# Is Feeling Believing?
# Insights on Introspection, Illusions, Delusions, and the Relationship Between Experience and Belief

Dissertation
zur Erlangung des akademischen Grades Dr. phil.,
genehmigt durch die
Fakultät für Humanwissenschaften
der Otto-von-Guericke-Universität Magdeburg

von Chiara Caporuscio
geb. am 18.10.1996 in Udine

Gutachter: Jun-Prof. Dr. Sascha Benjamin Fink
Gutachter: Prof. Dr. med. Dr. phil. Henrik Walter
Gutachterin: Prof. Lisa Bortolotti
Eingereicht am: 28.03.23
Verteidigung der Dissertation am: 11.07.23

# Table of Contents

## Abstract (Deutsch)

**Kurzfassung zur Dissertation mit dem Thema "Is feeling Believing? Insights on Introspection, Illusions, Delusions, and the Relationship Between Experience and Belief" vorgelegt von Chiara Caporuscio**

Diese Dissertation ist als eine Sammlung von eigenständigen Aufsätzen angelegt, die das Verhältnis zwischen Erfahrung und Glauben untersuchen. Erfahrung ist definiert als die subjektive Qualität des Mentalen, während Glaube eine propositionale Einstellung ist, die sich auf einen Zustand oder eine Art und Weise bezieht, wie die Welt sein könnte. Erfahrung und Glaube sind eng miteinander verwoben, aber manchmal weicht die Erfahrung von der Realität oder der Glaube von der Erfahrung ab. In dieser Arbeit gehe ich auf verschiedene Fälle ein, in denen Erfahrung und Glaube auseinanderklaffen, und untersuche die Rolle des Handelns bei der Überbrückung der Kluft zwischen Erfahrung und Glaube.

In Kapitel 1 führe ich in das Thema der Dissertation ein und gehe kurz auf die Rolle der Philosophie in den Kognitionswissenschaften ein. In den beiden folgenden Kapiteln geht es um introspektive Überzeugungen, die manchmal als direkter Zugang zur Erfahrung angesehen werden und daher als Goldstandard in der Bewusstseinswissenschaft gelten. In Kapitel 2 wird argumentiert, dass dieses "introspektive Privileg" auf einen Mangel an Klarheit in Bezug auf das Konzept der Introspektion zurückzuführen ist und dass introspektive Überzeugungen die Erfahrung falsch erfassen können. Kapitel 3 untersucht die Möglichkeit introspektiver Wahnvorstellungen, d. h. falscher pathologischer Überzeugungen, die introspektive Fehler beinhalten. Kapitel 4 befasst sich mit der Cornsweet-Illusion als Beispiel für eine visuelle Täuschung, bei der Erfahrung und Überzeugung in Bezug auf dasselbe Objekt entkoppelt sind, und argumentiert, dass die Beibehaltung einer hierarchischen und einheitlichen Sichtweise von Wahrnehmung und Kognition immer noch die Erklärung widersprüchlicher Wahrnehmungen und Überzeugungen ermöglicht. Kapitel 5 untersucht die Beziehung zwischen Glaubensänderungen und psychedelischen Erfahrungen und argumentiert, dass Handlungen eine wichtige Rolle dabei spielen, unsere Überzeugungen über uns selbst in die Realität umzusetzen. In Kapitel 6 integriere ich die vorangegangenen Kapitel, indem ich Implikationen,

allgemeine Schlussfolgerungen und zukünftige Richtungen aufzeige. Der Dissertation folgt ein Anhang mit zwei Buchrezension über Überzeugungen und Rationalität.

Ich bin der Meinung, dass die Untersuchung der Beziehung zwischen Erfahrung und Glauben von großer Bedeutung für die Bewusstseinsforschung und die Psychiatrie ist. Introspektive Überzeugungen und Berichte sind eine wichtige Quelle für Informationen über die mentalen Zustände und Erfahrungen einer Person; Allerdings sind introspektive Berichte als Maßstab nicht vollständig zuverlässig. Die Beziehung zwischen Erfahrungen und Überzeugungen ist auch deshalb so wichtig, weil sie sich gegenseitig stark beeinflussen, wobei die Erfahrungen die Überzeugungen nähren und die Überzeugungen die Erfahrungen formen. Dies ist besonders wichtig bei der Untersuchung der Ursachen von Wahnvorstellungen, da die Beziehung zwischen bizarren Erfahrungen und wahnhaften Überzeugungen immer noch umstritten ist. Die Auswirkungen auf die psychische Gesundheit und die Therapie gehen jedoch weit über wahnhafte Störungen hinaus: Positive oder traumatische Erfahrungen können gesunde oder schädliche Veränderungen in den Überzeugungen auslösen, die zu sich selbst erfüllenden Prophezeiungen werden können, indem sie neue Erfahrungen prägen. Wenn man versteht, wie Erfahrungen und Überzeugungen zusammenwirken, kann man die psychiatrischen Interventionen unterstützen.

## Abstract (English)

### Abstract for the dissertation titled "Is Feeling Believing? Insights on Introspection, Illusions, Delusions, and the Relationship Between Experience and Belief", written by Chiara Caporuscio

This dissertation is set out as a collection of self-standing essays exploring the relationship between experience and belief. Experience is defined as the subjective quality of the mental, while belief is a propositional attitude referring to a state of affairs or a way the world could be. Experience and belief are deeply intertwined, but sometimes experience deviates from reality or belief deviates from experience. Throughout this thesis, I explore different cases where experience and belief come apart, and examine the role of action in bridging the gap between experience and belief.

In Chapter 1, I introduce the topic of the dissertation and briefly discuss the role of philosophy in the cognitive sciences. The following two chapters focus on introspective beliefs, that are sometimes seen as a direct way to access experience and therefore are considered a gold standard in consciousness science. Chapter 2 argues that this "introspective privilege" is due to a lack of clarity around the concept of introspection, and that introspective beliefs can incorrectly capture experience. Chapter 3 explores the possibility of introspective delusions, namely false pathological beliefs that involve introspective errors. Chapter 4 looks at the Cornsweet Illusion as an example of a visual illusion where experience and belief about the same object are decoupled, and argues that maintaining a hierarchical and unified view of perception and cognition still allows to explain conflicting percepts and beliefs. Chapter 5 explores the relationship between belief change and psychedelic experiences, and argues that action plays an important role in turning our beliefs about ourselves into reality. In Chapter 6, I integrate the preceding chapters by identifying implications, general conclusions, and future directions. The dissertation is followed by an Appendix containing two book reviews about beliefs and rationality.

