex of minimal necessary conditions for something being a belief; the second step is to show that the states under discussion meet these conditions.

A mental state, *M*, qualifies as a belief when at least the following conditions are satisfied: (i) *M* is semantically evaluable (true/false), (ii) M is capable of being acquired through perception, (iii) *M* is capable of interacting with desires in such a way as to cause behaviour, (iv) *M* is 'inferentially promiscuous' (viz. available to be used as a premise in reasoning, see Stich, 1978).

It should be clear that the states discussed in the above experiments are consistent with both (i) and (ii); statements like 'a twyrin is a doctor' are perceptually learned and have truth conditions. To show that these states also meet (iii) and (iv), Mandelbaum cites a further experiment designed to test the Spinozan theory. In such experiment, Gilbert, Tafarodi, and Malone (1993) had participants sit at a computer screen with two lines of text scrolling horizontally across it, one beneath the other. The top line contained two crime reports describing unrelated robberies. Participants were told that the information presented in black font was true information, and that the information presented in red was false information. In one report, the false information exacerbated the gravity of the crime, and in the other it attenuated the gravity of the crime. The bottom line did not contain any words, but only a string of crawling digits. Half of the participants (the uninterrupted participants) were told to disregard any such digit, while the other half (the interrupted participants) were told to press a button whenever the digit 5 appeared. After doing so, participants were asked to play the role of a trial judge and adjudicate a prison term between 0 and 20 years for each perpetrator.

The main finding was that the prison terms recommended by uninterrupted participants were only minimally affected by false information, while false information significantly affected the prison terms recommended by participants in the interrupted condition. Interrupted participants recommended a sentence twice as long when the false

information was aggravating rather than extenuating for the perpetrator. A comparable pattern was seen in the participants' ratings of the perpetrators' likeability, dangerousness, and the likelihood of benefiting from counselling, with interrupted participants being significantly affected by the nature of the false information they had received (Gilbert et al., 1993).

Now for the implications of these findings. Gilbert and colleagues' experiment not only shows that participants in the interrupted condition were significantly more likely to encode false information as true, as predicted by the Spinozan model. It also shows that the attitudes participants formed on the basis of false information served as premises in the reasoning which informed the participants' recommendations for prison terms and their feelings toward the perpetrators. Mandelbaum takes this as evidence that the attitudes in question fit his criteria for belief, on the grounds that they are inferentially promiscuous and active constituents in producing behaviour (Mandelbaum, 2014).

5.4. Spinozan Doxasticism: The Basics

I have now sketched the fundamentals of the Spinozan theory of cognitive architecture. My guiding question in what follows will not be 'is the Spinozan architecture accurate?', but rather, 'what would the consequences of it be for the question of whether delusions are beliefs?' The primary focus will be on how a Spinoza-inspired doxasticism (hereafter simply Spinozan doxasticism) can explain the failures of rationality characteristic of delusions more fully than standard doxasticism, as defended by, for example, Bortolotti (2010, 2012).

If we accept the Spinozan framework, then a simple modus ponens is all we need to conclude that delusions are beliefs.

1. If one entertains the proposition *p*, then one believes *p*.
2. Being deluded that *p* involves entertaining *p*.

Being deluded that *p* involves believing *p*.

This is not as trivial as it may look initially, for as we will see, it can integrate key features of delusion in an explanatory fashion. But let us take things one step at a time.

Consider the Capgras delusion, in which people report that an imposter has replaced a spouse or relative. From the point of view of the Spinozan theory the proposition a person asserts in uttering 'That is not my wife' is automatically believed, simply by being entertained.

It does not matter what propositional attitude the person has. All that matters is that having a propositional attitude (or reporting about that attitude) involves entertaining a proposition, and entertaining a proposition (as per the Spinozan view) causes belief.[40]

Critically, the vast majority of philosophers who embrace non-doxasticism assume that the cognitions involved in delusions are propositional attitudes (Matthews, 2013). That is to say, delusional reports are expressions of mental states consisting of a representation of a proposition (e.g., 'my wife is an imposter') and an attitude towards that proposition ('I *imagine* that', 'I *accept* that', etc.). On some more parsimonious versions of non-doxasticism, delusions are identified with standard everyday propositional attitudes (ones that we already countenance) such as imaginings, or perhaps combinations of these, such as *imagining that p* and *believing that one believes that p* (Currie and Ravenscroft, 2002; Currie and Jureidini, 2001; Currie, 2000). On other versions, delusions are identified with hitherto undiscovered propositional attitudes, for example, mental states intermediate between imagination and belief, what Andy Egan (2008a) calls *bimaginings*. Regardless of which type of non-doxasticism one embraces, the Spinozan architecture would render doxasticism inescapable, since whatever attitude one has toward a proposition, it would result in the proposition being believed. The more interesting considerations, to which I now turn, concern how the Spinozan architecture can accommodate those features of delusion that are seen as most indicative of their rationality failures, and thus most problematic for doxasticism: unresponsiveness to evidence, circumscription, and double-bookkeeping (Dub, 2017).

5.5. Paradigmatic Features of Delusions

5.5.1. Unresponsiveness to Evidence

There are two senses in which delusions are said to be unresponsive to evidence: they are formed on the basis of insufficient evidence and they are maintained in the face of counterevidence. It is not clear that *all* delusions are equally unresponsive to evidence in the ways the definition suggests.[41] What is certain, however, is that *most* are. Consider a classic

[40] This brings us naturally to the following question: why do not people who interact with the deluded person also adopt the delusion? Were all else equal, they too should entertain the delusional proposition (e.g., 'that is not his wife'), and thus believe it. But all else is not equal. As we will see (Section 5.7.1.), explaining why they do not believe it is going to be part of the story for why the deluded person does. A further, related, question is why the deluded person can't drop his delusion by just hearing someone deny its content (e.g., by hearing the misidentified person say 'I am your wife'). This too is addressed in Section 5.7.1.

[41] The 'evidence' for some delusions comes in the form of irregular perceptual experiences (e.g., Davies et al., 2001; Maher, 1999; Stone and Young, 1997). Also, some delusions are sensitive to arguments and evidence to the contrary (e.g., Bortolotti, 2010).

case of delusional jealousy, in which a man claims that his wife is unfaithful to him because the fifth lamp-post along on the left is unlit (Sims, 2003, p. 119). The 'evidence' the man adduces to prove that he is being cheated on is not merely insufficient to support his claim, but altogether irrelevant. There is clearly no intelligible link between his wife's infidelity and the lamp-post being unlit, suggesting that the man's delusion lacks evidential support. The clinical literature also abounds with examples of delusions that are resistant to change, even when counterevidence becomes available. For instance, in one striking case of anosognosia for hemiplegia (i.e., the inability to recognise impairments resulting from brain damage), one patient reports that she can clap despite the fact that her left hand is visibly motionless and no sound is audible (Berti et al., 1998, pp. 29–30).

5.5.2. Circumscription

Another feature of delusions that has been brought to bear against their status as beliefs is the circumscribed role they play in one's cognitive economy. Circumscription is especially evident in monothematic delusions like Capgras. Delusions are circumscribed in three respects, the first of which is *inferential* (Campbell, 2001; Hamilton, 2007). This is not to say that people never make inferences from their delusions. On the contrary, if you ask a person with Capgras how and why loved ones are not who they appear to be, she will often elaborate on the content of her delusional state. Such elaborations, however, are typically disconnected from the rest of the person's worldview (Stone and Young, 2007), suggesting that the delusion must play a weak inferential role in the person's reasoning. Monothematic delusions are generally only elaborated when solicited with why-questions, but many other delusions that occur in schizophrenia (so-called 'polythematic') are elaborated spontaneously in intricate ways. It has been noted that even much elaborated delusions may exhibit some degree of circumscription (Sass, 1994), as evidenced by the phenomenon of double-bookkeeping (see next subsection).

The second respect in which delusions are circumscribed is *behavioural*, where this means that they tend to be causally inert in respect of behaviour (Bortolotti, 2012; Currie and Ravenscroft, 2002; Currie, 2000; Egan, 2008a; Young, 2000). Subjects with delusions often do not act on them and display inconsistent verbal and non-verbal behaviours. Suppose I tell you that an imposter who looks just like my mother has taken her place. If I truly believe that my mother has been replaced, then you would expect me to act in ways that are consistent with and can be explained by my belief. For instance, you might expect me to file

a missing person report, run away from the imposter, take some steps to locate my mother, and try to figure out how the switch occurred. However, people who experience the Capgras delusion may fail to do any of these things, with some even carrying on friendly interactions with the imposter (Lucchelli and Spinnler, 2007). Similarly, paranoid patients who accuse the nursing staff of trying to poison them may eat their meals without complaint and readily take oral medications as prescribed (Sass, 1992, p. 274).

The third respect in which delusions are circumscribed is *affective*, which is to say that they are often characterised by a lack of appropriate accompanying affect (e.g., the kind of emotional comportment we would expect from someone who genuinely believe the things they profess to believe, see Stone and Young, 1997; Sass, 1994). Using again the example of Capgras delusion, we can suppose that if your mother went missing, you would probably be concerned about her whereabouts. If you believed a person to be an imposter, you would most likely respond with rage and angry outbursts. However, most Capgras subjects show little concern about the welfare of their loved one and seem not much troubled by the presence of the imposter (Alexander et al., 1979)[42].

5.5.3. Double-Bookkeeping

The term 'double-bookkeeping' refers to the fact that people who sincerely avow their delusions operate on two parallel but separate clusters of representations and retain a roughly accurate sense of which is which (Sass, 1994). On the one hand, there are the true beliefs which they hold about objective reality. On the other, there are the delusional 'beliefs' they profess about some inner subjective reality. Double-bookkeeping applies equally well to psychotic breaks and delusional episodes in schizophrenia, some of which are elaborated and polythematic, as to monothematic delusions. An individual with schizophrenia might have a complex delusional system involving undercover FBI agents conspiring to kill her, yet at the same time take long walks at night by herself. Similarly, as we have seen, individuals with

[42] It should be noted that the degree to which delusions are behaviourally and affectively circumscribed can vary considerably from case to case. It is not as if delusions *never* motivate behaviour and affect in the way we would expect them to. Clinical studies have reported cases in which Capgras subjects grow extremely distressed and act against the alleged imposters, at times with deadly consequences (e.g., Christodoulou, 1977; De Pauw and Szulecka, 1988) This, however, does not make circumscription any less problematic from the point of view of doxasticists. Merely citing the existence of conflicting cases does not really explain what needs explaining, which is why delusions are *so often* weakly behaviour guiding and not accompanied by emotional responses appropriate to their content (Dub, 2017).

Capgras might confidently assert that imposters are substituting for their wife or husband, and yet fail to even search for their displaced spouses.

5.6. Why Spinozan Doxasticism?

Many have found that the above features of delusions cast doubt on doxasticism. The focus is on the concept of belief as used in folk psychology, the everyday practice of ascribing mental states to self and others as a way of predicting and explaining behaviour. Anti-doxastic arguments tend to move from an 'interpretationist' understanding of folk psychology, according to which when we ascribe mental attitudes to others, we aim to explain their behaviour given those mental attitudes. For instance, if you see me take an umbrella and a raincoat when I leave home in the morning, you can legitimately ascribe to me the belief that it is raining. A key idea behind interpretationism is that we must assume that a system is rational if we wish to explain the behaviour of that system via attitude ascription (e.g., Davidson, 1982, 1985; Dennett, 1987). One implication of this is that we can ascribe to a system only rational beliefs. Delusions, goes the anti-doxasticist, fail to be rational in several respects: they are often unsupported by reasons, inconsistent, and partially otiose. Therefore, it is concluded, delusions cannot qualify as beliefs.

We have seen that the intuitive strength of anti-doxastic arguments stems from the interpretationist assumption that one needs to be rational in order to be ascribed beliefs. The standard defence argument, developed at book-length by Bortolotti (2010), is to reject the rationality constraint on belief ascription. Bortolotti (2010) is concerned to adjudicate between 'the good and the bad of interpretationism' (p. 261). She thinks it is good that we understand belief in terms of our belief-ascribing practices. This means there is a threshold of interpretability that marks the minimum for a mental state to qualify as belief. But she also thinks it is bad that we idealise belief ascription by identifying such a threshold with a general backdrop of rationality (Bortolotti, 2010, p. 261).

As such, her strategy is to offer for each delusion a belief with a comparable type if not degree of irrationality. If it can be shown that the attitude in question is one that we are happy to ascribe as a belief to others, then rationality should not be invoked as a constraint on ascription. This has two implications for the status of delusions as beliefs. First, if we refuse to impose rationality constraints on belief ascription, then we cannot legitimately deny belief status to delusions on the ground of irrationality. Second, if the belief in question is irrational in a way comparable to delusions, then there would be reason to conclude that

delusions are beliefs. In support of this argument, Bortolotti (2010) cites numerous examples in which we, as interpreters, are willing to ascribe irrational beliefs. Many people express self-deceptive beliefs (e.g., beliefs motivated by desires and emotions) which can be very resistant to critical engagement. It is also common for people to maintain superstitious beliefs (e.g., ones involving supernatural causation) that are inconsistent with the other beliefs they have (e.g., their commitment to a scientific world view). Again, people often fail to act on their self-reported beliefs. For instance, they might express the belief that sex without a condom can lead to HIV infection, but admit to not using condoms (Aronson, 1999). These cases violate rational constraints no less than delusions do: they are unresponsive to evidence, badly integrated with one's other beliefs, weakly behaviour guiding. And yet, we are comfortable ascribing them as beliefs. If this is true, the story goes, we can interpret delusions as beliefs too.

The standard defence makes a convincing case for the conclusion that delusions can be ascribed as belief states. Still, the question can be asked whether doxasticism can be satisfactorily defended solely on folk-psychological grounds. Indeed, it may be thought that a satisfactory defence of doxasticism should also enable us to see why delusions, in virtue of being beliefs, deviate from rationality in the way they do. We seem to give up that explanatory component of doxasticism if we focus on just the criteria for the ascription of beliefs. For doxasticist to avoid this outcome, beliefs would need to be accounted for in a viable model of cognitive architecture. I will argue that the Spinozan theory provides such a model, and in so doing has the explanatory component won: if belief fixation is Spinozan, unresponsiveness to evidence, circumscription, and double-bookkeeping are natural— or at least predictable—results.

5.7. Delusions as Spinozan Beliefs

5.7.1. Evidence-Less/Resistant Beliefs

Spinozan belief fixation is arational, and therefore, by definition, unresponsive to evidence. Because of our cognitive architecture, we are set up with dispositions to immediately believe any propositions we happen to token, without weighing evidence first.[43] We acquire beliefs

[43] If 'evidence' were understood loosely to mean 'information' or 'data', one could claim that Spinozan belief fixation is hyperresponsive to evidence, because every proposition represented is thus believed. Here, however, 'evidence' should be understood in the epistemological sense, namely as something that makes a difference to what is reasonable for one to believe (e.g., Kelly, 2014). So, when I say that Spinozan belief fixation is 'unresponsive to evidence', I mean that believing occurs irrespective of what one is justified in believing.

in a brute causal way, via processes that work below the psychological level and which, as such, are impossible to counteract psychologically (Levy and Mandelbaum, 2014).

In principle, we are free to reject our newly acquired beliefs. The Spinozan model indicates that disbelief and suspending judgment are possible, albeit only as modifications of an initially untested belief. However, we have seen that our capacity to reject false beliefs can be overridden by an increase in cognitive load (i.e., extra mental processing imposed on the cognitive system). There are several factors that can increase load, most of which are woven into normal daily activities. Two notable examples are mind-wandering and distraction. In everyday life we are exposed to a vast amount of information, while our focus of attention is continuously switching between external stimuli and internal thoughts. Merely dividing attention between features presented in the same or different modalities is a primary source of cognitive load. In the study by Gilbert and colleagues discussed above (1993), subjects in the interrupted condition were told to read the text of the crime reports and concurrently search for the digit 5 whenever it appeared in the number line. The aim of this experimental design was to create a condition of split attention between two visual cues. The result was that the prison terms recommended by interrupted subjects were significantly affected by propositions they knew to be false. Beliefs were not simply acquired in an 'evidence-less fashion', but also retained in the face of disconfirming evidence (Mandelbaum and Quilty-Dunn, 2015).

Another factor that has been used in different studies to induce cognitive load is time pressure. Interestingly enough, it has been suggested that increased time pressure can make it difficult to suppress stereotype-based beliefs, even in circumstances where we most want to inhibit them (Huebner, 2009). For example, one study found that subjects who were admonished not be sexist in completing sentences with a missing word (e.g., 'Women who go out with a lot of men are …') were more likely to make sexist completions when load was induced by asking for immediate responses (Wegner et al., 1993)

In sum: resistance to belief revision is to be expected in situations where an increase in cognitive load due, for example, to heightened attention or time pressure depletes cognitive resources that are needed to override false beliefs.

There is suggestive evidence that one such situation is encountered, in a particularly acute and exacerbated fashion, among delusional and delusion-prone individuals. Several studies have focused on content-specific attentional biases in delusions. The most commonly used paradigm to study attentional bias is the emotional Stroop test (e.g., Gotlib and

McCann, 1984), in which participants are instructed to name as quickly as possible the colour in which affect-laden and neutral words are printed, while ignoring their semantic meaning. If subjects' performance on affect-laden words is observed to be markedly slower relative to neutral words, attentional bias is inferred; this is because delays in colour naming are taken as indicative of greater attentional capture by the meaning of words. Bentall and Kaney (1989) studied the Stroop performance of people with persecutory delusions and found that their subjects responded most slowly to threat-related words (e.g., 'kill', 'spy', 'pain'), an effect which was not observed for psychiatric (depressed) and normal controls. This finding has been replicated subsequently (Fear et al., 1996). Attentional bias for delusion-specific information was further demonstrated in the case of a woman, JK, who was convinced that she had died (Cotard delusion), and that members of her family were not who they seemed (Capgras delusion). JK showed disproportionately longer colour-naming times for tests lists containing death- and duplicate-related words relative to sets of neutral words. After her delusions had disappeared, however, her colour-naming times did not vary across test and neutral lists (Leafhead et al., 1996).