I believe that studying the relationship between experience and belief is of great importance for the study of consciousness and psychiatry. Introspective beliefs and reports are a crucial source of information about someone's mental states and

experiences; however, relying on introspective reports as a measure of experience may not be entirely reliable. The relationship between experience and belief is also important because of the strong influence they have on each other, with experiences feeding beliefs and beliefs shaping experiences. This is particularly relevant in studying the causes of delusions, where the relationship between bizarre experiences and delusional beliefs is still controversial. However, the implications for mental health and therapy extend far beyond delusional disorders: positive or traumatic experiences can trigger healthy or damaging changes in beliefs, which can turn into self-fulfilling prophecies by shaping new experiences. Understanding how experience and belief interact can therefore inform and aid mental health interventions and therapy.

# Chapter 1.
# Introduction

## 1.1. Background

This dissertation is a collection of self-standing essays exploring the multifaceted relation between experience and belief. When I say experience, I mean phenomenal experience: the subjective quality of the mental, the what-it-is-like to be someone at a particular moment in time (Nagel 1974). My current experience includes the what-it-is-like to see a cat, to hear it purring, to feel the light pressure on my lap, and to feel comfort for its presence. Beliefs, instead, are propositional attitudes that refer to a state of affairs or the way the world could be, and they involve a certain degree of confidence (Schwitzgebel 2021). I believe that there is a cat sleeping on my lap if I would confidently say so when asked, and if that belief is integrated with my other beliefs and desires in order to guide action: if I believe that it will wake up if I move, and I do not want it to wake up, I will try to stay as still as possible.

Experience and belief are deeply intertwined: my belief that a cat is sitting on my lap is largely motivated by my experience of it. Even in situations where the belief is more removed from experience, there is still often a link between the two: I might not have directly seen Biden becoming the president of the US, but I have heard people I trust talking about it, or read newspaper articles. In an ideal scenario, our experience would correctly represent reality, and our beliefs would be accurate inferences from our experiences. However, sometimes experience deviates from reality, or belief deviates from experience.

Experiences that are decoupled from reality are known as hallucinations and illusions. Illusions are non-pathological perceptual misinterpretations of a sensory stimulus: for example, a stick half-submerged in water will appear as bent. Hallucinations are errors in perception that cause a perception-like experience in the absence of sensory stimulus (MacPherson 2013). While this term is mostly used in reference to visual hallucinations, it includes experiences from other sensory modalities, like auditory and olfactory hallucinations, and experiences that are not easily reducible to one sensory modality, like experiencing one's normal surroundings as confusing and alien during a panic attack, or feeling phantom pain in a limb that does not exist anymore (Sacks 2012). In all of these cases, there is a mismatch between reality and experience.

Subjects who are hallucinating or are victims of an illusion are sometimes unaware of it. If someone spiked their drink with a powerful hallucinogen, they might believe that a queen on a horse has entered the room (at least, until the horse starts breathing fire). In this case, the experience is dissociated from reality and drives astray the belief.

However, beliefs are not always fully determined by an experience taken at face value. Instead, they are in part formed by taking into account different sources of evidence, generating different hypotheses and weighting their consistency to the rest of our beliefs and the particular experience that needs to be explained (Connors and Halligan 2015; 2020) The suspicion that they have ingested a hallucinogen or the belief that fire-breathing horses are unlikely to find at their local pub can help the hallucinating subject doubt their experience and block this experience from penetrating their beliefs.

This also works the other way around. Someone might have an experience that correctly represents reality, but doubt it because of how unlikely it seems - maybe a horse-riding fire illusionist really has entered their local pub. Alternatively experience and belief can both misrepresent reality, but in different ways: they hallucinate a horse-riding queen and come to believe that it's a manifestation of death and it's out to get them.

In this dissertation, I will look at several cases where experience and belief come apart. I will start from introspective beliefs, which according to their reputation are direct and infallible depictions of experience. I will argue that this is not the case: even when introspecting, it is possible to form beliefs that are removed from the experiences they are about. This is true in regular, daily cases of belief formation (Chapter 2), but it might also be the case for some clinical delusions (Chapter 3). Introspective beliefs that do not correctly represent experience are false; when forming beliefs about the external world, however, it might be epistemically good to distrust experience. In Chapter 4, I will look at some visual illusions in which we are aware that our experience is tricking us, resulting in true beliefs that diverge from our experience of the same object. Finally, I will look at the role of action in bridging the gap between experience and belief. For this, I will refer to a specific case study: psychedelic-assisted therapy (Chapter 5). In Chapter 6, I integrate the preceding chapters by identifying implications, general conclusions, and future directions

### 1.2. Motivation

I address here three questions regarding the motivation and relevance of this work. Why should we care about experience? Why should we care about belief? And most importantly, why should we care about the relationship between the two?

The first question is the one with the most thoroughly defended answer in the recent history of philosophy of mind; for this reason, I will keep my answer here short. The subjective nature of mental phenomena is one of the most often discussed unresolved mysteries of modern neuroscience. According to some, whether a subject is conscious or capable of consciousness has implications of whether they should be considered an ethical subject with rights and responsibilities (Madva 2021; Bernàth 2021). Furthermore, our first-person experience *feels* important. As Anil Seth (2021) puts it:

“Imagine a future version of me [...] offers you the deal of a lifetime. I can replace your brain with a machine that is equal in every way, so that from the outside, nobody could tell the difference. This new machine has many advantages - it is immune to decay, and perhaps it will allow you to live forever. But there’s a catch. Since future-me isn’t sure how real brains give rise to consciousness, I can’t guarantee that you will have any conscious experience at all, should you take up this offer [...]. I suspect you wouldn’t take the deal.”
(Seth 2021, p. 3).

Not only consciousness in general is important – we also care about the specific properties of some specific experiences. The phenomenology of mental illnesses can improve our understanding of patients’ lived experiences, and thus facilitate diagnosis and treatment.  Phenomenology can sometime tell us more about the brain: classifying the geometric patterns that commonly recur in hallucinogenic visuals has improved our understanding of the visual cortex (Bresslof et al. 2002). Investigating variability in different people's experience can improve our understanding of neurodiversity.  Not all differences in experience come with obvious differences in behaviour, and taking interest in people's first-person perspective can lead to the discovery of anomalies that would otherwise go unnoticed: some examples of differences in experience that were only properly understood in the past couple of decades are aphantasia, namely the lack of visual imagery (Zeman et al. 2015) and synesthesia, namely a blending of sensory modalities (Ramachandran  and Hubbard 2001; Jewanski et al. 2020).