Abnormalities in selective attention—particularly in attentional inhibition—have also figured importantly in schizophrenia research. One of the most consistently observed features in early schizophrenia (i.e., prior to the onset of frank psychotic symptoms like delusions) is the inability to screen out irrelevant stimuli from the environment, resulting in sensory overload and increased distractibility. Patients say such things as 'I can't shut things out'; 'Everything seems to go through me'; 'I am attending to everything at once and as a result I do not really attend to anything' (McGhie and Chapman, 1961, p. 104). Failure to inhibit distractors has been explained in terms of aberrant assignment of 'salience' to neutral stimuli, driven by an excess of dopamine release outside the proper context (Kapur, 2003; Kapur et al., 2005). The effect of this is to make events attention-grabbing that would otherwise be inconspicuous, creating a sense of uncertainty and unpredictability, as well as a sense of urgency and time pressure to resolve it.

Experiences of aberrant salience, so understood, are involved in the formation of some, but not all, delusions. For example, when salience is misattributed to unrelated or coincidental phenomena (e.g., a causal conversation between a couple of window-shoppers), these phenomena may be interpreted as pertaining specifically to oneself, giving rise to delusions of reference and persecution. Other delusions, however, are accompanied by highly specific experiences, which cannot simply be captured in terms of aberrant salience.

To be sure, these experiences may well be said to be phenomenally salient, if this means that the subject feels her attention being drawn to their contents. But even so, they are not *just* experiences of salience—they are not just about imbuing meaning to any stimulus or event that is currently occurring.

For example, people with Capgras delusion have an abnormally reduced affective responsiveness to familiar faces (Ellis et al., 1997; Hirstein and Ramachandran, 1997; Brighetti et al., 2007), with the result that persons whose faces are familiar feel unfamiliar. In another misidentification delusion, the Fregoli delusion (the belief that strangers are in fact familiar persons in disguise), people have an abnormally heightened responsiveness to unknown faces, and thus feel as if strangers are familiar to them (Langdon et al., 2014). In Cotard, there is a general flattening of affective responsiveness to all perceptual inputs, which could lead to the delusion of being dead (Young and Leafhead, 1996). For our purposes here, it does not matter how exactly delusions arise from such unusual experiences. The relevant question for us is whether such unusual experiences expend many cognitive resources. And the answer is certainly yes, not only because they are surprising, and thus in need of explanation, but also because they are often persistent and distressing for the person who undergoes them.

The take-away point is that delusional fixity—the fact that delusions are retained in the face of counterevidence—is predicted by the Spinozan architecture. If delusions are formed and maintained under conditions of mental load and time urgency, as seems the case, then we should expect these factors to interfere with the effortful processing needed for belief revision. This is because, for the Spinozan, effortful processing is resource-dependent, and so can be disrupted when people operate under cognitive load or with scarce cognitive resources available.

Taken together, the above considerations also help us address two questions that are potentially worrisome for Spinozan doxasticism (see fn. 40). One is the question: why does not everyone who entertains delusional propositions (e.g., people who live or are in close contact with the deluded individuals) end up being believing them? For instance, why does not the doctor who hears her patient say 'I am dead' believe that she has actually died? The second question is: why can't someone stop believing a delusional proposition by simply entertaining its negation (for instance, as a result of hearing friends and family members say, 'You are alive and well')?

The answer to the first question should be clear by now. Not everyone who entertains a delusional proposition is doing the same amount of cognitive work, and arguably delusional and delusion-prone subjects (e.g., first-episode patients with schizophrenia) have a much higher cognitive load than any healthy individual around them. Normally, when we consider outlandish propositions, we can immediately reject the correspondingly acquired beliefs. And that is because we have enough cognitive resources to allow for the rejection process to get off the ground.

But this is not always the case, as we have seen. Think, for example, of the attention biases toward delusion-salient stimuli shown by some delusional subjects. Not only are these subjects overwhelmed by attentional demands, and thus less apt to reject their initial beliefs, but they preferentially encode material that serves to maintain them (Bentall and Kaney, 1989). Or again, consider what happens in first-episode schizophrenia, where patients are unable to leave out irrelevant stimuli to the point where they feel flooded by information. These patients will be so distracted while entertaining a delusional proposition that they may utterly lack the requisite cognitive energy to reject the newly acquired belief.

Now let us complicate things a bit. Suppose that friends and family members are under cognitive load of a comparably high level to that undergone by the deluded individual. Does this mean that they will too believe the delusion upon entertaining the delusional proposition? This is a difficult question, and one which I cannot fully address. Still, I can make some tentative remarks. One possibility is this: the person in question continues to believe the delusional proposition but her credence in it is weaker than an outright belief, or simply not enough for the relevant belief to become consciously occurrent. Perhaps credence-boost fails to occur because the person thinks the deluded individual is not reliable, or because, absent experiential abnormalities, background beliefs have the power to prevent it. An opposite possibility is that the person gets credence-boost from entertaining the delusional proposition, with the result that she too ends up endorsing the delusional belief. This may happen as a consequence of the overall psychological situation the person is in. For instance, if the person has a paranoid predisposition, and the theme of the asserted proposition is paranoid (e.g., 'someone is listening to me through my phone'), her credence may arise more easily to an outright belief.

What about the second question: why is it that one does not reject the delusional belief after entertaining the negated delusional proposition? The answer here is threefold. First, attentional biases, aberrant salience, and anomalous experiences are cognitively

burdensome enough to make the rejection a proposition lot harder, and this is true even if one entertains its negation. Secondly, persistent exposure to information carried in anomalous experiences (e.g., '*x* feels unfamiliar') keeps bringing up delusional hypotheses anew (e.g., '*x* is unfamiliar'), and, as already noted, evidence suggests that belief increases with repeated exposure to the believed proposition (Fazio et al., 2019). Finally, as we also noted, the Spinozan view allows for the possibility of inconsistent beliefs, which means that one may believe the negation of a delusional proposition *p*, without necessarily disbelieving *p*.[44]

5.7.2. Fragmentation, Circumscription, and Double-Bookkeeping

Before concluding this section, let us see how the Spinozan theory allows for circumscription and double-bookkeeping whilst retaining the doxastic status of delusions.

One way that doxasticists can attempt to accommodate circumscription is by appeal to fragmentation.[45] As long as beliefs are fragmented, they can be stored without having to guide all of our behaviour in all contexts. Instead, since they are stored in separate compartments of the mind, they can drive different bits of our behaviour in different contexts. To say that fragmented beliefs would guide behaviour in some, but not all, contexts is equivalent to saying that they are *behaviourally* circumscribed. And, of course, the same thing can be said, mutatis mutandis, concerning emotional responses and *affective* circumscription. Finally, different fragments can contain inconsistent information, which would explain why some of our beliefs are *inferentially* circumscribed and not integrated with the rest.

For doxasticists the importance of this is that it bolsters their argument by way of the following reasoning. There is a clear correlation between fragmentation and circumscription. So if our belief system is in fact fragmented, then circumscription supports rather than undermines the continuity between delusions and beliefs (Bortolotti, 2010, p. 89).

[44] The following subsection will spell out in more detail how this might happen.

[45] Bayne and Pacherie (2005) argue that behavioural circumscription is excusable by reference to non-standard features of the circumstance one finds oneself in. For example, it is difficult to say which action or attitude would be consistent with a person's belief that someone is inserting thoughts into her mind. Clearly, however, not every case of behavioural circumscription can be excused in this way: it is plausible to expect of someone who believes her mother has been replaced by an imposter to report her missing. But there is more. Bayne and Pacherie's point is not just that it is hard to know what to expect in some cases of delusion. Rather, the point is that even in cases where the behaviour we would expect is not manifested, that can be excused: it might be that a person does not turn to the police because she knows that the imposter is a perfect lookalike and she fears she will not be taken seriously. This is plausible, but cannot be generalised across all the behavioural dispositions that delusional individuals fail to manifest. What should excuse a person who believes she is being poisoned from eating her food? It is hard to imagine anyone choosing the possibility of severe illness or death over being sectioned or whatever else.

One possible objection is that the idea of beliefs being fragmented is an ad hoc stipulation to save doxasticism: (P1) Belief fragmentation is typically defended on the grounds that beliefs are not always consistent, deductively closed, and effective in guiding an agent's behaviour (e.g., Egan, 2008b); (P2) The main limitation of doxasticism, as standardly devised, is that it merely asserts that, and sheds no light on why, belief has those features; (P3) The only thing to which doxasticism can appeal in order to explain why belief has these features is that belief is fragmented; (C) Therefore, the appeal to fragmentation is not revealing but rather begging the question.

While this objection has some force against the standard defence, it has no, or at least considerably less force, against Spinozan doxasticism. The reason is that the Spinozan architecture is particularly well placed to accommodate a fragmented picture of belief. For the Spinozan, we automatically believe propositions before being able to reject them. In order to reject *p*, we must already believe *p*. Rejecting *p* can only take place if a conflict is detected between *p* and other beliefs we hold. Since we already believe *p*, the very existence of a conflict hinges on the possibility of inconsistent beliefs. As such, the second stage of the Spinozan model, the evaluation of newly acquired beliefs, makes no sense unless our beliefs are fragmented. Moreover, since the rejection process is often disrupted by cognitive load, many inconsistencies are left unresolved.

What has this got to do with doxasticism? Assume the Spinozan model is true. Then clearly our belief system ought to be fragmented, because, necessarily, many of our beliefs are inconsistent with the rest, and often remain so, even when conflicts are detected. Given this, an appeal to fragmentation here is not susceptible to the charge of being ad hoc, because it is based on a principled model of belief fixation. Consequently, there is a stronger case to be made that delusions are beliefs.[46]

With these observations in mind, let us turn to double-bookkeeping. How do we explain someone's being aware of the delusional nature of their delusions, and yet failing to

[46] In order to avoid confusion, let me emphasise that I am not trying to argue that the Spinozan theory provides the best explanation (tout court) for belief fragmentation. My point is just that Spinozan doxasticism fares better than standard doxasticism in this respect. As I have been arguing, standard doxasticism does not tell us why key features of delusions, such as circumscription and inconsistency, or equivalently, double-bookkeeping, are features of beliefs. What the standard doxasticist can do is appeal to belief fragmentation: beliefs are circumscribed and mutually inconsistent because they are fragmented. The problem with this is that it may strike one as circular, since circumscription and inconsistency are themselves among the best evidence supporting fragmentation. Spinozan doxasticism avoids the charge of circularity by giving an account of belief fixation that 'ensures' fragmentation.

make up their minds about them? Note that the awareness that is involved by double-bookkeeping can be implicit or explicit. Some subjects might act or react contrary to their delusions, without actually consciously recognising the tension. This is the case, say, with a Capgras subject who states that his wife has been replaced by an imposter and yet continues to live in friendly terms with her (Rose et al., 2014). Other subjects, by contrast, might report conflicting beliefs in the course of a single conversation. Even though they explicitly recognise such conflicts, they do nothing to resolve them. For example, McKay and Cipolotti (2007) describe the case of a young woman with Cotard delusion, LU, who claimed to be dead. When asked how she would know when someone is dead, LU replied that dead people lay motionless with their eyes closed. Later in the interview, she recognised the inconsistency between her being dead and yet being able to move and speak, but she continued to maintain that she was dead.

Again, I stress that the question we need to ask is not (at least not only) whether there are quotidian states that exhibit double-bookkeeping and which we are prone to attribute to others as beliefs. The question, rather, concerns how double-bookkeeping is tied to those states being beliefs. In answering this question, the doxasticist can appeal once more to fragmentation: the phenomenon of double-bookkeeping is difficult to square with a unified web of belief, but naturally explained by a fragmented mind that allows for coexistence of dissonant beliefs. Still, we might feel that this leaves out something important about why subjects are indifferent to dissonance among their beliefs. One can grant that ordinary believers might be inconsistent, and yet expect them to restructure their beliefs once the inconsistency is brought into awareness (Bendana and Mandelbaum, forthcoming). Consider a case where an agent believes p and believes $\neg p$, but experiences no conflict because only the fragment containing p is activated at the time. Since no inconsistency is detected, there is no incentive for the agent to restore coherence in her belief system. Now, imagine another scenario with the same individual, but where p and $\neg p$ are both in activated fragments. Here, we would expect the two fragments to be rendered consistent, since the agent is co-attending to mutually conflicting beliefs. The worry, then, is that appeal to fragmentation alone may not be sufficient to account for double-bookkeeping. To see how Spinozan doxasticism might address this worry, recall that for the Spinozan, rejecting beliefs is a breakdown prone process, one which stalls under cognitive load. What happens when the rejection process goes awry? There are at least two possibilities, each arguably corresponding to one of the two senses of double-bookkeeping outlined above.

One possibility is that the agent *thinks* she has discarded some belief in virtue of its inconsistency with the rest, when in fact she has not. Consequently, she refrains from using the belief for conscious planning and verbal behaviour, but continues to access it in low-level behaviour. For instance, one may explicitly disavow any form of racial prejudice, yet nevertheless continue to act in racist ways. This is similar to the case of the Capgras subject who denies his own wife being his wife, but never ceases to treat her as such. The idea is simple. Since the agent thinks she has already re-established coherence in her belief system, she no longer experiences dissonance. However, due to a failure of the revision process, the apparently 'discarded' belief persists and exerts a continuing impact on her behaviour.

A second possibility is that the agent realises that something has gone amiss with the revision process, and remains aware of having inconsistent beliefs, but ultimately tolerates such inconsistencies. This might be what happens with LU in the example above, whereby the belief that dead people are speechless coexists with (the verbal expression of) the belief that one is dead. It also may explain the continuity between cases like LU's and non-delusional cases of superstition. People readily combine superstitious beliefs about magic and supernatural causation (e.g., divine intervention) with a scientific stance toward the world. Although they know that such beliefs are irreconcilable, they often fail to decisively resolve the conflict in one way or another, even when called upon to do so (Vyse, 2014)

If this suggestion is on the right track, then a Spinozan architecture could be used to clarify not only why beliefs are fragmented and variously circumscribed, but also why people retain inconsistencies between coactive beliefs, resulting in double-bookkeeping.

5.8. Remaining Concerns

I have argued that Spinozan doxasticism can accommodate features of delusions which many have felt are telling against their status as beliefs. More importantly, I have argued that Spinozan doxasticism is more explanatory than standard doxasticism, because it explains why delusions, in virtue of being beliefs, have the features they do. As already indicated implicitly, but as I will now make more explicit, to say that delusional beliefs have these features *in virtue of being beliefs* is not to say that *all* of our beliefs have these features. Of course, as Bortolotti (2010, 2018) has convincingly argued, some non-delusional beliefs, such as superstitious beliefs, share many of the key features of delusional beliefs: among other things, they are very resistant to change, even when counterevidence becomes available. However, not all of our beliefs are of that kind. Your belief that it is raining outside is swiftly dislodged when

you look outside and see that it is raining. So the question arises, if beliefs really are Spinozan, what explains why delusional beliefs have features that most non-delusional beliefs do not have? The answer lies with the fact, pointed out in Subsection 5.7.1., that the former typically arise and are maintained under conditions of high cognitive load (i.e., attentional demands and time pressure), combined with unusual, persistent, and distressing experiences (which too are load inducing). The upshot, then, is that Spinozan beliefs need not always display the kinds of features that obtain in cases of delusion, but they do if a particular set of circumstances come together.

Before closing, I have two further issues to address. The first is whether delusions actually fit the distinctive profile of Spinozan beliefs. The second is why doxasticists in general should want to accept such an unorthodox conception of belief.

To the first task: Spinozan beliefs are unconscious propositional attitudes which are not available to introspection; delusions, on the contrary, are conscious, at least in the sense that they are manifested in consciousness by an occurrent thought. How does one reconcile these viewpoints into a single conception of delusions as beliefs? My answer to this is that the conflict in question is not a genuine one. Spinozans deny that we are introspectively aware of beliefs qua beliefs. But they do not deny that we can introspect the content of beliefs. So, it may well be that delusions are unconscious beliefs, in the sense that they are content-carrying mental states whose content one can introspect, but whose functional role one cannot.

Qua beliefs, delusions are contents with a certain functional role. This includes their characteristic function in one's mental economy (e.g., their relation to evidence, other kinds of intentional states, and behaviour). These relations are not themselves conscious. Thus, one could conclude that delusions are not conscious tout court. All one is aware of is their belief content, yet a content is, in itself, not a belief.

Having said that, it is very much up for grabs whether *all* forms of delusion are unconscious beliefs whose content is present to mind and available for verbal report. Some delusional statements, while sincere, might not be genuine belief reports, that is, reports expressing actual belief contents.

Recall (Section 5.3.3.) that there are two ways in which the Spinozan thinks we find out what we believe. One way is by a direct introspecting act which makes us aware of the content of our beliefs. When this happens, what we sincerely report believing is largely coincident with what we actually believe. Another way proceeds via self-interpretation. We

infer what we believe by observing our own behaviour, by considering what seems more reasonable, or through other sorts of interpretive strategies. In this case, the 'beliefs' that we report having may well end up being nothing over and above what we actually believe. But most often they are the products of confabulations made up on the spot, which have little if anything in common with our actual beliefs. This is because self-interpretation is influenced by a multiplicity of factors, especially social (e.g., group identity) and motivational (e.g., reducing anxiety). As such, belief reports are often calibrated to fit the beliefs and reactions of those around us.

In essence, there is a distinction to be drawn between beliefs and belief reports. The former are brute architectural matters (Mandelbaum, 2014); our minds are designed to automatically believe any propositions to which they are exposed. The latter express judgments or claims about the contents of beliefs, and as such they can be genuine (e.g., direct introspective reports) or spurious (e.g., mere endorsements). It is not my aim here to consider which delusions are genuine belief reports and which are not. That will depend, inter alia, on how delusional hypotheses arise in the first place. To illustrate, consider the following two scenarios.

(i) The sight of your wife does not evoke the characteristic feeling of familiarity that normally accompanies the recognition of her face. You find passing thoughts popping into your mind, among which the implausible hypothesis (i.e., 'that woman who looks like my wife is not really my wife'). Tokening the proposition causes you to believe it, but only weakly at a nonconscious level. After days or weeks, you raise your credence in that proposition, until it becomes available in the form of a judgment (reporting your actual belief) that your wife is not really her.

(ii) Other types of delusion might be analysed as the combined result of personal and social concerns, plus the self-evaluation of being relatively immune to such concerns. Concerns might arise with regard to your autonomy and power in relation to others, for instance about the trustworthiness of friends and associates (paranoia), your class status (grandiosity), or your appealingness or love-worthiness (erotomania). Your struggle to establish an emotionally acceptable self-definition makes it so that you are caught in increasingly arbitrary and idiosyncratic interpretations of yourself and the world, which slowly transform into prolific delusional systems. Some have suggested that these interpretations appear idiosyncratic to others because they are developed in isolation from

social relations, perhaps due to impaired Theory of Mind skills (Bora et al., 2009; Bentall, 2018)

Obviously, these are oversimplifications, but nonetheless useful ones for our current purposes. Thinking in Spinozan terms, the former scenario (i) corresponds to a case where an agent rightly reports what she believes, the latter (ii) to cases where an agent merely infers what she takes herself to believe. This distinction brings about an additional advantage of Spinozan doxasticism. It is flexible enough to account for the heterogeneity present in the class of delusion. While some delusional reports may reflect the content of one's actual beliefs, others may be the products of an online elaboration, in which on-the-spot hypotheses are constructed and defended with arguments. It is important to realise, however, that when an agent takes herself to believe that *p*, she automatically believes that *p*. This means that even delusions that arise as *mere* belief reports (i.e., ones that are not reflective of actual belief contents) are liable to become beliefs.

To the second task: standard doxasticists like Bortolotti think delusions are beliefs where what they mean by belief is a state whose fixation can be explained on a psychological level, a level where talk of rational and irrational inferences makes sense. Why should they care if on some other view delusions are 'beliefs', but belief fixation is a brute causal mechanism operating below the psychological level? I make two points about this.

First, for Bortolotti no less than for the Spinozan, beliefs come cheap.[47] In the place of rationality constraints, Bortolotti (2010) offers more realistic features of beliefs that she thinks are to guide everyday interpretation:

> i. Beliefs have *some* inferential relations with other beliefs, wishes, desires, etc.
> ii. Beliefs display *some* sensitivity to evidence and argument.
> iii. Beliefs can be, but need not be, manifested in behaviour.
> iv. Beliefs can be self-ascribed (i.e., acknowledged as one's own), and their content can be endorsed (i.e., defended with reasons).

Each of these features fits squarely with the Spinozan conception of belief. From the foregoing discussion, it should be obvious that Spinozan beliefs are (i) inferentially

[47] Note that this is not unique to Bortolotti. Doxasticists in general are likely to be liberal about belief as compared to non-doxasticists, who tend to be stricter (e.g., Ichino, 2020; Archer, 2013). Here, I focus on Bortolotti as the most prominent representative of doxasticism.

promiscuous and able to (iii) cause behaviour. It should also be clear that at least some Spinozan beliefs are (ii) sensitive to evidence. For one thing, they can be revised in the wake of new information, although the process of revision is often short-circuited by cognitive load. For another, they can be formed, albeit contingently, based on evidence. Think for example of perceptual beliefs, whereby one believes that *p* based on a perceptual experience with the content that *p*. Finally, we have seen that, (iv) while beliefs are acquired in ways not available to introspection, agents can ascribe beliefs to themselves by introspecting their contents. In doing so, they also endorse such contents, which is to say that they are disposed to defend them with reasons or using rhetorical strategies.

Second, Spinozan doxasticism yields two results that all doxasticists should like. First, it entails that delusions are the same kind of attitudes as beliefs concerning garden-variety facts, such as the belief that there is some leftover pizza in the fridge. The entailment is based on the following reasoning. To have any attitude toward a propositional content *p* is equivalent to believing *p*. In having a delusion that *p*, one is psychologically related to *p*. Therefore, being deluded that *p* is the same kind of attitude as believing that there is some leftover pizza in the fridge. Secondly, it offers a simple and intuitively satisfying explanation of why the key features of delusion should be taken as features of belief: delusions are the way they are as a result of the way belief fixation actually works.

5.9. Conclusion

The Spinozan theory of belief fixation has it that the mere act of representing a proposition leads to immediately believing it. Minds like ours are such that they cannot *merely* represent a proposition. Rather, propositions are believed as quickly as they are represented. I have argued here that this view has important consequences for the debate over the doxastic status of delusions. Specifically, if we accept this view, it gives a more robust defence of doxasticism than the dominant one in the literature, offered by Bortolotti. Doxasticism has been criticised on the grounds that delusions fail to conform to certain rationality standards which we expect beliefs to conform to. Against this objection, Bortolotti points to typical cases of belief ascription which also fall foul of rationality standards. However, some may say this is putting the cart before the horse. Doxasticism is supposed to make us understand why delusions qua beliefs behave as they do. It does not do that if it only focuses on belief ascription. I have shown that Spinozan doxasticism, unlike the standard doxastic defence, is capable of providing such an understanding.

5.10. References

Alexander, M. P., Stuss, D. T. and Benson, D. F. (1979). 'Capgras Syndrome: A Reduplicative Phenomenon'. *Neurology*, 29(3), pp. 334–334.

Appel, M. and Richter, T. (2007). 'Persuasive Effects of Fictional Narratives Increase Over Time'. *Media Psychology*, 10(1), pp. 113–134.

Archer, S. (2013). 'Nondoxasticism about Self-Deception'. *Dialectica*, 67(3), pp. 265–282.

Aronson, E. (1999). 'Dissonance, Hypocrisy, and the Self Concept'. In E. Harmon-Jones and J. Mills (eds.) *Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology*. Washington: American Psychological Association, pp. 103–126.

Bayne, T. and Pacherie, E. (2005). 'In Defence of the Doxastic Conception of Delusions'. *Mind and Language*, 20(2), pp. 163–188.

Bendana, J. and Mandelbaum, E. (forthcoming). 'The Fragmentation of Belief'. In D. Kinderman, A. Onofri and C. Borgoni (eds.) *The Fragmented Mind*. Oxford: Oxford University Press.

Bentall, R. P. and Kaney, S. (1989) 'Content Specific Information Processing and Persecutory Delusions: An Investigation Using the Emotional Stroop Task'. *British Journal of Medical Psychology*, 62, pp. 355–364.

Bentall, R. P. (2018). 'Delusions and Other Beliefs'. In Bortolotti, L. (ed.) *Delusions in Context*. Cham: Springer International Publishing, pp. 67–95.

Bentall, R. P., Kinderman, P. and Kaney, S. (1994). 'The Self, Attributional Processes and Abnormal Beliefs: Towards a Model of Persecutory Delusions'. *Behaviour Research and Therapy*, 32(3), pp. 331–341.

Berrios, G. E. (1991). 'Delusions as 'Wrong Beliefs': A Conceptual History'. *British Journal of Psychiatry*, 159(S14), pp. 6–13.

Berti, A., Ladavas, E., Stracciari, A., Giannarelli, C. and Ossola, A. (1998). 'Anosognosia for Motor Impairment and Dissociations with Patients Evaluation of the Disorder: Theoretical Considerations'. *Cognitive Neuropsychiatry*, 3(1), pp. 21–43.

Bora, E., Yucel, M. and Pantelis, C. (2009). 'Theory of Mind Impairment in Schizophrenia: Meta-Analysis'. *Schizophrenia Research*, 109(1–3), pp. 1–9.

Bortolotti, L. (2010). *Delusions and Other Irrational Beliefs*. Oxford: Oxford University Press.

Bortolotti, L. (2012). 'In Defence of Modest Doxasticism About Delusions'. *Neuroethics*, 5(1), pp. 39–53.

Bortolotti L. (2018). 'Delusions and Three Myths of Irrational Belief'. In L. Bortolotti (ed.) *Delusions in Context*. Cham: Palgrave Macmillan.

Brighetti, G., Bonifacci, P., Borlimi, R. and Ottaviani, C. (2007). "Far from the Heart Far from the Eye': Evidence from the Capgras Delusion'. *Cognitive Neuropsychiatry*, 12, pp. 189–197.

Campbell, J. (2001). 'Rationality, Meaning, and the Analysis of Delusion'. *Philosophy, Psychiatry, and Psychology*, 8(2), pp. 89–100.

Christodoulou, G. N. (1977). 'The Syndrome of Capgras'. *British Journal of Psychiatry*, 130(6), pp. 556–564.

Coltheart, M. (2013). 'On the Distinction between Monothematic and Polythematic Delusions'. *Mind and Language*, 28(1), pp. 103–112.

Corlett, P. R., Krystal, J. H., Taylor, J. R. and Fletcher, P. C. (2009). 'Why do Delusions Persist?' *Frontiers in Human Neuroscience*, 3(Jul), pp. 1–9.

Currie, G. (2000). 'Imagination, Delusion and Hallucinations'. *Mind and Language*, 15(1), pp. 168–183.

Currie, G. and Jureidini, J. (2001). 'Delusion, Rationality, Empathy: Commentary on Martin Davies et al'. *Philosophy, Psychiatry, and Psychology*, 8(2), pp. 159–162.

Currie, G. and Ravenscroft, I. (2002). *Recreative Minds: Imagination in Philosophhy and Psychology*. Oxford: Oxford University Press.

Davidson, D. (1982). 'Paradoxes of Irrationality'. In R. Wollheim and J. Hopkins (eds.) *Philosophical Essays on Freud* (pp. 289–305). Cambridge: Cambridge University Press.

Davidson, D. (1985). 'Incoherence and Irrationality'. *Dialectica*, 39(4), pp. 345–354.

Davidson, D. (1974). 'Psychology as Philosophy'. Reprinted in D. Davidson (1982) *Essays on Actions and Events* (pp. 229–238). Oxford: Oxford University Press.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

De Pauw, K. W. and Szulecka, T. K. (1988). 'Dangerous Delusions. Violence and the Misidentification Syndromes'. *British Journal of Psychiatry*, 152(Jan.), pp. 91–96.

Dub, R. (2017). 'Delusions, Acceptances, and Cognitive Feelings'. *Philosophy and Phenomenological Research*, 94(1), pp. 27–60.

Egan, A. (2008a). 'Imagination, Delusion, and Self-Deception'. In T. Bayne and J. Fernandez (eds.) *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation*. Hove: Psychology Press.

Egan, A. (2008b). 'Seeing and Believing: Perception, Belief Formation and the Divided Mind'. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 140(1), pp. 47–63.

Ellis, H.D., Young, A.W., Quayle, A. and de Pauw, K. W. (1997). 'Reduced Autonomic Responses to Faces in Capgras Delusion'. *Proceedings of the Royal Society: Biological Sciences*, B264, pp. 1085–1092.

Fazio, L. K., Brashier, N. M., Payne, B. K. and Marsh, E. J. (2015). 'Knowledge Does Not Protect Against Illusory Truth'. *Journal of Experimental Psychology: General*, 144(5), pp. 993–1002.

Fazio, L. K., Rand, D. G. and Pennycook, G. (2019). 'Repetition Increases Perceived Truth Equally for Plausible and Implausible Statements'. *Psychonomic Bulletin and Review*, 26, pp. 1705-1710

Fear, C., Sharp, H. and Healy, D. (1996). 'Cognitive Processes in Delusional Disorders'. *The British Journal of Psychiatry,* 168(1), pp. 61–67.

Fodor, J. (1983). *Modularity of Mind.* Cambridge, MA: MIT Press.

Frankish, K. (2013). 'Delusions: A Two-Level Framework'. In Broome, M. and L. Bortolotti (eds.) *Psychiatry as Cognitive Neuroscience* (pp. 269–285). Oxford: Oxford University Press.

Frankish, K. (2016). 'Playing Double: Implicit Bias, Dual Levels, and Self-Control'. In M. Brownstein and J. Saul (eds.) *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology* (pp. 23–46). Oxford: Oxford University Press.

Freeman, D., Garety, P. A., Kuipers, E., Fowler, D. and Bebbington, P. E. (2002). 'A Cognitive Model of Persecutory Delusions'. *British Journal of Clinical Psychology*, 41(4), pp. 331–347.

Freeman, D. and Garety, P. A. (1999). 'Worry, Worry Processes and Dimensions of Delusions: An Exploratory Investigation of a Role for Anxiety Processes in the Maintenance of Delusional Distress'. *Behavioural and Cognitive Psychotherapy*, 27(1), pp. 47–62.

Friedman, J. (2013). 'Suspended Judgment'. *Philosophical Studies*, 162(2), pp. 165–181.

Gendler, T. S. (2007). 'Self-Deception as Pretense'. *Philosophical Perspectives*, 21(1), pp. 231–258.

Gerrans, P. (2014). *The Measure of Madness: Philosophy of Mind, Cognitive Neuroscience, and Delusional Thought.* Cambridge, MA: MIT Press.

Gilbert, D. T. (1991). 'How Mental Systems Believe'. *American Psychologist*, 46(2), pp. 107–119.

Gilbert, D. T., Krull, D. S. and Malone, P. S. (1990). 'Unbelieving the Unbelievable: Some Problems in the Rejection of False Information'. *Journal of Personality and Social Psychology*, 59(4), pp. 601–613.

Gilbert, D. T., Tafarodi, R. W. and Malone, P. S. (1993). 'You Can't Not Believe Everything

You Read'. *Journal of Personality and Social Psychology*, 65(2), pp. 221–233.

Gotlib, I. H. and McCann, C. D. (1984). 'Construct Accessibility and Depression: An Examination of Cognitive and Affective Factors'. *Journal of Personality and Social Psychology*, 47, pp. 427–439.

Green, M. C. and Brock, T. C. (2002). 'In the Mind's Eye: Transportation-Imagery Model of Narrative Persuasion.' In M. C. Green, J. J. Strange and T. C. Brock (eds.) *Narrative impact: Social and cognitive foundations* (pp. 315-341). Mahwah: Lawrence Erlbaum

Hamilton, A. (2007). 'Against the Belief Model of Delusion'. In M. C. Chung, Fulford, K. W. M. and Graham, G. (eds) *Reconceiving Schizophrenia* (pp. 217–34). Oxford: Oxford University Press.

Hasson, U., Simmons, J. P. and Todorov, A. (2005). 'Believe It or Not: On the Possibility of Suspending Belief'. *Psychological Science*, 16, pp. 566–571.

Hirstein, W. and Ramachandran, V. S. (1997). 'Capgras Syndrome: A Novel Probe for Understanding the Neural Representation of the Identity and Familiarity of Persons'. *Proceedings of the Royal Society of London B: Biological Sciences*, 264, pp. 437–444.

Holroyd, J. (2016). 'What Do We Want from a Model of Implicit Cognition?'. *Proceedings of the Aristotelian Society*, 116(2), pp. 153–179.

Huebner, B. (2009). 'Troubles with Stereotypes for Spinozan minds'. *Philosophy of the Social Sciences*, 39(1), pp. 63–92.

Ichino, A. (2020). 'Superstitious Confabulations'. *Topoi*, 39, pp. 203-217.

Kapur, S. (2003). 'Psychosis as a State of Aberrant Salience: A Framework Linking Biology, Phenomenology, and Pharmacology in Schizophrenia'. *American Journal of Psychiatry*, 160(1), pp. 13–23.

Kapur, S. (2004). 'How Antipsychotics Become 'Anti-Psychotic' – from Dopamine to Salience to Psychosis'. *Trends in Pharmacological Sciences*, 25(8), pp. 402–406.

Kelly, T. (2014). 'Evidence'. In E. N. Zalta (ed.) *Stanford Encyclopedia of Philosophy* (Fall).

Kinderman, P. (1994). 'Attentional Bias, Persecutory Delusions and the Self-Concept'. *British Journal of Medical Psychology*, 67(1), pp. 53–66.

Langdon, R., Connaughton, E. and Coltheart, M. (2014). 'The Fregoli Delusion: a Disorder of Person Identification and Tracking'. *Topics in Cognitive Science,* 6(4), pp. 615–31.

Leafhead, K. M., Young, A. W. and Szulecka, T. K. (1996). 'Delusions Demand Attention'. *Cognitive Neuropsychiatry*, 1(1), pp. 5–16.

Levy, N. (2015). 'Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements'. *Noûs*,

49(4), pp. 800–823.

Levy, N. and Mandelbaum, E. (2014). 'The Powers that Bind: Doxastic Voluntarism and Epistemic Obligation'. In Matheson, J. and Vitz, R. (eds.) *The Ethics of Belief* (pp. 15–32). Oxford: Oxford University Press.

Lewis, D. (1982). 'Logic for Equivocators'. *Noûs*, 16(3), p. 431.

Lucchelli, F., and Spinnler, H. (2007). 'The Case of Lost Wilma: A Clinical Report of Capgras Delusion'. *Neurological Sciences*, 28(4), pp. 188–195.