Why care about belief? A lot of the philosophical efforts in the past fifty years have centered around defining rationality: the art of coming up with and maintaining good beliefs that serve us well and get us closer to the truth (Williams 1982; Harman 1986; Foley 1992). This type of work has mostly targeted how we should ideally form beliefs rather than how we actually form beliefs: the philosophical conception of a "good belief" is informed by the platonic idea of rationality rather than by actual belief-formation processes. For example, beliefs are procedurally rational if they are well-integrated with the rest of one's beliefs and intentional states, epistemically rational if they adequately respond to evidence, and agentially rational if they guide or inform action (Bortolotti 2009). These are prescriptive norms, not meant to capture real cases of belief formation.

This has started to change in recent decades, as philosophers have ventured outside of the ivory tower to work side-by-side with the empirical sciences - for example, by studying real-life cases of imperfect belief formation like delusions to inform our understanding of non-pathological beliefs.  Delusions are defined as "false beliefs due to incorrect inference about external reality" in DSM-III and IV, while the DSM-V describes them as"fixed beliefs that are not amenable to change in light of conflicting evidence". According to Bortolotti (2009), delusional beliefs deviate from the norms of rationality in ways that are not categorically different from non-delusional beliefs. Our understanding of delusion can inform our understanding of belief-formation, and vice versa (Connors and Halligan 2015; 2020).

The same angle has been used to look at other cases of imperfect beliefs - conspiracy theories, for example. Levy's *Bad Beliefs* (2021), and Bermudez' *Frame it Again* (2020)[1] are recent books that look at contemporary cases of imperfect beliefs, such as climate change skepticism, not from the ideal rationality angle, but with the intention of understanding some of the real, imperfect mechanisms that inform our belief formation processes. Levy's book talks about the practice of deferring beliefs to one's own social environment, a strategy that has no place in the Englightenment-based ideal of rationality, according to which humans are supposed to question each and every belief they did not form themselves. Levy instead notes how social deference is a very important strategy to form good beliefs about complex matters where the first-hand evidence is too complicated to be understood by a layman, such as climate change; however, when this strategy is used in a polluted epistemic environment such as the one we live in, where fake news and partesan information is hard to distinguish from reputable sources, it can foster the development of bad beliefs and conspiracy theories. Bermudez talks in a similar way about framing effects, namely, the tendency to value the same thing differently when it is accompanied by a different narrative. There is no space for framing effects in the bayesian ideal of rationality, that is based around internal consistency; however, Bermudez argues, frame-sensitive reasoning might lead to more rational decisions in complex, multifaceted situations like political decisions and clashes of values.

I have argued that we should care about experience and about belief. But why should we care about their overlap? The first reason is that the most important source of information about others' mental states is their beliefs about them. One of the biggest challenges of consciousness research is accessing someone else's mental experience (Nagel 1974). If we are trying to understand what another person is experiencing, we have two options: we can ask them, or we can try to figure it out ourselves, using either our mindreading abilities (Carruthers 2017) or third person methods, such as brain scans (Pauen  and Haynes 2021). If we are asking them, our understanding of their mental states will be dependent on their introspective beliefs about their own experience. Not only that, introspective reports are the gold standard that second and third-person methods are measured

[1] For more about these two books, see the Appendix of this dissertation.

against. However we try to measure the mental, we start from the assumption that our beliefs about our experience give us direct access to the experience itself.

It is controversial whether this assumption is warranted. It is questionable whether introspective reports have actually earned their reputation of infallible measures of experience, and a number of studies have highlighted the limitations of this method (Jack and Roepstorff 2003; Schwitzgebel 2002 2008; Pronin 2019). Turning one's experience into an introspective belief entails translating a non-verbal experience it into concepts, which can open space for errors and misunderstandings. If someone says they are visualizing, we assume that we know what their experience is like. But how can we be sure that "visualizing" means the same thing for an aphant and a person with vivid visual imagery?

Another reason to study belief and experience together is the strong influence they have on each other. Our experiences feed our beliefs, our beliefs in turn influence the way we experience the world. The second process is known as cognitive penetration, namely, the influence of higher-order cognitive states on perceptual experience (Macpherson 2012 2017). At the same time though, perception and belief maintain some independence, as becomes clear in cases where we don't believe what we see, or we have experiences conflicting with our worldview.

The relationship between experience and belief is of particular importance when studying the aetiology of delusions. Delusions are often accompanied by bizarre experiences; however, the relationship between the two is controversial. The most influential understanding of delusional belief formation in recent years has been empiricism: the idea that delusions are caused by powerful hallucinations, that drive belief formation astray (Bayne and Pacherie 2004a 2004b; Noordhof and Sullivan-Bissett 2021). Rationalism, on the other hand, defends that delusions involve a "top-down disturbance in some fundamental beliefs of the subject, which may consequently affect experiences and actions" (Campbell 2001, p. 89).

The implications for mental health and therapy extend far beyond delusional disorders: as traumatic experiences can trigger damaging beliefs about oneself and the world, meaningful experiences can be powerful instruments of change, and lead to the development of healthier beliefs about the self and the world. At the same time, beliefs can shape and structure one's experience, making it possible to fall into a loop of self-fulfilling prophecies: if someone believes they are incapable of leaving their bed, the way they will act in the world and sample their environemnt is going to be such that their experiences will conform to the belief and thus confirm it. Taking a closer look at how belief and experience influence each other can help us understand these dynamics and develop effective interventions.

## 1.3. Overview

In the first two essays of my dissertation, I look at introspective beliefs. The reason why I focus on introspection is that it´s a particular subset of belief formation that has been traditionally given a privileged epistemic status. The belief that it's a rainy day outside is formed by weighing different sources of evidence, such as the view from my window or the weather forecast. It might be influenced by motivational factors, such as my desire to have a picnic later in the day. It is prone to error - for example, my upstairs neighbour watering the plants on their balcony might have caused me to jump to conclusions - and it can be revised and updated when new evidence comes in. My belief that I'm feeling happy, on the other hand, has been regarded by a long philosophical tradition as being fundamentally different: direct and infallible depictions of experience.

In **Chapter 2,** I argue that this is not the case: introspective beliefs can incorrectly capture experience. The way we form most beliefs about our inner world is often not so different from the way we form beliefs about the external world - and, like the latter, it can go astray.

I argue that the so called "introspective privilege" has to do with a lack of clarity around the concept of introspection. Some introspective beliefs can be plausibly defended as being highly protected from error because they are exclusive, i.e. they are only determined by their target mental state and nothing else. Judgements of this kind are "I am feeling this" (Gertler 2012) or "This is R", where R is a phenomenal concept purely constituted by the experience (Chalmers 2003). However, this also makes them uninformative: their infallibility is not helpful for our everyday goals of communicating our mental states to others, guiding action or learning something about ourselves. The informative beliefs that better capture daily instances of introspection in practical use, like “I am feeling a throbbing pain” cannot be infallible in this way, as they require relating the phenomenal character of our mental state to other concepts and experiences.

In the rest of the chapter, I compare informative introspection with regular belief formation. I employ a 5-stage cognitive account of belief formation put forward by Connors and Halligan (2015; 2020), according to which beliefs arise in response to a distal trigger, namely, a precursor (stage 1). Then, different hypotheses to explain the precursor are formulated in a search for meaning (stage 2) and evaluated (stage 3). The hypothesis that better explains the precursor given the rest of our beliefs becomes accepted as a new belief (stage 4) and affects new beliefs and lower-level processes (stage 5). In this process, non-pathological errors and delusions can arise when something goes wrong in stages 2 and 3: for example, when we lack the background knowledge that would help us formulate the right hypothesis, when our background beliefs are false and lead us astray, or when our biases or motivational factors lead us to favour the wrong hypothesis.

I argue that the same five stages are likely to be needed for the formation of informative introspective beliefs, meaning that false or missing background beliefs, biases or motivational factors can mess with stages 2 and 3 and lead them astray. For example, a psychiatric patient lacking the notion of intrusive thoughts might be unable to formulate the right hypothesis about their mental state, and thus mistake them for desires; someone angry at a friend for petty reasons might decide to favour the hypothesis that they are perfectly calm because they don’t want to be the kind of person that holds unmotivated grudges. Informative introspective beliefs have similar failure conditions as beliefs about the external world.

In **Chapter 3**, I turn to questions about pathological errors in introspective belief formation. According to the DSM IV[2], a delusion is a pathological failure in belief formation “based on incorrect inference about external reality [...]”. A prima facie reason to maintain the external reality condition is the presumed infallibility of

[2] This definition has been changed in the DSM-V (2013) to exclude the reality condition, because of controversial *prima facie* counter-examples. However, the question of whether such counterexamples actually entail false introspective beliefs is controversial (see Chapter 3 for an extended discussion).

introspection. However, if Chapter 2 is on the right track, we could be as dramatically wrong about our internal world as we are about our external world: our beliefs about our own experience could be not only false but delusional. The possibility of introspective delusion raises questions about the relationship between experience and belief in delusional belief formation and deserves further investigation.

Chapter 3 addresses one further possible reason to resist introspective delusions. It is easily observable in clinical cases that delusions often come together with hallucinatory experiences, raising the question of whether delusional beliefs result from taking a really bizarre but really powerful experience at face value. If so, delusional beliefs might be mistaken about the external world, but never about one's own experience.

I argue that the possibility of introspective delusions depends on how the relationship between hallucinatory experience and delusional belief is spelled out. Three main accounts have been put forward in the literature to explain the relationship between delusional belief and experience: endorsement, according to which the experience is taken at face value to form the belief, explanationist, according to which the experience causes a search for meaning and explanations that eventually gives rise to the belief, and rationalist, according to which the belief influences the experience and gives rise to hallucinations. I will argue that introspective delusions are not possible in endorsement accounts, but they are in explanationist or rationalist accounts. To illustrate this point, I will present and spell out an endorsement, explanationist and rationalist account of three candidate introspective delusions: Anton-Babinski Syndrome, thought insertion and supernumerary limb delusion.

Not all delusions will be explained uniformely by the same account. Different delusions are likely to vary in how close or distant they are to experience. The only way to investigate whether a particular delusion is an introspective delusion or not is on a case-by-case basis: as our methods of accessing experience from a second and third-person perspective become more advanced and precise, it wil be possible to measure experience and belief independently and compare the two. This will raise important ethical issues: it is crucial to point out that critically evaluating introspective reports should never mean overwriting the patient's perspective.

It is not always an epistemic mistake to form beliefs that are not in line with experience. In **Chapter 4**, we look at visual illusions, a compelling case of true belief diverging from experience of the same object. An example is the cornsweet illusions: even though we experience A as being darker than B, we form the belief that they are the same color upon knowing how the illusion works. This belief is maintained even though the conflicting experience persists.

This can be explained by modular theories of mind, according to which perceptual systems are informationally encapsulated, meaning that they do not have access to the same evidence. Some theories like Predictive Processing relax the modularity assumption and imply that perception and belief are not independent of each other, and that there is no clear-cut boundary between the two. It is thus puzzling how Predictive Processing can account for a persisting divergence between perception and belief.

Recent insights concerning the neural implementation of Predictive Processing may help elucidate this. Specifically, prior information is proposed to be approximated by mechanisms in both the top-down and bottom-up streams of information processing. While the former is context-dependent and flexible in updating, the latter is context-independent and difficult to revise. We propose that a stable divergence between perception and belief may emerge when flexible prior information at higher hierarchical levels contradicts inflexible prior information at lower ones. This allows Predictive Processing to account for conflicting percepts and beliefs while still maintaining a hierarchical and unitary conception of cognition.

In **Chapter 5**, I look at the relationship between belief change and powerful experiences, specifically within the context of psychedelic-assisted therapy. Belief change is a central part of psychotherapy: especially significant are beliefs about the self, as a big part of therapy consists in replacing harmful or self-loathing narratives with positive or productive ones. Recent literature, such as Chris Letheby's Philosophy of Psychedelics (2021) has argued that psychedelic experiences combined with psychotherapy might be a helpful aid to this process.

During the psychedelic experience, patients seem to acquire new beliefs which improve their mental well-being. However, are such beliefs true or are they powerful comforting delusions? Letheby believes that metaphysical beliefs resulting from psychedelic experiences are likely to be false, but it's not these that mediate therapeutic success. Changes in self-related narratives are the key to mental health improvements following psychedelic sessions.

I argue that changes in beliefs about the self are likely to be epistemically more successful than changes in beliefs about the external world, but not because people are more likely to form true beliefs about themselves about psychedelics. The real reason for this success is the capacity of the new beliefs to turn into self-fulfilling prophecies when people act according to them. When our beliefs about ourselves do not match our experience of ourselves, action can be the bridge between the two.

In **Chapter 6**, I bring the previous chapters together by highlighting implications, general conclusions and future directions. This dissertation takes seriously the possibility of fallacious beliefs about our experiences. The epistemic status of introspection is only partially rescued by self-shaping. One of the consequences of this is skepticism towards introspective incorrigibility, namely the claim that introspective reports can in principle never be corrected by third-person methods. It's important not to conflate this with claims about the ethics of treating introspective reports, such as the claim that it is acceptable for psychiatrists to overwrite the patient's perspective.