Madva, A. (2016). 'Why Implicit Attitudes are (Probably) not Beliefs'. *Synthese*, 193(8), pp. 2659–2684.

Mandelbaum, E. (2010). *The Architecture of Belief: An Essay on the Unbearable Automaticity of Believing*. PhD Diss. Chapel Hill: University of North Carolina.

Mandelbaum, E. (2014). 'Thinking is Believing'. *Inquiry*, 57(1), pp. 55–96.

Mandelbaum, E. (2016). 'Attitude, Inference, Association: On the Propositional Structure of Implicit Bias'. *Noûs*, 50(3), pp. 629–658.

Mandelbaum, E. and Quilty-Dunn, J. (2015). 'Believing without Reason: or: Why Liberals Shouldn't Watch Fox News'. *The Harvard Review of Philosophy*, (22), pp. 42–52.

Marsh, E.J. and Fazio, L.K. 'Learning Errors from Fiction: Difficulties in Reducing Reliance on Fictional Stories'. *Memory and Cognition*, 34, pp. 1140–1149.

Matthews, R. J. (2013). 'Belief and Belief's Penumbra'. In Nottelmann, N. (ed.) *New Essays on Belief: Constitution, Content and Structure* (pp. 100–123). London: Palgrave Macmillan.

McGhie, A. and Chapman, J. (1961). 'Disorders of Attention and Perception in Early Schizophrenia'. *British Journal of Medical Psychology*, 34(2), pp. 103–116.

McKay, R. and Cipolotti, L. (2007). 'Attributional Style in a Case of Cotard Delusion'. *Consciousness and Cognition*, 16(2), pp. 349–359.

Pylyshyn, Z. W. (1989). 'Computing in Cognitive Science'. In Posner, M. I. (ed.) *Foundations of Cognitive Science* (pp. 49–91) Cambridge, MA: MIT Press.

Quine, W. V. (1960). *Word and Object*. Cambridge, MA: MIT Press.

Quine, W. V. and Ullian, J. S. (1970). *The Web of Belief*. New York: Random House.

Richter, T., Schroeder, S. and Wohrmann, B. (2009). 'You Don't Have to Believe Everything You Read: Background Knowledge Permits Fast and Efficient Validation of Information'. *Journal of Personality and Social Psychology*, 96, pp. 538–558.

Rose, D., Buckwalter, W. and Turri, J. (2014). 'When Words Speak Louder than Actions:

Delusion, Belief, and the Power of Assertion'. *Australasian Journal of Philosophy*, 92(4), pp. 683–700.

Ross, L., Lepper, M. R. and Hubbard, M. (1975). 'Perseverance in Self-Perception and Social Perception: Biased Attributional Processes in the Debriefing Paradigm'. *Journal of Personality and Social Psychology*, 32(5), pp. 880–892.

Sass, L. (1992). *Madness and Modernism: Insanity in the light of Modern Art, Literature, and Thought*. New York: Basic Books.

Sass, L. (1994). *The Paradoxes of Delusion: Wittgenstein, Schreber, and the Schizophrenic Mind*. Ithaca: Cornell University Press.

Schwitzgebel, E. (2012). 'Mad belief?' *Neuroethics*, 5(1), pp. 13–17.

Sims, A. (2003). *Symptoms in the Mind*. 3rd edn. London: Elsevier.

Skurnik, I., Yoon, C., Park, D. C. and Schwarz, N. (2005). 'How Warnings about False Claims Become Recommendations'. *The Journal of Consumer Research*, 31, pp. 713–724.

Stalnaker, R. (1984). *Inquiry*. Cambridge: Cambridge University Press.

Stich, S. P. (1978). 'Beliefs and Subdoxastic States'. *Philosophy of Science*, 45(4), pp. 499–518.

Stone, T. and Young, A. W. (2007). 'Delusions and Brain Injury: The Philosophy and Psychology of Belief'. *Mind and Language*, 12(3–4), pp. 327–364.

Street, C. N. H. and Kingstone, A. (2016). 'Aligning Spinoza with Descartes: An Informed Cartesian Account of the Truth Bias'. *British Journal of Psychology*, 108, pp. 453–466.

Vyse, S. (2014). *Believing in Magic. The Psychology of Superstition*. New York: Oxford University Press.

Unkelbach, C. and Greifeneder, R. (2018). 'Experiential Fluency and Declarative Advice Jointly Inform Judgments of Truth'. *Journal of Experimental Social Psychology*, 79, pp. 78–86.

Wedgwood, R. (2017). *The Value of Rationality*. Oxford: Oxford University Press.

Wegner, D. M., Coulton, G. F. and Wenzlaff, R. (1985). 'The Transparency of Denial: Briefing in the Debriefing Paradigm'. *Journal of Personality and Social Psychology*, 49(2), pp. 338–346.

Wegner, D. M., Erber, R. and Zanakos, S. (1993). *On Trying Not To Be Sexist*. Unpublished manuscript.

Young, A. W. (2000). "Wondrous Strange': The Neuropsychology of Abnormal Beliefs'. *Mind and Language*, 15(1), pp. 47–73.

Young, A. W. and Leafhead, K. M. (1996). 'Betwixt Life and Death: Case Studies of Cotard

Delusion'. In P. W. Halligan and J. C. Marshall (eds.) *Method in Madness: Case Studies in Cognitive Neuropsychiatry* (pp. 147–171). Taylor and Francis: East Sussex.

Chapter 6: Delusions and Perceptual Justification

6.0. Abstract

According to a popular view, the 'endorsement' model, the process by which some delusions are formed is one of taking perceptual experience at face value. In this paper I ask whether delusions so formed enjoy any degree of epistemic justification. To give the best chance of answering affirmatively, I approach this question by drawing on perceptual dogmatism, since this view seems, at first blush, most likely to allow that their justification is at least defeasible. I argue, that, despite appearances to the contrary, such delusions are completely unjustified, even by the standards of perceptual dogmatism.

6.1. Introduction

People with monothematic delusions make claims that seem at best unlikely and at worst patently absurd. In the Cotard delusion, one asserts that one is dead, that one is missing internal organs, or that one's body is rotting away. In the Capgras delusion, one says that a loved one has been replaced by an imposter. In the Fregoli delusion, one reports that several unfamiliar individuals are actually the same familiar individual in disguise.[48]

Delusions are standardly characterised in epistemic terms, as beliefs that are unsupported by the available evidence and maintained despite counterevidence (American Psychiatric Association, 2013).[49] Although agreement prevails that delusions involve malfunctions somewhere in the processes leading up to belief fixation, opinions differ on where exactly these malfunctions are located (for an overview, see e.g., Bortolotti, 2018).

An influential view in philosophy and cognitive science, originally put forward by Brendan Maher (1974), is that delusions arise as subjectively adequate responses to highly unusual experiences, such that the content of the former is intelligible in terms of the content of the latter. Call this view 'empiricism' (Campbell, 2001). Maher's distinctive brand of empiricism has been dubbed 'explanationist' (Bayne and Pacherie, 2004) because it conceives of delusions as generated by normal inferential processes to explain (a relatively unspecific)

[48] It is commonplace in considering delusions to draw a distinction between so-called *monothematic* and *polythematic* ones. For more, see Coltheart (2013). Monothematic delusions revolve around a single theme, whereas polythematic concern multiple themes. 'Delusions' henceforth refers only to monothematic delusions.
[49] I will assume throughout this paper that delusions are beliefs, because this is also assumed in the strand of the literature that I focus on. For arguments in support of the doxastic status of delusions, see e.g., Bayne and Pacherie, 2005; Bortolotti, 2010, 2012.

experience. On Maher's account, for example, subjects with Capgras initially arrive at misidentification beliefs like 'that's not Mum' as a way to explain the fact that their mother's face feels oddly unfamiliar (Ellis and Young, 1990). A second inferential step is required to derive the impostor belief from the notion that (as subjects believe is true) the person before them is really a stranger who looks just like her.

An alternative version of empiricism, known as the 'endorsement' model (Bayne and Pacherie, 2004; Aimola Davies and Davies, 2009), is that delusions are beliefs that derive directly, or non-inferentially, from perceptual experience. The idea here is that one has a perceptual experience with content that *p*, where *p* is the delusional content or something very close to that. What is needed to lead from experience to belief is that one should treat *p* as veridical or endorse it. Applied to the Capgras delusion, the endorsement model, in its most usual form, is profitably factorised into two distinct theses: (i) a subject *S* has a perceptual experience representing that *x* (who happens to be *S*'s mother) is not her mother; (ii) this experience is accepted as veridical and *S* comes to believe that *x* is not her mother. Here too, the impostor belief occurs through inferential elaboration based on the observed visual similarity between *x* and *S*'s mother (cf. Chapter 2). But it is important to recognise that the initial misidentification belief is non-inferentially grounded in perceptual experience.

It has been suggested that some cases of delusions conform more naturally to the explanationist model and others to the endorsement model, with most being hybrid combinations of explanationist and endorsement processes (Langdon and Bayne, 2010). Here I am interested in the implications that would hold if the endorsement model (henceforth abbreviated by EN) were true, for reasons that I will now explain. The main feature that distinguishes EN from the explanationist model is the account of delusions as *perceptual* beliefs rather than as *inferential* beliefs. By 'inferential', I mean beliefs whose fixation is via inference from perception or previously acquired beliefs. By 'perceptual', I mean beliefs that satisfy the following description, which captures something like the default view among epistemologists (cf. Lyons, 2009, p. 87): one's belief that *p* is a perceptual belief if and only if one believes that *p* because one has a perceptual experience with the content that *p*.[50]

[50] Whether perceptual experiences have propositional content is a controversial issue in the philosophy of perception. That means any representationalist theorist who denies that perceptual experiences have propositional content, or any non-representationalist theorist who denies that perceptual experiences have content, would reject this conception of perceptual belief. Since my focus in this article is on EN, and since EN rests on the assumption that perceptual experiences have propositional content, I will pass over these matters here.

Perceptual beliefs are often accorded a privileged, if not exclusive, status in epistemology, because they are said to enjoy immediate prima facie justification, of which more below. For now, suffice it to say that *immediate* justification is justification that does not depend on the justification of other beliefs, whereas *prima facie* justification is defeasible justification that becomes *ultima facie* provided one has no countervailing evidence. Suppose you are looking at the snow pile up outside your window and form the belief that it is snowing. Your belief that it is snowing seems an excellent candidate for being immediately prima facie justified, since its justification appears both defeasible (you might conceivably be hallucinating) and independent of other justified beliefs.

This opens up the possibility that at least some delusions (namely, those formed solely via endorsement processes) have the same justificatory status. Since endorsement theorists claim EN is most suitable to misidentification delusions (henceforth abbreviated as MDs) (Langdon and Bayne, 2010), I will be concerned here with one such case, namely the Capgras delusion.[51] In a slogan: if one's perceptual experience represents that the person confronting one is not the person one knows, then it provides one with immediate prima facie justification for believing that content.

The central question for this paper is whether this idea can be parlayed into a sound argument. Whatever answer is given has wide-ranging consequences. Let *d* be the delusional proposition 'that [currently perceived] person is not that [familiar] person'. If at t_1 your experience *E* provides immediate prima facie justification for believing *d*, then at t_1 *E* gives you good reason to believe that *d*. Suppose now that, at t_1, you form the belief that *d* on the basis of *E*. It follows that your belief is held for at least one good reason. This reason may cease at t_2 to be a good reason (or not, in which case you have ultima facie justification for *d*). If, on the other hand, *E* does not even provide prima facie justification for *d*, then *E* gives you no good reason to believe that *d*. Accordingly, a belief that *d* held solely on the basis of *E* would be unjustified.

There are two broad ways that one might try to support the claim that MDs are immediately prima facie justified by experience. First, one might claim that a perceptual experience *E* with the delusional content that *p* justifies believing that *p* in virtue of some externalist conditions, which come apart from the individuation of *E*'s phenomenology.

[51] Throughout the paper I work under the assumption that MDs arise from endorsement processes. Although my main focus in what follows will be on the Capgras delusion, I believe my arguments apply, mutatis mutandis, to all other MDs.

Well-known, natural examples include, but are not limited to, *reliability* (*E* justifies believing that *p* only if it arises from a truth-conducive process), *veridicality (E* justifies believing that *p* only if it veridically represents that *p*), and *aetiology* (*E* justifies believing that *p* only if it has the right kind of causal history).[52] Since MDs are arguably based on non-veridical perceptions, which in turn, are the effects of dysfunctional and unreliable mechanisms, they are very unlikely to satisfy the foregoing externalist conditions.[53] Second, one might claim that a perceptual experience *E* with the delusional content that *p* justifies believing that *p* in virtue of *E*'s phenomenology alone. The theory of perceptual justification that underpins this view I will call *perceptual dogmatism*, PD for short. According to this theory, so long as you have a perceptual experience with the content that *p* in combination with the appropriate sort of phenomenology, you thereby have immediate prima facie justification for believing that *p*.[54] Note that PD per se is not committed to what Elijah Chudnoff has dubbed 'phenomenalism' (2013, p. 88), for which if your perceptual experience justifies you in believing that *p* it does so because, and only because, of its relevant phenomenology.

PD is more likely than externalist theories to allow that a subject may have immediate prima facie justification to believe MDs, as it does not require for justification that the source of belief be truth-conducive. It only says that having a *p*-experience with the right kind of phenomenology relative to *p* is sufficient for immediate prima facie justification of the belief that *p*. Perhaps for this reason, EN is sometimes regarded as expressing a dogmatist epistemology, where the idea is that delusions are prima facie justified given the experiences

---

[52] I borrow this classification from Smithies (2019). For literature in defence of the reliability constraint, see Goldman (1979), Sosa (1991), Burge (2003) and Lyons (2009). For literature in defence of the veridicality constraint, see McDowell (1995), Williamson (2000), Littlejohn (2009), Pritchard (2012), and Schellenberg (2013). For literature in defence of the aetiology constraint, see Siegel (2012, 2017), McGrath (2013, and Teng (2018).

[53] According to reliabilism, what makes justified beliefs justified is their relation to a process (or indicator) that is reliably conducive to producing true beliefs. Whether a process is deemed reliable depends on the ratio of true to false beliefs it produces. One difficulty is that every *token* belief-forming process is an instantiation of different process *types* that vary widely as to their degree of generality and reliability. The question then arises: which process type has to be reliable in order to make the resulting belief justified? The latter is known as the Generality Problem for reliabilism (Conee and Feldman, 1998). Now consider the putative process of belief formation in the Capgras delusion—say, *S*'s coming to believe, upon perceiving that some person *x* is not her mother, that *x* is not her mother. Did *S* form this belief by means of a reliable process? If we consider the broadest process type that was instantiated by the token process, namely, perception, we might be tempted to answer in the affirmative, for perception is a reliable source of belief. However, if we consider narrower process types, i.e., the process of relying on abnormally diminished responsiveness to familiar visual stimuli (Ellis et al., 1997; Ellis et al., 2000), we will find that *S*'s belief was formed in a highly unreliable fashion. This makes it difficult to see how such a belief could count as justified by reliabilist lights, since it is unclear why reliability at a broad level should offset unreliability at the narrower level.

[54] Notable advocates of PD include Pryor (2000, 2004), Huemer (2001, 2006, 2007), Tucker (2010, 2013), Brogaard (2013, 2016), Bengson (2015), Chudnoff (2011, 2012, 2013, 2016), Chudnoff and Didomenico (2015).

they are endorsements of (Gibson and Bradley, forthcoming). In what follows, however, I argue that the perceptual experiences on which MDs are based do not even prima facie justify, not even by the lights of PD, because they lack the right kind of phenomenology. The more general point is this: even if MDs might be formed via endorsement processes, i.e., in the manner typical of perceptual beliefs, that does not guarantee that they are epistemically well-formed. Therefore, endorsement interpretations of MDs should not be taken to indicate commitment to their prima facie justifiedness.

Here is the plan for the paper. I first more fully characterise PD. I next consider two notions of justification-conferring phenomenology which have surfaced in the literature. According to the first notion, for every perceptual experience with content *p*, if it has assertive force with respect to *p*, then it gives you immediate prima facie justification for *p*. According to the second notion, for every perceptual experience with content *p*, if it has presentational force with respect to *p*, then it gives you immediate prima facie justification for *p*. I concede that the former (and only the former) is compatible with the aetiology of MDs: the experiences in question may have assertive force with respect to misidentification propositions like 'This person here present is not (the same person as) Mum', but necessarily lack presentational force with respect to such propositions. On its face, this leaves it open that at least the assertive force enjoyed by the deluded subject should help confer immediate prima facie justifiedness to MDs. However, I argue that, on balance, we should reject the view that assertive force is sufficient for immediate prima facie perceptual justification. Assertive force is supposed to explain why perception has a distinctive epistemic role that marks it out from beliefs, namely, its capacity to prima facie justify without standing in need of justification. Since there are cases where, arguably, beliefs have assertive force, it is unclear why those beliefs could not be just like perceptual experience in conferring prima facie justification without being justified themselves. I consider the prospects of biting the bullet and accepting that assertive force grounds in beliefs the same justificatory power it grounds in perceptual experience. Yet, I contend, this would lead to an implausible expansion of the class of propositions that one is immediately prima facie justified in believing. I close by clarifying some implications of this argument for the relationship between EN and PD.

6.2. Perceptual Dogmatism

Mia is hospitalised after sustaining a head injury that disrupted the heightened affective responsiveness normally elicited by the sight of a familiar individual. When she sees her mum

at the hospital soon after her accident, the sensory input does not initiate the correct emotional response. Consequently, she has a perceptual experience as of the visitor not being her mum.[55] Is Mia immediately prima facie justified in holding the belief that the visitor is not her mum?