### 1.4. Methodology

This is a philosophical dissertation. I have not conducted an experiment, nor gathered any novel empirical data. However, it is not purely conceptual work, with the only aim of a-priori investigation of folk psychological concepts and armchair reflections. While the methods are those of philosophy, the aim is to contribute to discussions that are of interest to philosophers and empirical scientists alike.

One way to achieve this aim is uncovering philosophical assumptions that are in tension with empirical data, but still inform and guide much of the empirical research. Empirical studies about the mind cannot be independent from philosophical assumptions, because how we gather data and what conclusions we draw from them depend heavily on our conceptions of the mind and of philosophical concepts such as experience and belief.  As Fodor (1972) writes:

"Psychologists have not been able to stop doing philosophy, for one cannot think seriously about mentation without eventually dealing with the sorts of issues that presented themselves to Locke, Hume, Berkeley, and Kant. But they have often managed to stop noticing when they are doing philosophy, and from not doing it consciously, it is a short step to not doing it well".

This is the scope of Chapter 1, where I investigate one philosophical assumption that underlies a lot of empirical research: the special epistemic status traditionally granted to introspection, our ability to form beliefs about our own experiences through the first-person perspective. Because of the philosophical assumption that introspection is infallible, or at least protected from error, introspective reports are considered the gold standard in consciousness studies. Investigating the truth of this assumption is of great value not only for philosophers, but also to those scientists that wish to conduct empirical research about experience.

Another way to foster interdisciplinary collaboration is to use philosophical methods to generate empirically testable hypothesis. By exploring the conceptual space given certain assumptions, philosophy has the potential of putting forward hypotheses that have not been formulated yet, and can meaningfully contribute to the scientific landscape. This is what I attempt to do in Chapter 3: by rejecting or suspending judgement on the assumption of introspective infallibility, we open up the possibility of delusions about one's own experience. Such a possibility is at least in theory empirically testable - or at least, it will be in the future once we improve our methods of accessing experience independently from introspective reports.

Thirdly, philosophers can attempt to reconcile complex and puzzling data with existing theories. One example of this effort is Chapter 4 of this dissertation, which was co-authored by two philosophers and two neuroscientists. In the paper, we investigate whether a unified theory of the mind like Predictive Processing is in tension with empirical cases of persisting cases of divergence between perception and belief, and whether this tension can be resolved.

These are only some of the ways in which philosophy and empirical science can work together while investigating one common object. Theoretical analysis and collection of new data should proceed hand in hand, in order to avoid the accumulation of meaningless data or the development of philosophical hypotheses that are too detached from reality. However, interdisciplinary research presents important challenges, as it requires trust and communication between researchers with different areas of expertise. Building a common framework with vocabulary and practices shared across disciplines is one of the challenges of philosophy in cognitive sciences.

.

.

Most of the essays that this thesis is made up of have been published elsewhere before. Chapter 2 was published as a sole-author paper in the Review of Philosophy and Psychology. Chapter 3 is currently bring revised for publication in the volume "Routledge Handbook of the Philosophy of Delusion", edited by Ema Sullivan-Bissett. Chapter 4 was co-authored with Sascha B. Fink, Philipp Sterzer and Joshua M. Martin and it was published in Consciousness and Cognition. Chapter 5 was published as a sole-author paper in Philosophy and the Mind Sciences, as part of the Simposium on Chris Letheby's Philosophy of Psychedelics. The two book reviews in the Appendix were published by Philosophical Psychology, and the second one was co-authored with Kathleen Murphy-Hollies.

## 1.5. References

American Psychiatric Association, (APA). (1995). *Diagnostic and statistical manual of mental disorders (4th ed.)*. Arlington, VA: Author.

American Psychiatric Association, (APA). (2013). *Diagnostic and statistical* *manual of mental disorders (5th ed.)*. Arlington, VA: Author.

Bayne, T., & Pacherie, E. (2004). Bottom-up or top-down: Campbell's rationalist account of monothematic delusions. *Philosophy, Psychiatry, & Psychology*, 11(1), 1-11.

Bayne, T., & Pacherie, E. (2004). Experience, belief, and the interpretive fold. *Philosophy, Psychiatry, & Psychology*, 11(1), 81-86.

Bernáth, L. (2021). Can Autonomous Agents Without Phenomenal Consciousness Be Morally Responsible? *Philosophy & Technology* 34, 1363–1382. https://doi.org/10.1007/s13347-021-00462-7

Bermúdez, J. L. (2020). *Frame it again: New tools for rational decision-making.* Cambridge University Press.

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

Bressloff, P. C., Cowan, J. D., Golubitsky, M., Thomas, P. J., & Wiener, M. C. (2002). What geometric visual hallucinations tell us about the visual cortex. *Neural computation*, *14*(3), 473-491.

Campbell, J. (2001). Rationality, meaning, and the analysis of delusion. *Philosophy, Psychiatry, & Psychology*, 8(2), 89-100.

Carruthers, P. (2017). Mindreading in adults: Evaluating two-systems views. *Synthese*, *194*, 673-688.

Connors, M.H., & Halligan, P. W. (2015). A cognitive account of belief: a tentative road map. *Frontiers in psychology*, *5*, 1588.

Connors, M. H., & Halligan, P. W. (2020). Delusions and theories of belief. *Consciousness and Cognition*, *81*, 102935.

Fodor, J. (1972). Some reflections on LS Vygotsky's Thought and language. *Cognition*, *1*(1), 83-95.

Foley, R. (1992). The epistemology of belief and the epistemology of degrees of believing. *American Philosophical Quarterly*, 29, 111-122.

Harman, G., (1986), *Change in View*, Cambridge: Cambridge University Press.

Jack, A., Roepstorff, A. (2003). *Trusting the subject? The use of introspective evidence in cognitive science.* Imprint Academic, Exeter, UK; Charlottesville, VA

Jewanski, J., Simner, J., Day, S. A., Rothen, N., & Ward, J. (2020). The evolution of the concept of synesthesia in the nineteenth century as revealed through the history of its name. *Journal of the History of the Neurosciences*, 29(3), 259-285.

Levy, N. (2021). *Bad Beliefs: Why They Happen to Good People*. Oxford University Press.

Macpherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research, 24-62.*

Macpherson, F. (2013). The philosophy and psychology of hallucination: an introduction. *Hallucination: Philosophy and psychology*, 1-38.

Macpherson, F. (2017). The relationship between cognitive penetration and predictive coding. *Consciousness and Cognition*, *47*, 6–16. https://doi.org/10.1016/j.concog.2016.04.001

Madva, Alex. (2019). Equal Rights for Zombies? Phenomenal Consciousness and Responsible Agency. *Journal of Consciousness Studies.* 26. 117-40.