The answer depends on one's theory of justification. Were justification to depend on the satisfaction of externalist conditions, the answer would most likely be *No*, since MDs do not seem to be the sort of thing that are formed via a reliably truth-conducive kind of cognitive process. If one is a dogmatist about perceptual justification, the answer is possibly *Yes*, but the warrant for that answer would have to be looked for in the phenomenology of Mia's experience. Before we turn to this issue (in Sections 6.3. and 6.4.), we must say more about PD.

PD can be factored into two distinct theses (adapted from Teng, 2018, p. 638):

> **The Immediacy Thesis**: having a perceptual experience (*E*) with *p* as a content provides immediate prima facie justification for believing that *p*.
>
> **The Sufficiency Thesis**: having the right kind of phenomenal character is sufficient for any experience (e.g., perceptual, memorial, imaginative, etc.) to confer immediate prima facie justification on belief.[56]

Several clarifications should be noted. To begin with, the immediacy thesis and the sufficiency thesis are separable: one might accept the former without the latter; that is, one might accept that perceptual experiences provide immediate prima facie justification, while denying that they do so in virtue of their phenomenology alone (Goldman, 2008; Lyons, 2009).[57] Furthermore, the sufficiency thesis should be kept separate from the notion that having a distinctive phenomenology is the *only way* that any experience can immediately prima facie justify one in believing something (as in phenomenalism).

Second, proponents of PD think of perceptual experiences as states with truth-evaluable propositional contents. They thus assume that it is possible for beliefs and

[55] Although this is a fictional example, it closely resembles real-life cases of Capgras delusion (e.g., Ramachandran and Blakeslee, 1998; Brighetti et al., 1997) and it comfortably fits into standard accounts of its formation (Ellis and Young, 1990; Ellis et al., 1997; Ellis et al., 2000).

[56] Most of those who identify themselves as 'dogmatists' (see fn. 54 above) endorse both theses (there are exceptions though, see Silins, 2008).

[57] Here and in what follows, I will use 'in virtue of…' to indicate logical sufficiency rather than a relation of causal/explanatory relevance.

perceptual experiences to share the same contents. These contents are not about sense-data or subjective qualities of experience; they are about observables (e.g., objects, properties, or events) in the world (Pryor, 2000; Moretti, 2015). It is open to debate which sorts of propositions your perceptual experience represents, and accordingly, gives you justification for believing.

Suppose you are watching an apple rolling along your desk. Does your perceptual experience represent that there is an apple in front of you? Or does it represent that there is a red thing of a certain shape and texture in front of you? For the sake of discussion, I will assume that perceptually represented propositions do not only concern shapes, textures, and shades of colour, but also entry-level categories (e.g., apples, cars), causal relations, and more subordinate level categories (e.g., Granny Smith, BMW), including personal identity (e.g., Mum, Bill).[58]

Third, some philosophers have adopted a *disunified* view of perception, according to which your overall 'perceptual' state consists of two separable components: a *sensory* component and a *cognitive* response to that component.[59] I will use 'sensations/sensory experiences' to pick out the former and 'seemings' to pick out the latter, although different philosophers use different terminology to mark out this distinction. On this view, for example, when you see a bird flying past, you have (i) a nonconceptual sensory experience of a dark shape in motion and (ii) an accompanying seeming which categorises the sensory information and tells you that it is a bird, it is a crow, etc. For defenders of the disunified view, PD is a thesis about perceptual experience conceived as the overall state which comprises sensations and seemings.

Fourth, PD claims that some perceptual justification is prima facie and immediate. It is prima facie in the sense that it can be defeated (overridden or undermined) by further evidence (Pollock, 1974, pp. 39-46). If I seem to see you jogging down the street, and then your wife tells me that you were sick in bed all day with the flu, that constitutes an *overriding* defeater for my apparent experience. If I seem to hear whirring skate wheels in my living room, and then I learn that I have been slipped a hallucinogenic drug, that constitutes an *undermining* defeater for my apparent experience. In both cases, your apparent experiences

[58] For helpful discussion of arguments for and against allowing perception to have high-level contents, see papers collected in Hawley and Macpherson, 2011, and Siegel, 2010, 2017.
[59] For defence of the disunified view, see Brogaard (2013), Lyons (2005, 2009, 2015), Reiland, (2014, 2015), Tucker (2010). For criticism, see Chudnoff and Didomenico (2015).

provide you with prima facie justification for the relevant beliefs. However, the defeaters prevent the prima facie justification from becoming ultima facie justification.

Moreover, perceptual experience is sometimes a source of immediate justification. To illustrate, it is useful to draw a contrast between *mediate* and *immediate* justification. A source gives you mediate justification if and only if it gives you justification in a way that depends on your having justification to hold other beliefs. For example, you have justification from your experience to believe that Fiona dyed her hair blonde only conditional on your independent justification to believe that Fiona was brunette. A source gives you immediate justification if and only if it gives you justification in a way that does not depend on your justification for other beliefs. Think about how your experience plausibly can justify your belief that Fiona looks blonde without reliance on any further beliefs.

Fifth, PD can be (and has been) interpreted as an internalist version of modest foundationalism. Roughly, epistemic internalism holds that whether you are justified in believing *p* supervenes on factors that are internal to your mental life (*mentalism*) or which are accessible by reflection (*accessibilism*).[60] PD is internalist at least in the mentalist sense that consciousness suffices for justification. In addition, PD is modestly foundationalist in that it affirms that there is a class of perceptual beliefs about the external world whose (prima facie) justification does not depend on other justified beliefs. It is here that PD has much of its appeal. By allowing that we can be justified by our perceptual experiences alone, the dogmatist halts the regress of justifications that would otherwise lead to scepticism.

Finally, epistemologists usually assume a distinction between having justification for believing a proposition (i.e., *propositional* justification) and justifiably believing that proposition (i.e., *doxastic* justification). The former obtains if you have justification that supports believing *p*, regardless of whether you do so believe. The latter obtains if you are propositionally justified in believing *p* and you properly base your belief on your propositional justification to believe *p*.

Suppose that both you and I see a kid shoplifting a sandwich at a grocery store. You believe him guilty on the basis of your visual experience. I believe him guilty because he looks suspicious. We both have propositional justification from our experience to believe the kid is guilty. But only your belief is doxastically justified because only you based your belief on that which propositionally justifies it. PD is typically construed as a claim about propositional

[60] The terms 'mentalism' and 'accessibilism' were first used by Conee and Feldman (2001). For more discussion on different versions of internalism see Pryor (2001).

justification. Thus, from now on, when I use the term 'justification', I will be speaking of propositional rather than doxastic justification.[61]

6.3. Justification-Conferring Phenomenology

Recall the question raised at the start of Section 6.2. Assuming that Mia experiences her visitor as not her mom, does it thereby follow that Mia has immediate prima facie justification for believing that the visitor is not her mum? For the dogmatist, the answer lies in what it is in virtue of which perceptual experience delivers immediate prima facie justification. There are two main proposals available in the literature, each of which picks up on a particular type of phenomenology (Teng, 2018). On the first view, it is sufficient for perceptual experience to justify your believing *p* that it has *assertive* force with respect to *p*. On the second view, it is sufficient for your perceptual experience to justify your believing *p* that it has *presentational* force with respect to *p*.

In this section, I look at each of these in turn. Before doing so, however, we should note two desiderata for any account that grounds prima facie justificatory immediacy in the phenomenology of perceptual experience. One desideratum is to explain how perceptual experience can be unlike mental states which are standardly seen as unable to confer justification without themselves being justified (e.g., beliefs), to do any epistemic work at all (e.g., desires), or to play any role in the justification of non-modal external world beliefs (e.g., deliberate imaginings).[62] Let us call this the *distinctiveness* desideratum (following Ghijsen, 2014). The other desideratum is to explain why having the relevant phenomenal force would make experience have prima facie justificatory immediacy. Let us call this the *epistemic significance* desideratum.

To a first approximation, assertive force is a way of representing contents about the external world. When an experience represents its content assertively, it does not simply represent it as true; it represents it in a way that assures us of its truth. One finds this notion expressed, albeit under different labels, in the works of various writers. Jim Pryor describes it in terms of 'the feeling of *seeming to ascertain* that a given proposition is true' (2004, p. 357). He says that perceptual experience is such that it makes it feel as though you can just tell that

[61] To ease the presentation below, I will sometimes use expressions (e.g., 'justified belief', 'justified in believing') that lend themselves most naturally to an account of doxastic justification. The reader should bear in mind that even when I use such expressions I will be focusing on propositional justification.
[62] For an exception, see Kind (2018).

its content obtains. So, for example, if you are having an experience of something red, that experience makes it feel as though you can just tell that there is something red.

Other writers think of assertive force as a property that pertains specifically to seemings. William Tolhurst writes, 'seemings have the feel of truth, the feel of a state whose content reveals how things really are' (1998, p. 299). Chris Tucker says that when it perceptually seems to you that *p*, you have an experience that feels as though it 'recommends' its content as true or 'assures' you of the content's truth (2010, p. 530). Michael Huemer argues that seemings represent their contents *as being actualised,* where the term 'actualised' is intended to capture the property of being the case with respect to the world (2001, pp. 77-79).

The 'assertiveness' account of perceptual justification, as we might call it (following Teng, 2018), purports to satisfy both of the above desiderata. On the one hand, it supposedly captures the distinctive phenomenal character of perceptual experience. When you consciously see a red balloon ahead, your experience assures you that there is a red balloon ahead. This, however, does not happen when you deliberately imagine or desire that there is a red balloon ahead. Likewise, in believing or judging that there is a red balloon ahead, you (at least allegedly) lack any such feeling of assurance. Or, if you have it, you do only derivatively so—derivative of your visual experience. This is not to say that beliefs and judgments are not, in some sense, assertive. The idea is that experience involves a feeling of being *struck* by the truth of something, as opposed to merely asserting that something is true (Cowan, 2018, p. 223).

On the other hand, this (alleged) phenomenal difference between perceptual experience and other mental states promises to explain why only the former deliver immediately prima facie justified beliefs. That is because only the former represents its contents in such a way as to make one feel assured of their truth. It is by virtue of this feeling of assurance that perceptual experience can serve as a regress-stopping state that gives justification without needing it in turn.

An alternative but, in principle, complementary account of why perceptual phenomenology is justification-conferring is offered in recent work by Elijah Chudnoff (2012, 2013, 2016). According to Chudnoff, an experience's capacity to immediately prima facie justify belief stems from its presentational force, where this is conceived as a correlation between two kinds of phenomenal properties. He writes:

> What it is for an experience of yours to have presentational [force] with respect to *p* is for it to both (1) make it seem to you that *p* and (2) make it seem to you as if this experience makes you sensorily (e.g., visually) aware of a truth-maker for *p* (Chudnoff, 2013, p. 37).[63]

In order to better appreciate Chudnoff's view here, it will help to unpack a little his thinking with respect to each of the properties above, namely (1) 'seeming to you that *p*' and (2) 'seeming to be, or feeling as if you are, aware of a truthmaker for *p*'. Let us take these in turn. (1) simply picks out the experience's property of having a representational content *p*. To say that an experience makes it seem that *p* is just to say that the experience represents that *p*. Suppose you look in front of you and see a red balloon flying away. So long as your visual experience represents that there is a red balloon over there, it also seems to you there is a red balloon over there. (1) by itself does not imply assertive force, although of course one's perceptual experiences can and often do represent their contents assertively.

While (1) is a matter of having a perceptual experience representing that *p*, (2) is a matter of seeming to be sensorily aware of an item (e.g., individual, event, property) that serves as a truthmaker for *p*. The first thing to note is that seeming to be sensorily aware of an item is not the same as being sensorily aware of an item simpliciter. The former is a purely phenomenal property, the latter depends on extra-conscious conditions, such as that the intended item really exists. When you see a red balloon in good light, you stand in a sensory awareness to a mind-independent object (i.e., the balloon) and a property that this object instantiates (i.e., redness). This is not necessarily so with seeming sensory awareness, since you can seem to see a red balloon even if there is no red balloon for you to see. It should also be clear that (1) and (2) report on two distinct phenomenal properties. Say the red balloon is enveloped by mist. You might visually represent that there is a balloon and that is red: imagine glimpsing a reddish speck through the mist. Nonetheless, you do not see a red balloon, nor does it seem to you as if you see one. Secondly, (2) makes use of the notion of a truthmaker for a proposition. Chudnoff does not have much to say about just what a truthmaker is, but he seems to have in mind something like the following principle (cf. Armstrong, 1997): *x* is a truthmaker for a proposition *p* if and only if *x* exists and the

[63] Two things are noteworthy here. First, Chudnoff talks of 'phenomenology' rather than (as I do) 'phenomenal force'. I have slightly altered his terminology so that it meshes comfortably with the terminology that I have been using. Second, and more importantly, Chudnoff intends the content of any *p*-experience with presentational force to be self-referential, namely, such that it makes it seem to you as if *this experience* makes you aware of a truthmaker for *p*. Still, for ease of explanation I will mostly use the shorter phrase 'makes it seem to you as if you are aware of a truthmaker for *p*'.

existence of *x* entails the truth of *p*. In this sense, for example, the redness inhering in a red balloon is a truthmaker for the proposition that the balloon is red.

Once again let us take an illustrative example to fix ideas. Suppose you have a visual experience representing that a balloon is red. Suppose further that you seem to be visually aware of a red balloon just by having the experience. Your experience here is one in which it seems to you that *p*—a balloon is red—and in which you seem to be aware of an item—a red balloon—that makes *p* true. What presentational force requires is that these two phenomenal properties be so correlated in experience: in addition to representing that *p*, an experience must 'be felt as making you aware of the chunk of reality' that makes it the case that *p* (Chudnoff, 2016, p. 288).

On the resulting view of perceptual justification, if a perceptual experience has presentational force with respect to *p*, then it immediately prima facie justifies believing that *p*. As it turns out, this is quite a demanding view. Consider a case where perceptual experience allegedly lacks presentational force (adapted from Markie, 2005, pp. 356-7). Suppose you are an expert gold miner panning for gold. After all the lighter rocks are washed out of your pan, your eyes fall on a yellow pebble, which happens to be a gold nugget. As you look at the pebble, you have a visual experience that represents it as being gold. Imagine that you form the correct belief that the pebble is gold on the basis of your visual experience. Are you thereby immediately prima facie justified in holding that belief? Chudnoff thinks not, because he claims you cannot seem to see a truthmaker for the proposition that the pebble is gold. When you visually experience a gold pebble, you seem to see the yellowness of the pebble, perhaps even its golden colour, but the gold of which it is made does not itself seem to be seen. For Chudnoff, this would involve seeming to see something like the molecular make-up in virtue of which the pebble is gold, which is implausible (Chudnoff, 2013, p. 91).

Lurking here is the more general issue of the admissible contents of perceptual experience. Chudnoff does not reject the possibility of high-level perception. But this needs qualification. To that end, it helps to distinguish between propositions we seem to perceive to be so full stop, and propositions we seem to perceive to be so only derivatively, in virtue of seeming to see that other propositions are so (Pryor, 2000, pp. 538–539). Using Pryor's terminology, we may call these propositions 'basic' and 'non-basic' respectively. Chudnoff takes the latter distinction to be coextensive with the distinction between low-level contents (i.e., propositions featuring low-level properties) and high-level contents (e.g., propositions featuring high-level properties). He says that perceptual experience does represent high-level

contents, such as that the pebble is gold, but only *basically* represents low-level contents, such as that the pebble is yellow. Moreover, he adds, perceptual experience has presentational force only with respect to their low-level contents (Chudnoff, 2013, Ch. 1). The reason he gives is that, since high-level properties (e.g., being gold), unlike low-level ones (e.g., being yellow), cannot be detected by visual transducers, we are not, nor do we even seem to be, visually aware of them.[64]

Before concluding this section, there is one final point that is worth taking up. At least at first pass, Chudnoff's account fares no worse than the assertiveness account with respect to the desiderata sketched above. As for the distinctiveness desideratum, what (allegedly) sets apart perceptual experiences from other mental states, like beliefs and deliberate imaginings, is their presentational force. Chudnoff concedes that some cases of deliberate imagination may appear to have presentational force. One such case is that of a person who propositionally imagines that *p* ('there is a tiger'), while at the same time objectually imagining *F*, where *F* is something that makes it true that *p* (the tiger itself).

Chudnoff's response to this is that objectually imagining an *F* does not make it seem as if it makes you aware of an *F*. He argues roughly as follows: (1) If an experience seems to make you aware of an *F*, then *F* is represented as actual—as really being there. (2) It is possible to objectually imagine an *F* without representing it as actual. Therefore, (3) objectually imagining an *F* is not a way to seem to be aware of an *F* (Chudnoff, 2012, p. 92).[65] By the same reasoning, you lack presentational force in believing that *p*, since your belief that *p* does not seem to make you aware of truthmaker for *p* (again, unless one assumes that seeming is derivative from your visual experience).

As for the epistemic significance desideratum, we have already seen that a perceptual experience immediately prima facie justifies you in believing that *p* in virtue of having presentational force with respect to *p* (Chudnoff, 2012, p. 66). On Chudnoff's account, it is not sufficient for an experience to halt a justificatory regress that it assures you of its contents'

[64] Some care is needed here. Chudnoff seems to identify low-level properties with those that can be directly transduced by the sensory modalities. There are obvious candidates (e.g., colour, motion, shape, texture), but beyond the obvious candidates, it is very much a matter of dispute. Indeed, Chudnoff treats as visually detectable properties that are often grouped together as 'high-level', like, for example, natural kind properties (e.g., being a tiger, see below). Accordingly, he thinks propositions such as 'there is a tiger' are contents with respect to which your experience can have presentational force, for you can seem to be visually aware of something being a tiger.

[65] This argument has been challenged (Ghijsen, 2014; Teng, 2018). The most straightforward challenge is to point out that although you can imagine *F* without representing it as actual, you can also imagine *F* and represent it as actual, in which case (3) does not follow (Ghijsen, 2014 p. 1556).

truth. It is also necessary that it makes you seem to be aware of the very items in virtue of which such contents are true.

So far, we have considered two dogmatist accounts of what makes perceptual phenomenology justification conferring. Settling which of these accounts is superior is well beyond the scope of this paper, and so I will remain neutral on that question. The question that will concern us in what follows is whether either of these makes MDs justified. To anticipate my answer here, my main contention will be that, in a situation where a Capgras subject, Mia, looks at the misidentified person, her mum, Mia's experience may have assertive force, but does not have presentational force, with respect to the proposition that the person is not her mum. So, if we accept the assertiveness account of perceptual justification, then we may accept that Mia would be immediately prima facie justified in believing that the person standing at her bedside is not her mum. However, there are reasons to be wary of drawing this conclusion. Despite assertions to the contrary by its supporters, the assertiveness account fails by the lights of both our above desiderata.[66] Or so I will argue.

6.4. Are MDs Immediately Prima Facie Justified?

In this section I explore the chief motivation for presuming that some delusions, MDs, are justified by perception. This motivation is fleshed out by means of two theses.

> **Misidentification Psychological Thesis**: some of our perceptual experiences of other people misrepresent them as being not who they are or as being who they are not.
>
> **Misidentification Dogmatist Thesis:** if one has a perceptual experience (*E*) as of another person being not who she is, or being who she is not, and (*E*) has the right kind of phenomenology with respect to its content, then (E*)* immediately prima facie justifies believing that person is not who she is, or is who she is not.[67]

By the Misidentification Psychological Thesis, the delusional hypothesis is already part of the representational content of Mia's perceptual experience. That is: Mia has a perceptual experience that makes it seem as if her mum is not really her mum. By the Misidentification

[66] Note that I am sticking with my neutrality regarding which account of a justification-conferring phenomenology is correct. The fact that the assertiveness account is found wanting should not be read as implying the superiority of Chudnoff's account over the latter.

[67] I framed these theses in terms of people/persons, but they equally apply to objects, things like limbs and places.

Dogmatist Thesis, her belief that her mum is not really her mum is immediately prima facie justified by her perceptual experience, provided this has the right kind of phenomenology. The question we have to ask, then, is: does it?

6.4.1. (Presentational) Dogmatist Thesis

Let us begin by considering the following variant of PD:

> **(Presentational) Dogmatist Thesis (PDT):** if one has a perceptual experience (*E*) with the content that *p*, and (*E*) has *presentational force* with respect to *p*, then (*E*) immediately prima facie justifies believing that *p*.

I will argue that PDT does not hold for experiences like Mia's and so presentational force cannot be appealed to justify her belief. Put in premise-conclusion form, my argument is this:

> (P1)**:** If a perceptual experience has presentational force with respect to *p* it both makes it seem to one that *p* and makes it seem as if you are sensorily aware of a truthmaker for *p*.
>
> (P2): Mia's perceptual experience can at best make it seem as if the relevant person is not her mum, but it cannot make it seem to her as if she is visually aware of a truthmaker for the proposition that the relevant person is not her mum.
>
> Conclusion: Perceptual experiences like Mia's lack presentational force.

Grant (P1) as true by definition. (P2) is the key premise on which the argument turns. The conclusion follows from (P1) and (P2) and is a rejection of PDT applied to Mia's experience.

The defense I offer of (P2) runs as follows. Suppose Mia has a brother named Paul. Both Mia and Paul know their mum by sight. It is plausible that when Mia and Paul look at their mum from the same angle in the same lighting conditions, then they could enjoy the same visual appearance. What visually appears to Mia also visually appears to Paul. However, it is also plausible that the seen person only seems to be Mum to Paul, whereas it could seem to Mia that the seen person is not Mum. If this is correct, then Paul's experience makes it seem as if it makes him visually aware of a truthmaker for the proposition 'that's Mum' (namely, the object Mum). By contrast, Mia's experience does not make it seem as if it makes her visually aware of a truthmaker for the proposition 'that's *not* Mum'. Why? because her

experience too makes it seem as if it makes her aware of Mum, and Mum is not a truthmaker for 'that's *not* Mum'.

To drive the point home, however, we need to say more about Mia's experience, and more generally, about the sorts of experiences to which people with Capgras are responding. Remember that there should be a distinction between (a) experience making you visually aware of a truthmaker for *p* and (b) experience making it seem as if you are visually aware of a truthmaker for *p*. (a) is about what your experience visually relates you to, while (b) is about how your experience makes things visually appear to you. Now, Mia sees a person and she is her mum. So clearly, her experience makes her aware of Mum. But this does not by itself settle whether her experience also makes it seem as if it makes her visually aware of Mum. It could, instead, seem as if it makes her aware of someone other than Mum, which would be a truthmaker for the proposition that the person seen is not Mum.

A closer look casts doubt on that supposition. One distinguishing feature of Capgras (and other MDs, as we will see later) is that the misidentified person visually appears as who she is, which means that misidentification occurs not because of, but despite, her appearance. This is explained empirically via appeal to a dual route model of visual face recognition. For instance, an influential account formulated by Ellis ad Young (1990) suggests that we recognise familiar faces using two neural pathways: a 'ventral route', subserving conscious recognition of facial identity, and a 'dorsal route', responsible for the appropriate emotional response to familiar faces. In Capgras, the ventral route remains intact, but the dorsal route is impaired, the result being that the familiar faces are recognised through vision, but provoke little or no emotional reaction. Going back to our example, this means that when Mia sees her mum, she experiences a mum-like visual appearance. She can tell that the person standing by her bed looks just like Mum, and that is because she seems to see Mum. However, due to the absence of emotional arousal, she is unable to identify her as such.

The upshot is that not only is Mia visually aware of Mum, it also seems to her that she is visually aware of Mum. If it did not, she would not be able to consciously recognise Mum's appearance, which she clearly does. Since Mum is a truthmaker for 'that's Mum', and it seems to Mia that she is visually aware of Mum, Mia's experience does not make it seem as if she is visually aware of a truthmaker for 'that's not Mum'. This gives us reason to believe P2. And, by the argument sketched above, if P2 is true, then Mia's experience runs afoul PDT.

6.4.2. (Assertive) Dogmatist Thesis

A second variant of PD can be expressed in the following manner:

> **(Assertive) Dogmatist Thesis (ADT)**: if one has a perceptual experience (*E*) with the content that *p,* and (E*)* has *assertive force* with respect *p*, then (*E*) immediately prima facie justifies believing that *p*.

If we grant (as per the Misidentification Psychological Thesis) that Mia experiences her mum as being not really her, then I say there is no reason in principle to deny that such an experience is assertive in the sense elaborated above. In other words, it is possible that when Mia is looking at her mum, she has a feeling of assurance that her mum is not really there. If so, then, by ADT, the experience Mia has can justify a corresponding perceptual belief to that content. Or, put more generally: a person with Capgras is immediately prima facie justified in believing that the observed individual is not who they appear to be.

This line of thought can be resisted by denying ADT. If we can show that having an experience with assertive force does not suffice to provide immediate prima facie justification at the baseline, then we can show that the conclusion does not follow. In the rest of this section, I set out an argument to this effect.

The key claim in ADT is that some perceptual experience (*E*) provides immediate prima facie justification, in a way that does not depend on any justification for other beliefs. While beliefs cannot contribute to justification without themselves being justified, (*E*) justifies without needing to be justified by anything else. By implication, whatever makes (E*)* confer immediate prima facie justification has to be something that (*E*) has and beliefs do not. According to ADT, that is assertive force.

As mentioned earlier, an obvious objection to this view is that beliefs too have assertive force, at least in the sense that they represent their contents as being true. If you believe that there is no God, it is not for you an open question whether there is a God, it seems true to you that there is none. Thus, the objection goes, assertive force alone cannot be what sets (*E*) apart from beliefs. In reply, 'assertive dogmatists', as we may call those who accept ADT, could argue that this objection conflates 'mere' assertiveness with what we may call 'felt' assertiveness. The former property pertains to any mental states whereby one takes their contents to be true, whereas the latter pertains to mental states which are felt as recommending their contents as true. Since beliefs appear to lack such a felt character, the

reply goes, they can be said to be assertive in the first sense, but not in the second sense. Upon reflection, however, it is not obvious why this should be so, as I will now attempt to show.[68]

One point of clarification should be emphasised before we move on. Let C_b be the claim that at least some beliefs have assertive force, and let C_p be the claim that perceptual experiences do. Assertive dogmatists do not provide any rigorous *argument* for C_p. They simply rely on the *intuition* that perceptual experiences are phenomenally conscious in a way that make their contents feel true. Thus, if a plausible case can be made for C_b, then the burden is on them to explain why C_p is acceptable but not C_b.

6.4.3. Beliefs with Assertive Force

Joe has long believed that sensible objects continue to exist when unperceived, although the belief has never explicitly occurred to him before and although he no longer recalls its original source. Joe's belief is not a priori guaranteed of truth, but neither can it be verified or falsified by sense experience. One might call it 'fundamental' in the sense it is not derived from other beliefs Joe has. Now imagine asking Joe whether his desk continues to exist when unperceived. Assertive dogmatists will want to rule out the possibility that once the belief (or content thereof) is brought to mind, it assures him (or, at least, makes him feel assured) of its content's truth. I see no good reason to rule out this possibility, for reasons that will become clear in a moment.

The broad construal of the notion of assertive force is motivated by an analogy between perception and testimony (Teng, 2018, p. 644). The analogy should tell us what makes it the case that perception and testimony assure us that *p*, while belief does not. So, our first step will be to assess whether the analogy is apt, and if so, to what extent.

Compare two scenarios. In one, I form the belief that it is raining on the basis of seeing water pouring down outside my window. In the other, I form the belief that it is raining on the basis of your having said so. Thomas Reid (1788) was perhaps the first philosopher to note that we credit the deliverance of our senses in a way analogous to that in which we credit the testimony of others. Just as we are naturally inclined to believe things as they appear, so we are naturally inclined to believe the things we are told. But what do perception and testimony have in common that makes us so inclined? One response is: a

[68] From now on, when I speak of 'assertive force' I will mean 'felt assertiveness' in the sense just defined.

representation-as-true that *p*. Just as speakers represent-propositions-as-true by, for instance, claiming or asserting that *p*, so too does perception. The idea would then be that we find ourselves naturally inclined to believe that *p* whenever we understand a representation-as-true that *p*.

Assertive dogmatists go further, however, and say that perception *assures* us that *p* (or at least we feel it does). As we shall see, the analogy to testimony breaks down with respect to assurance. Compare: the fact that my perception represents-as-true *p* gives me assurance regarding the truth of *p* vs. the fact that a speaker's testimony represents-as-true *p* gives me assurance regarding the truth of *p*. The disanalogy is easy to appreciate. Testimony requires an action on the part of the speaker to induce belief in the hearer. Strictly speaking, therefore, it is not the testimony in itself that offers the hearer an assurance of *p*'s truth, but the speaker. There is no unanimity on how to understand the speaker's assurance (see e.g., Schmitt, 2010, for an overview). But it is generally agreed that it is a matter of the speaker presenting herself in a certain way to someone who interprets her as doing so. Perhaps the most common way assurance has been conceptualised is as the speaker's assumption of responsibility for her testimony (e.g., Ross, 1986; Moran, 2018; Hinchman, 2005; Weiner, 2003). As Moran puts it, 'when someone gives me her assurance that it's cold out, she explicitly assumes a certain responsibility for what I believe' (Moran, 2018, p. 44). Admittedly, then, the sense in which a speaker assures us as to the truth of what she says is not the same in which perception assures us of its content's truth. For in the case of testimony, assurance that *p* ultimately depends on the responsibility the speaker assumes for *p*'s being true.

Now this may seem a rather trivial point for us to make—perceptions are not agents, and as such they cannot literally assure us of anything (Tooley, 2013)—but it is actually illuminating. We have seen that the analogy between perception and testimony works only to an extent. Unlike perception, testimony that *p* is an act of assurance, where *p* is something the speaker takes responsibility for in giving her assurance. One might think, then, that all perception and testimony have in common is that they represent their contents as true. If this is all there is to the sense in which perception is analogous to testimony, it does a poor job of distinguishing between perception and belief, as beliefs are equally analogous to testimony. It seems indeed to be a platitude that believing that *p* involves representing *p* to be true. That leaves two possibilities: either assurance that *p* requires more than just representing-as-true that *p*, or it does not. If it does, then we need an account of what would do the trick. If it does not, then belief assures in the same way as does perception.

Assertive dogmatists, we saw, have a ready answer: while your apparent perception that *p* involves a *feeling* of being assured that *p*, your belief that *p* lacks it, for beliefs have no phenomenal character. But there are at least two problems about this answer.

First, assertive dogmatists need to tell us what, if not its property of representing-as-true that *p*, would enable perception to make you feel assured that *p*. Until this is addressed, there is pressure to explain why belief, which also represents-as-true that *p*, would not make you feel assured that *p*.

Second, the claim that beliefs have no phenomenal character needs qualification. It is widely held that most of one's beliefs are not occurrent at any given time but stored in unconscious long-term memory.[69] This feature of belief serves adaptive ends, allowing organisms to retain information beyond acquisition without having to learn it each time anew. Moreover, it fits remarkably well with our commonsense intuitions, for we take beliefs to persist through changes in the stream of consciousness, and despite interruptions of consciousness, such as periods of dreamless sleep or time-gaps in which we simply do not attend to them. Virtually anyone would agree that non-occurrent beliefs lack phenomenal character. There is nothing it is like to believe that *water is H_2O* when you dreamlessly sleep, just as, presumably, there is nothing it is like to believe that *water is H_2O* when you are awake but busy looking for a parking space. Still, perhaps there is something it is like to believe that *water is H_2O* when the belief is consciously occurrent, for instance, when someone asks you about the chemical formula for water. However, this itself is far from unproblematic. Many philosophers have denied that beliefs, qua beliefs, can ever be consciously occurrent.[70] The argument is straightforward. If beliefs are to survive psychological discontinuity, they cannot themselves be occurrent manifestations in the stream of consciousness. Were this to be the case it would mean that they would cease to exist as soon as the occurrence does. But this seems implausible; it seems implausible that you stop believing *water is H_2O* when you are dreamlessly asleep, or your attention is occupied elsewhere. Therefore, there is no such thing as a consciously occurrent belief, strictly so called. Rather, beliefs are standing states that only exist unconsciously and without phenomenal character.

Even those who deny that beliefs are ever conscious often concede that beliefs are disposed to cause conscious episodes functionally akin to themselves, typically first-order

[69] See Goldman (1999) and (2011). See also Mandelbaum (2014), Bendana and Mandelbaum, forthcoming, and Strick et al., 2011.
[70] See Crane, 2013; Mandelbaum, 2014; Carruthers, 2017; Smithies, 2019.

judgments with the same content as one's beliefs, or second-order judgments that one has a particular belief (see Smithies, 2019, Ch. 4). This is supposed to explain why we are under the illusion that our beliefs are often conscious. What we call 'conscious beliefs' are no other than judgments as to what we believe or as to the fact that we believe it (Crane, 2013, p. 167). Beliefs may not issue in judgments directly, however. Another possibility is that beliefs cause *cognitive* feelings that act as intermediaries between the propositions a person believes and her judgments that those propositions are true. By 'cognitive' here, I mean feelings that can be used as a guide for judging whether or not *p* obtains (Clore and Parrott, 1994). For example, L. Jonathan Cohen (1992) has argued that believing something is being disposed to feel that it is true *p* (p. 4). When the disposition is actualised (e.g., when one is asked about water's chemical structure), one is supposed to feel it true that *water is* H_2O and judge accordingly.

Cohen's feeling it true that *p* might just as well be interpreted as a feeling of being assured that *p*. So, there is here a loose sense in which belief is felt as assuring one of its content's truth. But we need to be clear about what it could mean for a belief to be *felt* in any suitable sense. Clear understanding requires a distinction between what we might call *phenomenal character* and *phenomenal disposition* (Smithies, 2019; Schwitzgebel, 2002). The phenomenal character of a mental state is what it is like to be in that mental state. The phenomenal disposition of a mental state is its disposition to cause phenomenally conscious episodes in a person's mind. As we already saw, beliefs themselves do not seem to be the sort of thing that can have phenomenal character. They do not in the strict sense feel like anything, let alone as assuring truth. However, even if this is correct, it does not follow that they cannot have a certain phenomenal disposition to make it feel as something in appropriate circumstances.

Let us go back briefly to our initial example with Joe. Presumably, there is nothing it is like for Joe to believe his desk continues to exist while unperceived. Yet, arguably, it is possible that when Joe is asked something referred to by the proposition in question, that triggers activation of the disposition to feel assured of its truth. In this sense, to say that a belief is felt as assuring that *p* is just to say that one feels assured that *p* in virtue of having the belief when the occasion demands. Why think it is in virtue of having the belief rather than in virtue of whatever was the source of the belief? One reason to do so stems from the empirical observation that we (like Joe) are often bad at remembering the source of our beliefs.

Consider one more example of a belief based on a forgotten source. Three years ago, Jake read a scientific article suggesting that (*r*) second-hand smoke is hazardous to children's health. He then formed a belief in *r*. Today, he has no recollection of how or why he formed such a belief. Still, he holds firmly to it. Whenever he attends to *r*-related matters, he affirms *r*'s truth in thought.[71] It seems fair to say that Jake's belief is stored in his mind and remains in existence even when *r* is not actively attended to. However, information about its source was lost prior to storage into long-term memory. If that is so, it is unlikely that the source of his belief could exert any influence over Jake's current mental states, let alone make him feel assured that *r*.