Naci, L. Sinai, L., Owen, A. M. (2017). Detecting and interpreting conscious experiences in behaviorally non-responsive patients. Neuroimage, 145 (Pt B) (2017), pp. 304-313, 10.1016/j.neuroimage.2015.11.059

Nagel, T. (1974). What is it like to be a bat?. *The philosophical review*, *83*(4), 435-450.

Pauen, M., & Haynes, J. D. (2021). Measuring the mental. Consciousness and Cognition, 90, 103106.

Pronin, E. (2009). The introspection illusion. *Advances in Experimental Social Psychology* 41: 1–67.

Ramachandran, V. S., & Hubbard, E. M. (2001). Psychophysical investigations into the neural basis of synaesthesia. Proceedings of the Royal Society, 268, 979–983.

Sacks, O. (2012). Hallucinations. Alfred A. Knopf.

Schwitzgebel, E. (2002). How Well Do We Know Our Own Conscious Experience. The Case of Visual Imagery. Journal of Consciouness Studies, 9 (5–6), pp. 35-53

Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. The Philosophical Review, 117 (2), pp. 245-273

Schwitzgebel, E. (2021) "Belief", *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2021/entries/belief/>

Noordhof, P., & Sullivan-Bissett, E. (2021). The clinical significance of anomalous experience in the explanation of monothematic delusions. *Synthese*, 199(3-4), 10277-10309.

Williams, John N. (1982). *Believing the Self-Contradictory.* American Philosophical Quarterly 19 (3):279 - 285.

Zeman, A., Dewar, M., & Della Sala, S. (2015). Lives without imagery—congenital aphantasia. *Cortex, 73*, 378–380.

## Chapter 2

## Introspection and Belief: Failures of Introspective Belief Formation

This work was published as "Caporuscio, C. (2021). Introspection and Belief: Failures of Introspective Belief Formation. *Review of Philosophy and Psychology*, 1-20." Slight revisions have been made to fit the dissertation format.

It was supported by funding from the Deutsche Forschungsgemeinschaft (RTG2386)

### 2.0. Abstract

Introspection has traditionally been defined as a privileged way of obtaining beliefs about one's occurrent mental states, and the idea that it is psychologically and epistemically different from non-introspective belief formation processes has been widely defended. At the same time, philosophers and cognitive scientists alike have pointed out the unreliability of introspective reports in consciousness research. In this paper, I will argue that this dissonance in the literature can be explained by differentiating between infallible and informative introspective beliefs. I will argue that the latter are formed similarly to beliefs about the external world, and are therefore susceptible to similar success and failure conditions. Understanding introspection as belief-like will help to locate possible sources of error in regular as well as in pathological cases, carrying relevant implications for the relationship between experience, belief, and delusion.

### 2.1. Introduction

Think of the following belief: “A storm is coming”. Your belief formation process is likely to be triggered by looking outside and noticing some dark clouds approaching. You know that clouds like these usually mean that it is going to rain soon. Then you might look for alternative evidence: you have a look at the weather forecast and read that evening showers are likely. All these sources of evidence are weighted together to infer the best possible interpretation: a storm is coming. You might then act according to this newly formed belief, for example by fetching the clothes drying on your balcony. If the evidence changes (for example, the sky gets cleared up by a sudden wind) you might re-evaluate your belief and update it. Despite this process of accumulation and assessment of evidence, your belief is still prone to ignorance and error: maybe the weather forecast was imprecise, or maybe your pessimistic attitude made you jump to conclusions about an innocuous passing cloud.

Now think of your belief that you are feeling anxious, or that you are in pain, or that you are having a visual experience of a certain kind. These are all introspective beliefs: beliefs that have as their object not the external world, but your occurrent mental experience. Are these beliefs radically different from your beliefs about the external world? Can you doubt and revise them the same way you can doubt and revise your belief that a storm is coming? Are they prone to the same errors, or do they benefit from a special epistemic status?

When trying to answer this question by appealing to experience, a dissonance emerges. On one hand, my occurrent mental states seem tangible and accessible in a way that no external fact can be. On the other hand, if I am asked to precisely describe what I am feeling, that certainty dissolves. How detailed is my experience outside of the center of my visual field? Is that tingling sensation I am feeling on my back pain, or is it itchiness caused by the fabric of my clothes? Am I anxious about a meeting I have in a couple of hours, or am I excited? Am I hungry, or am I just feeling peckish because I am bored?

Hohwy (2013) vividly describes the challenges we face when we try to answer questions on introspection's epistemic status from our subjective experience of it[3]:

> "When we introspect, the introspected state seems easily accessible, for example, the pain or colour experience is as it were right there; and introspection seems certain and sometimes beyond doubt [...]. But equally, when we introspect, it doesn't take much for the introspection to be elusive, fleeting, and uncertain: we are stumped for words when trying to describe precisely whether the experience was like this or like that; we find it hard to sustain an experience stably in introspection for any length of time and the experience often seems to slip out of grasp when we focus on its individual aspects. When we introspect it seems we harbour both attitudes: introspection seems both accessible and certain, and inaccessible and uncertain." (Hohwy 2013, p. 247)

This dissonance is mirrored in the philosophical debate about introspection. On one side, proponents of the difference thesis argue that introspection is psychologically and epistemically different from our capacity to acquire beliefs about the external world, and less prone to ignorance and error. A long philosophical tradition attributes to introspection at least some epistemic privileges, including infallibility, omniscience, incorrigibility, indubitability, truth-sufficiency or self-warrant(Descartes 1641; Locke 1690; Ayer 1956; Alston 1971; Chalmers 2003; Smithies 2012; Gertler 2012). On the other side, the unreliability of using introspection as a measure of conscious experience has often been highlighted, and empiricists and philosophers alike have pointed out how we often cannot trust our judgments about the contents of our minds (Schwitzgebel 2008; Pronin 2009).

[3] It should be noted that Hohwy's prediction error approach to introspection explains such dissonance differently than I propose. According to Hohwy, introspection is unconscious probabilistic inference of mental causes, which in turn are the current probabilistic winners of a perceptual or interoceptive inference. He argues that introspection feels certain because it targets a winning hypothesis that is represented as highly invariant and noise-free; however, trying to decompose the experience or focus on its individual aspects means decomposing the winning inference, which brings back noise and uncertainty (Hohwy 2013, p. 245–249). Instead, I will argue that the dissonance stems from a different degree of fallibility and protection from error between different types of introspective judgments.

In the first part of the paper, I will argue that this tension stems from a confusion between different types of introspective beliefs and judgments[4]. Some introspective judgments are indeed infallible, like “I am feeling this” (Gertler 2012). I will argue that the infallibility of such judgments derives from the fact that they are exclusively sensitive to the mental state they are about, and they do not depend on other sources of knowledge. For this reason, they are immune from error in a way regular beliefs are not. However, these judgments lack in other respects, such as the capacity to convey and communicate precise information about our conscious experience. If I try to make an informative judgment, for example one that describes my current experience as one of excitement rather than anxiety, I will lose infallibility, introducing the possibility of error. It is this last category of judgments that better captures what most people have in mind when they talk about introspection: the kind of introspective beliefs and reports that we need for self-knowledge, social cognition, psychiatry and consciousness research, or as Schwitzgebel calls it, “introspection in practical use” (Schwitzgebel 2011).

In the second part of my paper, I will focus on whether the difference thesis holds for informative introspection: are informative introspective beliefs fundamentally different or epistemically superior to beliefs about the external world? As pointed out by Smithies and Stoljar (2012), it is difficult to claim that introspection is psychologically similar to other cognitive faculties while maintaining that it is epistemically superior to such cognitive faculties. Following Schwitzgebel (2011), my strategy to undermine the epistemic difference thesis is to argue that, when introspection is informative, it has no relevant psychological difference from other ways of forming beliefs about the world. Using a recent example of a cognitive model of belief formation (Connors and Halligan 2015, 2020), I will argue that the same factors that can introduce error and ignorance when we form beliefs about our external world can do the same to our beliefs about our mental world. It may be harder to be mistaken about complex emotional states than about our basic phenomenal experiences (Peels 2016); however, in this paper I will make the case that no introspective judgment is completely shielded from the possibility or error.

[4] For the purposes of this paper, I will use both terms to refer to the products of introspection. The difference is subtle: beliefs are mental states, while judgments are mental acts. When a belief is formulated occurrently, it becomes a judgment. After being formulated, a judgment can become a background belief (Cassam 2010)

Looking at the psychological process behind informative introspection can provide fresh insight into potential sources of error: if introspection's success and failure conditions can be accounted for by a regular theory of belief, then failures of inference and rationality that disrupt our belief formation process can also disrupt our introspective process. This is directly relevant to the problem of using introspective reports in consciousness research, as it means that introspective reports should be handled with a certain amount of scepticism and awareness of potential influencing factors. Locating such factors will shed light not only on normal, daily introspective mistakes but also on potential pathological failures of introspective belief formation. Delusions are defined in the DSM-IV as "false beliefs based on incorrect inference about external reality [...]" (American Psychiatric Association 1995[5]); but if my account is on the right track, our beliefs about our internal reality can be just as irrational and wrong as those about our external reality.

## 2.2 The Dissonance of Introspection

### 2.2.1 Introspective Desiderata

For a belief or a judgment to qualify as introspective, in the way the term is used in contemporary philosophy of mind, it must meet some minimal criteria: it must be about our own current or recently passed mental experiences, and it must be obtained in a way that is first-person specific (Schwitzgebel 2010). These are necessary features of introspection; if a belief does not possess these features, it is not introspective. If we want to use introspective judgments as a measure of conscious experience, however, there are other features that are desirable: for example, informativeness and protection from error. I will argue that these features are not equally present in all introspective judgments, but are gradable. Introspective judgments can be placed on a spectrum from very informative to not informative at all, or from infallible to very prone to error.

[5] See Footnote 2

### Protection from Error

The first desirable feature of introspection is protection from error, derived from the Cartesian idea that at least some introspective judgments cannot be wrong. This captures the intuition that we have privileged access to our own experience, and therefore we have ultimate authority regarding our own mental states.

To be infallible, introspection needs to be an exclusive measure of experience, namely, it needs to be only determined by its target mental state and not by external influences such as confidence, ignorance, background beliefs or motivational factors. If a judgment about the mental state M is only determined by the presence of M, there is no room for error: whenever the judgment occurs, M must also have occurred. This strategy is successful in plausibly granting infallibility to such judgments: they are shielded from error because there is no place where error can enter the process.

Chalmers (2003) and Gertler (2012) provide some examples of what introspective judgments that are exclusively determined by their target mental states look like. Chalmers (2003) argues that we possess direct phenomenal concepts that are directly constituted by the phenomenal quality of the experience. Such concepts can be combined with appropriately aligned demonstrative concepts to form direct phenomenal beliefs: if R is the pure phenomenal concept constituted by an experience, "this is R" is a direct phenomenal belief. The content of direct phenomenal concepts and beliefs is "determined by the phenomenal character of [...] experience, in that it will vary directly as a function of that character in cases where that character varies while physical and other phenomenal properties are held fixed, and that it will not vary independently of that character in such cases" (Chalmers 2003, p.16).

A similar strategy is adopted by Gertler (2012). In her account, infallibility is granted to those introspective judgments that are exclusively grounded in an introspective demonstrative: "I am feeling this"[6]. In judgments of this kind, the epistemic intersects with the phenomenal: demonstratives are not epistemically

[6] "I feel this" might be open to external influences and background beliefs about what "feel" means than "This is R". However, both judgments are not open to error relative to the mental state they are about, that is expressed through a direct phenomenal concept or an introspective demonstrative.

rigid, meaning that whatever the content of my experience is, the judgment "I am feeling this" will be true.

Protection from error can be successfully grounded in exclusiveness: a judgment that makes reference to its target mental state by using an introspective demonstrative or a direct phenomenal concept is exclusively determined by that target mental state and therefore it is infallible, because its truth does not depend on anything other than the experience itself. However, protection from error is not the only scale to evaluate our introspective judgments. I will now turn to another desirable feature of introspection.

**Informativeness**

I use the term informative to refer to introspective beliefs and judgments that we can use to learn and share information about our mental states. Some examples of informative introspective judgments are "I am feeling a throbbing pain on the right side of my head", "I am feeling anxiety", or "I am having an experience of geometric visuals in my periphery". We need informativeness both for ourselves, to help us guide our own actions, and for interpersonal relations, to be able to share our experiences and mental states with others. Another context in which informativeness is fundamental is psychiatry: in order to identify pathological experiences and start a therapeutic process, psychiatrists need to be able to access the first-person experiences of their patients. A patient describing their emotional state to their therapist, or describing the content of their visual experience, is producing informative introspective judgments.