This point is bolstered by a great deal of research in social psychology on a phenomenon known as the 'sleeper' effect (Kumkale and Albarracìn, 2004). Here, the idea is that a message from a low-credibility source has a greater persuasive impact after a delay from onset of exposure. Sometimes a persuasive message is presented with a discounting cue—information that seriously questions the credibility of the message (e.g., message: 'migrants rob young Britons of jobs'; discounting cue: source of message is *The Sun* newspaper). While recipients may initially discount the message because of its source, their confidence in its truth increases with the passage of time. The explanation is that, over time, recipients remember the message but forget where it came from, and as such, that it had been discredited. This supports the view that assertive force can, in principle, arise independently of the source of a belief.

A few remarks are in order.

1. The appeal to felt assurance comports well with the phenomenological fact that we seem to find ourselves with beliefs that we did not even know we had, but which immediately strike us as true under appropriate circumstances. It may well be, as some have suggested, that beliefs are individuated by a disposition to judge that *p* or that one believes that *p*. But this cannot be the whole story: judgment is a mental act, the taking on of a committal attitude toward a proposition, and as such, it does not quite capture the sense of being struck by the truth of a belief someone already has.

2. The hypothesis under consideration is that some beliefs have assertive force in the sense just canvassed, not that all do. It is possible for there to be transient weakly held beliefs whose dispositions themselves are weak (Alston, 1996). These beliefs will hardly carry much

[71] Similar examples are used by Goldman (1999, p. 280; 2011, p. 260–61).

felt assurance when 'brought to mind'. So, we should expect the domain of beliefs with assertive force to be a limited set of the totality of one's beliefs. The natural candidates for one such set would be beliefs in 'hinge' or 'framework' propositions, e.g., 'the world has existed for quite a long time', 'trees do not give birth', 'bones are not made of wood', etc. (Campbell, 2001; Wittgenstein, 1969). But the set need not be seen as necessarily being so restricted, as we will discuss more below.

3. One might object that 'feeling assured' that *p* in virtue of so believing is nothing other than a thought that *p* is true (Tuomela, 2000). In other words, so-called 'feeling assured' is not really a feeling at all. If this is correct, then the kind of assurance one receives from believing that *p* is not the same kind of assurance one receives from perceiving, since only the latter is felt in the relevant sense. My response to this objection is that it equally applies to C_p. Assertive dogmatists have nothing to say on why felt assurance would consist in anything else than the thought that the content of one's perception obtains. For all that they tell us, the assurance you have when (say) looking at an apple might just consist in the thought that your perception is veridical (see Ghijsen, 2014). C_p, therefore, does not escape the objection.

Let me pause to take stock. My goal has not been to argue that C_b is any more plausible than C_p. I merely hope to have convinced you that C_b is no *less* plausible than C_p. If am right, at least some beliefs may give rise to the same phenomenology that assertive dogmatists regard as sufficient for experiences to confer immediate prima facie justification. This (I think) raises some worries for ADT, to which I now turn.

6.4.4. Against ADT

I said earlier that ADT is in trouble unless it can account for the distinctive epistemic role it assigns to perceptual experience (satisfying what I called the *distinctiveness* desideratum). In more detail, the idea is this. Take a simple example. Suppose a subject *S* believes that *p* (*the grass is wet*), and that *if p then q* (*if the grass is wet, then it rained last night*), thereby forming the belief that *q* (*it rained last night*). It is hard to see how *S* could have a justified belief in *q* other than by being justified in believing *p* and *if p then q*. This then appears to be a case where propositional states must be justified in order to justify beliefs. Now consider propositional states such as desiring or deliberate imagining. It is quite clear that neither of these states can, at least typically, justify external world non-modal beliefs. To simply desire or deliberately imagine a proposition, i.e., 'my wallet is stuffed with bank notes' provides no justification to

believe it. One way to explain this is to suggest that there are no unjustified justifiers: anything that makes a belief justified must itself be justified. Assuming (as many do) that there are no justified mental states other than beliefs, this amounts to saying that only beliefs can justify other beliefs, and that perceptual experiences cannot justify perceptual beliefs.[72]

One need not to accept this conclusion to appreciate the challenge it poses to ADT. The challenge is to explain what makes perceptual experience different from beliefs, deliberate imaginings, desires, etc., in virtue of which it serves as an immediate prima facie justifier without needing to be justified itself.[73] Assertive dogmatists allegedly have an answer to this question: what sets apart perceptual experience from belief and other propositional states is that it makes you feel assured of its contents' truth. I have argued, however, that there are plausible cases in which subjects enjoy this feeling with respect to *p* just by virtue of having the relevant belief that *p*. The question naturally arises, then, whether this implies (by ADT) that beliefs in such cases can serve as immediate prima facie justifiers. On its own, the answer is *No*. ADT is a claim about experiences, not beliefs. All the same, assertive dogmatists are hard-pressed to explain why assertive force can ground immediate prima facie justificatory powers in experiences but not in beliefs. Their explanation seems to be that, unlike experiences, beliefs must themselves be justified before they can confer any justification. But this begs the question by assuming what is at issue, that beliefs cannot confer justification solely in virtue of their assertive force. So, at least until evidence to the contrary is provided, ADT is open to the possibility of beliefs being immediate prima facie justifiers.

Here I think assertive dogmatists might just bite the bullet, and insist that in fact it is not much of a bullet to bite. They might say: 'That's true, some beliefs have assertive force, and are thereby capable of conferring immediate prima facie justification. But such beliefs are not counterexamples to the conclusion we wanted: that having assertive force makes experiences confer immediate prima facie justification. They just force us to be more inclusive on what counts as immediate prima facie justifier. What our thesis should say is that, for any mental state (not just experiences) with the content *p*, if it has assertive force, it supplies immediate prima facie justification for *p'*. Call this revised thesis *ADT'*.

---

[72] This is a very rough sketch of the argument Jack Lyons offers against the 'experientialist' approach to perceptual justification, which holds that sensory experiences themselves (not just facts about them) can play a justificatory role. For more see Lyons, 2009, pp. 74–75.

[73] See Ghijsen, 2014, for a congenial discussion of much the same point.

What, then, could possibly constitute a counterexample to ADT'? A counterexample to ADT' would be a case in which someone believes that *p*, feels assured that *p* in virtue of so believing, yet lacks immediate prima facie justification for *p*. This is what I shall now try to provide. I said earlier that our beliefs in framework propositions (bedrock beliefs, for short) are particularly good candidates to have assertive force. Only rarely (or even never) do these beliefs come to the forefront of our minds, and when they do, we often fail to recall their original sources. For example, if you have long believed that things exist unperceived, you may have forgotten you ever learned this. In that case, when you consciously entertain the belief, you may be aware of no information stored in memory that would support it. Nevertheless, your answer to the question 'Do you believe that things exist unperceived' would most likely still be a confident 'Yes'. One explanation for this is that your confidence has its source in a feeling of assurance. It thus seems plausible to say that you feel assured, when the question arises, that things exist unperceived. If so, assertive dogmatists are committed to the claim that beliefs like this can immediately (albeit defeasibly) justify themselves without standing in need of further justification. This makes ADT' akin to a sort of *phenomenal* doxastic conservatism (cf. Chisholm, 1980), which says that you have immediate prima facie justification for any belief you find yourself with, provided it has assertive force and you have no reason to think that it is false.

The preceding might seem acceptable in the case of bedrock beliefs. After all, bedrock beliefs are standardly assumed to be justification-makers which do not require support from other beliefs and are exempt from doubt. So, we can see how having one such belief that *p* might give you immediate prima facie justification for *p*. The problem is that bedrock beliefs are not the only ones that are suitable candidates for assertive force. Other likely candidates are, inter alia, beliefs with which one self-identifies (i.e., beliefs that are central to one's self-image) such as that one is intelligent, competent, ethical, etc. These beliefs are notoriously difficult to relinquish; even once we encounter disconfirming information, we find it hard to rid ourselves of them. Some philosophers have diagnosed such perseverance as deriving from a need to protect one's self-image (Mandelbaum, 2019). This seems right; but is not there any more to be said? For how is it, one might wonder, that we routinely disbelieve propositions about what matters most to us (e.g., 'my partner loves me', 'I am healthy'), even if that threatens one's sense of self? One plausible explanation is that we do not simply care about the occurrent belief contents under disconfirmation, thereby clinging to them. To a greater or lesser extent, we feel assured of their truth. That is

why we cannot be swayed from belief in them, despite evidence or well-reasoned arguments to the contrary.

This makes room for possibilities whereby unrealistic positive self-evaluations have assertive force. Indeed, there is evidence that the majority of people see themselves not just as good as average, but as above average across a number of dimensions (Alicke and Govorun, 2005; Brown, 1986; Alicke, 1985). Compared to our peers, most of us believe we are more intelligent, more competent, more ethical, and so on. Now, these beliefs of ours are often derived from inadequate evidence. Consider, for example, the following case. Let *p* be the proposition that I am more intelligent than most of my colleagues. Imagine I come to believe that *p* on the grounds that I can solve the Rubik's cube. Imagine, further, that in virtue of having a belief that *p*, I am brought to feel assured that *p*. ADT' yields the result that I am immediately prima facie justified in believing *p*. But am I really? I think not. By definition, I am immediately justified in believing that *p* only if this justification holds independent of inferential connections to other beliefs. At a minimum, however, my belief that *p* is inferred from my belief that I can solve the Rubik's cube and my unjustified belief that if I can solve the Rubik's cube, then I am more intelligent than most of my colleagues. Of course, it *could* be that my immediate justification for *p* derives exclusively from assertive force, in spite of the other beliefs I have. But on the other hand, it is unclear why we should think that my other beliefs (especially if unjustified) are irrelevant to *p*'s epistemic standing. At best, then, ADT' supports the possibility of *p*'s being immediately justified, not its actuality.

Second, it is difficult to see how I could even have prima facie justification for believing that *p*. Suppose in fact I have no defeater for my belief: I lack counterevidence to *p*, and I am not aware of any epistemic fault of mine in forming the belief. ADT' implies the (exceedingly implausible) conclusion that my belief is ultima facie justified.

If the above considerations are sound, we have a counterexample to ADT' and hence to ADT. I believe that *p* and feel assured that *p* in virtue of so believing, yet I am arguably neither immediately nor prima facie justified in believing that *p*. The argument I have offered could be cast as a modus tollens argument: ADT carries a commitment to ADT'; but ADT' is false, so ADT is false. Let me quickly recap the main points. Assertive dogmatists argue that, for any experience, if it has assertive force, it provides immediate prima facie justification for believing its content (ADT). The view that perceptual experiences have assertive force is no more (or less) plausible that the view that some beliefs do. So, the

question arises, can beliefs confer immediate prima facie justification in virtue of their assertive force alone? Admittedly, there could be something about beliefs such that, despite having assertive force, they cannot serve as immediate prima facie justifiers. But for that assertive dogmatists tell us, it is not at all clear what that could be. As long as they do not give up on ADT, then, they are forced to accept a stronger claim, namely, that for any mental state, if it has assertive force, it provides immediate prima facie justification for believing its content (ADT'). I have argued, however, that ADT' is false, for there are, plausibly, at least some beliefs with assertive force which provide neither immediate nor prima facie justification. Therefore, assertive force alone cannot be what gives a mental state prima facie justificatory immediacy.

6.5. Putting It All Together

I started our discussion in this paper by asking whether MDs receive any degree of perceptual justification. Most non-sceptical epistemologists recognise that we have immediate prima facie justification for believing that things are as they appear in perception. However, they differ over which features of perception would enable it to provide such justification. According to PD, a person having a perceptual experience with the content that *p* is thereby immediately prima facie justified in believing that *p*; the only requirement is that the perceptual experience has the right kind phenomenal force with respect to its content *p*.

Returning to our running example of MD, let *p* be the proposition that the person Mia sees in front of her is not her mum (assuming for the sake of argument that *p* is a proposition that Mia's perceptual experience *E* represents). I argued that dogmatists' accounts of phenomenal force (the assertiveness account and Chudnoff's account) provide no good reason to think that *E* would give Mia immediate prima facie justification for *p*. I defended two specific claims. First, even if it is true that experiences can serve as immediate prima facie justifiers in virtue of their presentational force, that is not true of *E*, for *E* lacks presentational force with respect to *p*. Second, even if *E* has assertive force with respect to *p*, that does not make *p* immediately (not even defeasibly) justified, for assertive force is not sufficient for the justificatory force of perceptual experience.

As we shall now go on to see, my argument extends beyond the Capgras case to virtually every instance of MD. Consider the following well known cases:

1. Fregoli delusion (FD): the subject believes that a nearby stranger is one and the same as some known individual despite not looking like her at all (Ellis and Szulecka, 1996).
2. Mirrored-self misidentification (MSM): the subject believes that the person she sees in the mirror is a stranger and not numerically the same as herself (Breen et al., 2001).
3. Somatoparaphrenia (SP): the subject believes that body parts contralateral to the side of a brain pathology, typically upper limbs, are not their own, but someone else's (Bottini et al., 2002).[74]

Suppose (again, for the sake of argument) that when subjects form their respective beliefs, they do so by taking perceptual experience at face value. In either case, erroneous identity propositions ('a=b' and 'a≠b') are what are believed. This means (as per PD) that in order for FD, MSM, and SP to be immediately prima facie justified, perceptual experience must meet at least one of two conditions: have (i) assertive force or (ii) presentational force with respect to the misidentification contents 'a=b' and 'a≠b'. We now can safely ignore (i), since we have seen that assertive force is insufficient for any experience to serve as an immediate prima facie justifier, regardless of its content. What needs to be determined is whether (ii) holds. I do not think it does. To see this, one has to appreciate that in FD, MSM, and SP, just as in Capgras, misidentified people (or objects) do not visually appear different from who (or what) they really are.

When the subject with FD takes some stranger to be a familiar individual, call him Sam, she does not seem to see Sam. On the contrary, she is well aware that the person seen looks nothing like Sam. This means her experience does not reach the requisite threshold of presentational force, because it does not make it seem as if she is visually aware of a truthmaker for 'the person before me = Sam'. In MMS, subjects misidentify their own reflection in the mirror. All the same, what visually appears to them when they gaze into the mirror is their own reflected image. This is confirmed by the finding that subjects usually acknowledge that the person in the mirror looks like them, although they might confabulate about minor physical differences to post hoc rationalise their beliefs (Barnier et al., 2008; Breen et al., 2000). Here too, the requisite conditions for presentational force are not met. Because what the subject seems to see in the mirror is

[74] Another instance of delusional misidentification is reduplicative paramnesia, where the subject believes that one and the same place exists in two or more physical locations simultaneously (Pick, 1903). I leave this aside here, as it is unclear to me how, if at all, such belief could arise via endorsement processes.

her own self-image, it does not seem to her as if she is visually aware of a truthmaker for the proposition 'the person in the mirror ≠ myself'. Finally, something similar can be said for SP. In SP, subjects can no longer identify their limbs as their own, but they nonetheless do seem to be visually aware of their own limbs, as confirmed by anecdotical observations. For example, Olsen (1937, cited in Vallar and Ronchi, 2008, p. 545) reports the case of a patient who, when he reassured her that her left limbs were her own, replied: 'I know they look like mine, but I can't feel that they are not, and I can't believe my eyes'. Once again, the relevant experience lacks presentational force, because the relationship between the objects of which the subjects seem to be visually aware (their own limbs) and the proposition 'the limbs I see ≠ my own limbs' is not truthmaking'.

The bottom line is that, even if we assume MDs are formed via endorsement processes, they cannot be made immediately prima facie justified in virtue of the experience's phenomenology alone. In a word: MDs are unjustified perceptual beliefs, even by the relatively undemanding standards of PD.

What are we to make of this? It is sometimes said that the endorsement account adopts the epistemological stance of the dogmatist. For instance, Quinn Hiroshi Gibson and Adam Bradley (forthcoming) write that 'underlying the endorsement theory is a certain epistemology of perception (…), namely one on which we are *prima facie* entitled to believe the immediate deliverances of perceptual experience'. The idea is this: the endorsement account works under the assumption that subjects are prima facie entitled to trust their aberrant experience and therefore to take them at face value. If they were not, they would not go on to believe based on those experiences, in which case it would be odd to describe delusions as endorsements of experience.

I have argued that having an experience with the same content as a resulting MDs does not alone give one any justification at all to believe it. So, the question arises, should we take this to weigh against the endorsement theory? I think not, and the temptation to think otherwise is due to a conflation between ordinary dispositions to believe what we perceive and being prima facie justified in so believing. It is one thing to ask whether by having an experience with the content *p* you are disposed to believe *p*; it is quite another thing to ask whether by having such an experience you are prima facie justified in believing that *p*. To see this, suppose you are trying to make yourself hallucinate rain. You look outside your window and deliberately undertake to imagine seeing raindrops falling from the sky. When you do in fact hallucinate rain, you form the belief that it is raining on the basis of that experience.

Here it seems safe to say that you would lack even prima facie justification for your perceptual belief, since you fabricated the experience for yourself (Teng, 2018). Yet nonetheless, you trust your experience and believe (automatically and non-inferentially) that it is raining. This scenario seems possible, and if it is, then we have a case in which we operate in the mode of taking experience at face value, without being even prima facie justified in doing so.

I see no reason why this should be different in the case of MDs. Even if MDs are completely unjustified that does not rule out that they are not formed in the manner the endorsement theory suggests, that is, by taking aberrant experience at its face value. This being so, it makes little sense to say that underlying the endorsement account is a dogmatist epistemology. The fact that MDs are not even prima facie justified does not count against the endorsement theory. Moreover, the considerations which motivate the endorsement account are independent of considerations of what is necessary, and what suffices, for perceptual justification.

6.6. Conclusion

I have granted that there are delusions which are formed by taking an aberrant experience as veridical. My guiding question in this paper has been whether some such delusions, MDs, enjoy any degree of positive epistemic status. To give a 'Yes' answer the best chance, I approached the question from the perspective of PD, since this seems on its face as the best fit with at least the defeasible justifiedness of MDs. Indeed, according to PD, having an experience of the right phenomenology, i.e., presentational or assertive force, is a sufficient condition for immediate prima facie justification, regardless of the experience's aetiology. I have argued that even if we adopt the epistemological stance of PD, there are reasons to think MDs are completely unjustified perceptual beliefs. First, MDs' contents are not propositions with respect to which the relevant experiences have presentational force. Second, assertive force itself is irrelevant to whether experiences justify us in believing their contents, so it does not matter that the relevant experience may have it. In sum, then, the prospects are dim for moving from the claim that MDs are formed by endorsing contents of experience to the claim that one thereby has prima facie justification for holding them. At the same time, one need not be prima facie entitled to endorse the content of one's experience in order to endorse it. This of course does not exclude that there might be delusions formed via endorsement processes that enjoy positive epistemic status, and anyone

interested in delusions would do well to consider whether there are. Still, we should not draw epistemological conclusions too quickly from the consideration that some delusions, if any, are formed in the manner typical of perceptual beliefs.

6.7. References

Aimola-Davies, A. and Davies, M. (2009). 'Explaining Pathologies of Belief'. In M. Broome and L. Bortolotti (eds.) Psychiatry as Cognitive Neuroscience: Philosophical Perspectives (pp. 285–323). Oxford: Oxford University Press.

Alicke, M. D. (1985). 'Global Self-Evaluation as Determined by the Desirability and Controllability of Trait adjectives'. *Journal of Personality and Social Psychology*, 49, pp. 1621–1630.

Alicke, M. D. and Govorun, O. (2005). '*The Better-Than-Average Effect*'. In M. D. Alicke, D. A. Dunning and J. I. Krueger (eds.) *Studies in self and identity. The Self in Social Judgment* (pp. 85–106). New York: Psychology Press.

Alston, W. P. (1996). 'Belief, Acceptance, and Religious Faith'. In J. J. and D. Howard-Snyder (eds.) *Faith, Freedom, and Rationality* (pp. 3–27). Lanham, MD: Rowman and Littlefield.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of mental Disorders: Fifth Edition*. Washington: American Psychiatric Press.

Barnier, A. J., Cox, R. E., O'Connor, A., Coltheart, M., Langdon, R., Breen, N. and Turner, M. (2008). 'Developing Hypnotic Analogues of Clinical Delusions: Mirrored-Self Misidentification'. *Cognitive Neuropsychiatry*, 13(5), pp. 406–430.

Bayne, T. and Pacherie, E. (2004). 'Bottom-up or Top-down? Campbell's Rationalist Account of Monothematic Delusions'. *Philosophy, Psychiatry, and Psychology*, 11, pp. 1–11.

Bayne, T. and Pacherie, E. (2005). 'In Defence of the Doxastic Conception of Delusions'. *Mind and Language*, 20(2), pp. 163–188.

Bendana, J. and Mandelbaum, E. (forthcoming). 'The Fragmentation of Belief'. In D. Kinderman, A. Onofri and C. Borgoni (eds.) *The Fragmented Mind*. Oxford: Oxford University Press.

Bengson, J. (2015). 'The Intellectual Given'. *Mind*, 124(495), pp. 707–760.

Bortolotti L. (2018). 'Delusions and Three Myths of Irrational Belief'. In L. Bortolotti (ed.) *Delusions in Context*. Cham: Palgrave Macmillan.

Bortolotti, L. (2012). 'In Defence of Modest Doxasticism About Delusions'. *Neuroethics*, 5(1), pp. 39–53.

Bottini, G., Bisiach, E., Sterzi, A. and Vallar, G. (2002). 'Feeling Touches in Someone Else's Hand'. *Neuroreport*, 13(2), pp. 249–52.

Breen, N., Caine, D., Coltheart, M., Hendy, J. and Roberts, C. (2000). 'Towards an Understanding of Delusions of Misidentification: Four Case Studies'. *Mind and Language*, 15(1), pp. 74–110.

Breen, N., Caine, D. and Coltheart, M. (2001). 'Mirrored-Self Misidentification: Two Cases of Focal Onset Dementia'. *Neurocase*, 7, pp. 239–254.

Brighetti, G., Bonifacci, P., Borlimi, R. and Ottaviani, C. (2007). "Far from the Heart far from the Eye': Evidence from the Capgras delusion'. *Cognitive Neuropsychiatry*, 12, pp. 189–197.

Brogaard, B. (2013). 'Phenomenal Seemings and Sensible Dogmatism'. In C. Tucker (ed.) *Seemings and Justification* (pp. 270–291). New York: Oxford University Press.

Brogaard, Berit. (2016). 'Staying Indoors: How Phenomenal Dogmatism Solves the Skeptical Problem Without Going Externalist.' In B. Coppenberg and M. Bergmann (eds.) *Intellectual Assurance: Essays on Traditional Epistemic Internalism* (pp. 85–104). Oxford: Oxford University Press.

Brown, J. D. (1986). 'Evaluations of Self and Others: Self-Enhancement Biases in Social Judgments'. *Social Cognition*, 4, pp. 353–376.

Burge, T. (2003). 'Perceptual Entitlement'. *Philosophy and Phenomenological Research*, 67, pp. 503-–548.

Campbell, J. (2001). 'Rationality, Meaning, and the Analysis of Delusion'. *Philosophy, Psychiatry, and Psychology*, 8(2–3), pp. 89–100.

Carruthers, P. (2017). 'The Illusion of Conscious Thought'. *Journal of Consciousness Studies*, 24(9/10), pp. 228–52.

Chisholm, R, (1980), 'A Version of Foundationalism'. *Midwest Studies in Philosophy*, 5(1), pp. 543–564.

Chudnoff, E. (2011). 'What Intuitions are Like'. *Philosophy and Phenomenological Research*, 82(3), pp. 625–654.

Chudnoff, E. (2012). 'Presentational Phenomenology'. In S. Miguens and G. Preyer (eds.) *Consciousness and subjectivity* (pp. 51–72). Berlin: Ontos Verlag.

Chudnoff, E. (2013). *Intuition*. New York: Oxford University Press.

Chudnoff, E. (2016). 'Epistemic Elitism and Other Minds'. *Philosophy and Phenomenological Research*, 96 (2), pp. 276–298

Chudnoff, E. and DiDomenico, D. (2015). 'The Epistemic Unity of Perception'. *Pacific Philosophical Quarterly*, 96, pp. 535–549.

Clore, G. L. and Parrott, W. G. (1994). 'Cognitive Feelings and Metacognitive Judgments'. *European Journal of Social Psychology, 24*(1), pp. 101–115.

Cohen, J. (1992). *An Essay on Belief and Acceptance*. Oxford: Oxford University Press.

Coltheart, M. (2013). 'On the Distinction Between Monothematic and Polythematic Delusions'. *Mind and Language*, 28 (1), pp. 2013–2112.

Conee, E. and Feldman, R. (1998). 'The Generality Problem for Reliabilism'. *Philosophical Studies*, 89, pp. 1–29.

Conee, E. and Feldman, R. (2004). *Evidentialism*. Oxford: Oxford University Press.

Cowan, R. (2018). 'Epistemic Sentimentalism and Epistemic Reason-Responsiveness'. In A. Bergqvist and R. Cowan (eds.) *Evaluative Perception* (pp. 219–236). Oxford: Oxford University Press.

Crane, T. (2013). 'Unconscious Belief and Conscious Thought'. In U. Kriegel (ed.) *Phenomenal Intentionality* (pp. 156–172). New York: Oxford University Press.

Davies, M. and Egan, A. (2013). 'Delusion: Cognitive Approaches—Bayesian Inference and Compartmentalization'. In K. W. M. Fulford, M. Davies, R. G. T. Gipps, G. Graham, J. Z. Sadler, G. Stanghellini and T. Thornton (eds.) *The Oxford handbook of philosophy and psychiatry* (pp. 688–727). Oxford: Oxford University Press.

Ellis, H. D. and Szulecka, T. K. (1996). 'The Disguised Lover: A case of Fregoli Delusion'. In P. W. Halligan and J. C. Marshall (eds.) *Method in madness: Case studies in Cognitive Neuropsychiatry* (pp. 39–50). East Sussex: Psychology Press.

Ellis, H. D. and Young, A. W. (1990). 'Accounting for Delusional Misidentifications'. *British Journal of Psychiatry*, 157, 239–248.

Ellis, H. D., Lewis, M. B., Moselhy, H. F. and Young, A. W. (2000). 'Automatic without Autonomic Responses to Familiar Faces: Differential Components of Covert Face Recognition in a Case of Capgras Delusion'. *Cognitive Neuropsychiatry*, 5(4), pp. 255–269.

Ellis, H.D., Young, A.W., Quayle, A. and de Pauw, K. W. (1997). 'Reduced Autonomic Responses to Faces in Capgras Delusion'. *Proceedings of the Royal Society: Biological Sciences*, B264, pp. 1085–1092.

Ghijsen, H. (2014). 'Phenomenalist Dogmatist Experientialism and the Distinctiveness Problem'. *Synthese*, 191(7), pp. 1549–1566.

Gibson, Q. H. and Bradley, A. (forthcoming). 'Monothematic Delusions and the Limits of Rationality'. *British Journal for the Philosophy of Science*.

Goldman, A (1999). 'Internalism Exposed'. *Journal of Philosophy*, 96(6), pp. 271–293.

Goldman, A. (1979). 'What Is Justified Belief?' In G. Pappas (ed.) *Justification and Knowledge* (pp. 1–25). Boston: Reidel.

Goldman, A. (2008). 'Immediate Justification and Process Reliabilism'. In Q. Smith (ed.) *Epistemology: New Essays* (pp. 63–82). Oxford: Oxford University Press.

Goldman, A. (2011). 'Toward a Synthesis of Reliabilism and Evidentialism? Or: Evidentialism's Troubles, Reliabilism's Rescue Package'. In T. Dougherty (ed.) *Evidentialism and Its Discontents* (pp. 254–280). Oxford: Oxford University Press.

Hawley, K. and MacPherson, F. (eds.) (2011). *The Admissible Contents of Experience*. Malden: Wiley-Blackwell.

Hinchman, T. (2005). 'Telling as Inviting to Trust'. *Philosophy and Phenomenological Research*, 70(3), pp. 562–87.

Huemer, M. (2001). *Skepticism and the Veil of Perception*. Lanham: Rowman and Littlefield.

Huemer, M. (2006). 'Phenomenal Conservatism and the Internalist Intuition'. *American Philosophical Quarterly*, 43, pp. 147–58.

Huemer, M. (2007). 'Moore's Paradox and the Norm of Belief'. In S. Nuccetelli and G. Seay (eds.) *Themes from G. E. Moore: New Essays in Epistemology and Ethics* (pp. 142–157). Oxford: Oxford University Press.

Kind, A. (2018). 'How Imagination Give Rise to Knowledge'. In F. Dorsch and F. Macpherson (Eds.) *Perceptual Memory and Perceptual Imagination* (pp. 227–246). New York: Oxford University Press.

Kumkale, G. T. and Albarracin, D. (2004). 'The Sleeper Effect in Persuasion: A Meta-Analytic Review'. *Psychological Bulletin*, 130(1), pp. 143–172.

Langdon, R. and Bayne, T. (2010). 'Delusion and Confabulation: Mistakes of Perceiving, Remembering and Believing'. *Cognitive Neuropsychiatry*, 15, pp. 319–345.

Littlejohn, C. (2009). 'The Externalist's Demon'. *Canadian Journal of Philosophy*, 39 (3), pp. 399–434.

Lyons, J. (2005). 'Perceptual Belief and Nonexperiential Looks'. *Philosophical Perspectives*, 19, pp. 237–56.

Lyons, J. (2015). 'Seemings and Justification'. *Analysis Reviews,* 75(1), pp. 153–64.

Lyons, J. C. (2009). *Perception and Basic Beliefs: Zombies, Modules, and the Problem of the External World*. New York: Oxford University Press.

Maher, B. A. (1974). 'Delusional Thinking and Perceptual Disorder'. *Journal of Individual Psychology*, 30, pp. 98–113.

Mandelbaum, E. (2014). 'Thinking is Believing', *Inquiry*, 57, pp. 55–96.

Mandelbaum, E. (2019). 'Troubles with Bayesianism: An Introduction to the Psychological Immune System'. *Mind and Language*, 34(2), pp. 141– 157.

Markie, P. (2005). 'The Mystery of Direct Perceptual Justification'. *Philosophical Studies*, 126, pp. 347–373.

McDowell, J. (1995). 'Knowledge and the Internal'. *Philosophy and Phenomenological Research,* 55, pp. 877–893.

McGrath, M. (2013). 'Phenomenal Conservatism and Cognitive Penetration: The 'Bad Basis' Counterexamples'. In C. Tucker (ed.) *Seemings and Justification: New Essays on Dogmatism and Phenomenal Conservatism* (pp. 225– 247). New York: Oxford University Press.

Moran, R. (2018). *The Exchange of Words: Speech, Testimony, and Intersubjectivity*. New York: Oxford University Press.

Pick, A. (1903). 'On Dreamy Mental States as a Permanent Condition in Epileptics'. *Brain*, pp. 260–67.

Pollock, J. (1974). *Experience and Justification*. Princeton: Princeton University Press.

Pritchard, D. (2012). *Epistemological Disjunctivism*. Oxford: Oxford University Press.

Pryor J. (2001) 'Highlights of Recent Epistemology'. *The British Journal for the Philosophy of Science*, 52, pp. 95–124.

Pryor, J. (2000). 'The Skeptic and the Dogmatist'. *Noûs*, 34, pp. 517–549

Pryor, J. (2004). 'What's Wrong with Moore's Argument?'. *Philosophical Issues*, 14, pp. 349–378.

Ramachandran, V. S. and Blakeslee, S. (1998). *Phantoms in the Brain: Human Nature and the Architecture of the Mind*. London: Fourth Estate.

Reid, T. (1788). *Essays on the Active Powers of the Human Mind*. Cambridge, MA: MIT Press, 1969.

Reiland, I. (2014). 'On Experiencing High-Level Properties'. *American Philosophical Quarterly*, 51, pp. 177–87.

Reiland, I. (2015). 'Experience, Seemings, and Evidence'. *Pacific Philosophical Quarterly*, 96, pp. 510-534.

Ross, A. (1986). 'Why Do We Believe What We Are Told?'. *Ratio*, 28, pp. 69–88.

Schellenberg, S. (2013). 'Experience and Evidence'. *Mind*, 122, pp. 699–747.

Schmitt, F. (2010). 'The Assurance View of Testimony'. In A. Haddock, A. Millar and D. Pritchard (Eds.) *Social Epistemology* (pp. 216–41). Oxford: Oxford University Press.

Schwitzgebel, E. (2002). 'A Phenomenal, Dispositional Account of Belief'. *Noûs*, 36, pp. 249–275.

Siegel, S. (2010). *The Contents of Visual Experience. New York: Oxford University Press.*

Siegel, S. (2012). 'Cognitive Penetrability and Perceptual Justification'. *Noûs*, 46 (2), pp. 201–222.

Siegel, S. (2017). *The Rationality of Perception*. New York: Oxford University Press.

Silins, N. (2008). 'Basic Justification and the Moorean Response to the Skeptic'. In T. Gendler and J. Hawthorne (eds.) *Oxford studies in epistemology* (Vol. 2, pp. 108–142). Oxford: Oxford University Press.

Smithies, D. (2019). *The Epistemic Role of Consciousness*. New York: Oxford University Press.

Sosa, E. (1991). *Knowledge in Perspective*. Cambridge: Cambridge University Press.

Strick, M., Dijksterhuis, A., Bos, M. W., Sjoerdsma, A., van Baaren, R. B. and Nordgren, L. F. (2011). 'A Meta-analysis on Unconscious Thought Effects'. *Social Cognition*, 29(6), pp. 738–762.

Teng, L. (2018) 'Is Phenomenal Force Sufficient for Immediate Perceptual Justification?' *Synthese*, 195 (2), pp. 637–656.

Tolhurst, W. (1998). 'Seemings'. *American Philosophical Quarterly*, 35(3), pp. 293–302.

Tooley, M. (2013). 'Michael Huemer and the Principle of Phenomenal Conservatism'. In C. Tucker (ed.) *Seemings and Justification: New Essays on Dogmatism and Phenomenal Conservatism* (pp. 306–27). New York: Oxford University Press.

Tucker, C. (2010). 'Why Open-Minded People Should Endorse Dogmatism'. *Philosophical Perspectives,* 24, pp. 529–45.

Tucker, C. (2013). 'Seemings and Justification: An Introduction'. In C. Tucker (ed.) *Seemings and justification* (pp. 1–27). New York: Oxford University Press.

Tuomela, R. (2000). 'Belief versus Acceptance'. *Philosophical Explorations*, 3(2), pp. 122–137.

Vallar, G. and Ronchi R. (2009). 'Somatoparaphrenia: a Body Delusion. A Review of the Neuropsychological Literature'. *Experimental Brain Research*, 192 (3), pp. 533–551.

Weiner, M. (2003). 'Accepting Testimony'. *The Philosophical Quarterly*, 53, pp. 256–64.

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford: Oxford University Press.

Wittgenstein, L. (1969). *On certainty* (ed.) G. E. M. Anscombe and G. H. von Wright, trans. D. Paul and G. E. M. Anscombe. Oxford: Blackwell