The informativeness of introspection is also important for research purposes: we use our phenomenal experience, that we can access introspectively, to formulate hypotheses about the workings of our mind that we can then test against empirical evidence. An example of an introspectively generated hypothesis that has then been supported with third-person methods is number-color synesthesia, in which showing numbers to synesthetic individuals elicits in them a perceptual experience of different colors associated with different numerals (Kriegel 2013). Third-person evidence of number-color synesthesia was only discovered relatively recently: by showing an array of numbers to synesthetic and control subjects, Ramachandran and Hubbard showed that the elicited colors, and not only the different shapes, had an effect on perceptual grouping in synesthetic individuals (Ramachandran and Hubbard 2001). However, the phenomenon of synesthesia has been known at least since the nineteenth century, thanks to the introspection of synesthetic individuals (Galton 1880). Without introspective judgments, it would have been impossible (or at least, much more difficult) to formulate a hypothesis that could be empirically tested, namely that the color-number association was a perceptual effect and not only a mnemonic or metaphorical association. For this purpose, informativeness is key: an uninformative introspective judgment like "I am feeling this" would not have achieved the same result, not even if it was coming from a synesthetic individual mentally pointing at an occurring number-color perceptual experience. What was needed was not an infallible judgment about the phenomenal experience in question, but one that could relate it to other concepts and experiences, in this case by conveying that it was similar to perception and different from memory, imagination, and metaphorical thinking.

### 2.2.2 Unpacking the Trade-off

Protection from error and informativeness are both continuous properties, meaning that introspective judgments may vary in how informative and error-prone they are. Exclusiveness and infallibility are instead dichotomous properties: they only concern judgments at the extreme end of the error-protection spectrum. A judgment like "I am feeling this" is infallible, but not informative; a judgment like "I am feeling a throbbing pain on the right side of my head" is highly informative but fallible. Most judgments will fall somewhere between the two extremes, and benefit from both properties to different degrees. However, it is still not clear how these two features are linked to each other.

I argue that the relationship between informativeness and protection from error can be understood as a trade-off. Informativeness derives from our capacity to conceptualize the phenomenal experience and interpret it in virtue of our background beliefs. As soon as any conceptualization, background belief, or external factor enters the introspective process, the introspective belief stops being exclusively determined by conscious experience. With exclusiveness, a degree of protection from error is also lost, because its truth does not only depend on the presence of the mental state, but on external factors and background beliefs: for the judgment “I am feeling a throbbing pain on the right side of my head” to be true, previous beliefs about what pain is and what it feels like, throbbing pain in particular, about where right and left are, and about the position of my head compared to the body should also be true. Instead, capturing the phenomenal experience purely through an introspective demonstrative like “I am feeling this” renders the judgment exclusive and infallible but uninformative, while interpreting and conceptualizing it adds informativeness but introduces fallibility (Fig. 1).

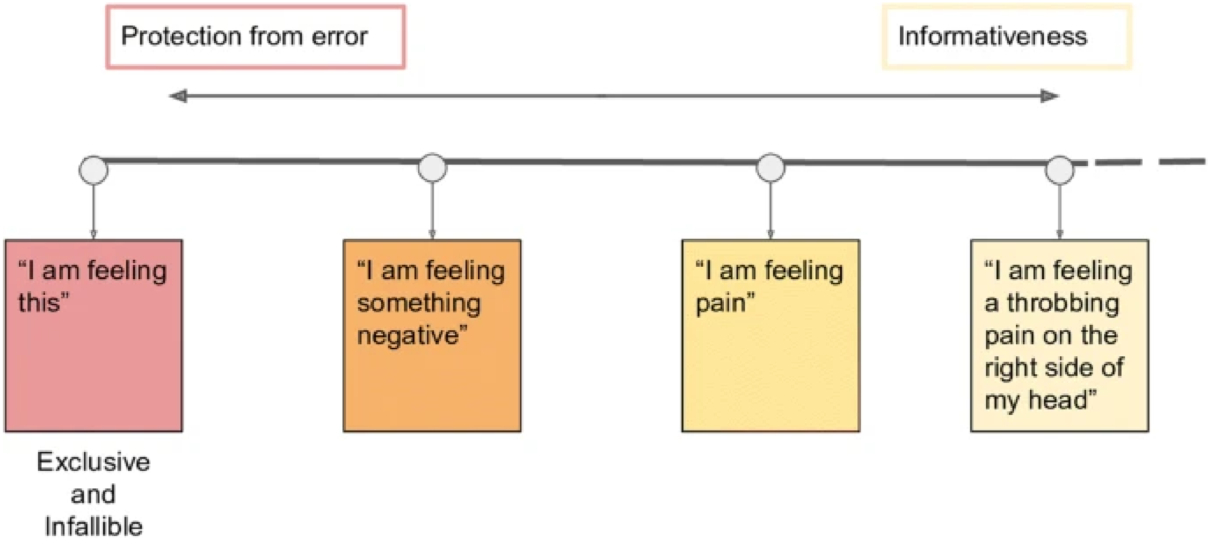

**Fig. 1:** The trade-off between informativeness and protection from error. This figure represents the trade-off between the two continuous properties that I believe to be at the center of the philosophical debate on introspection: informativeness and protection from error. “I am feeling this”, “I am feeling something negative”, “I am feeling pain”, “I am feeling a throbbing pain on the right side of my head” are all introspective judgments about the same mental state. Exclusiveness and infallibility only apply to the first judgment, while the last one is a highly informative belief that depends on many more assumptions, background beliefs, and inferences: the identification of a phenomenal state with the concept of throbbing pain, and the capacity to connect it to a specific bodily part. The second and third beliefs, instead, are somewhere in between: compared to “I am feeling this”, “I am feeling something negative” adds a little bit of information and introduces a small chance of error in attributing valence to the pure experience. “I am feeling pain” moves further towards the high informativeness and low protection from error end of the spectrum, but is still less informative and more secure than “I am feeling a throbbing pain on the right side of my head”

If this is on the right track, then infallibility only applies to a very restricted number of judgments, that inadequately represent our everyday experience of introspection and have very little use for practical purposes. I am always right when I say “I am feeling this”; however, most of our daily introspective judgments can be placed much further towards the “high informativeness-low error protection” end of the spectrum. Everyday instances of introspection are not infallible: I can be wrong when I say that I am having intrusive thoughts, that I am angry, jealous, or that I am feeling a burning pain. But does this mean they are formed similarly to regular beliefs, or do they still hold some kind of privilege? Informative introspection might be fallible, but should we expect it to fail under similar conditions as regular beliefs, or are we talking about a completely different process?

Relatively little attention has been given to informative introspective beliefs in the philosophical literature. One of the authors that has taken a closer look at introspective judgments that are not exclusive measures of experience is Schwitzgebel. In his 2011 paper, Schwitzgebel considers the kind of introspective judgments that are heavily informed by sources of knowledge other than the experience that they are about, resulting in the conclusion that most introspective beliefs do not involve one isolated introspective process, but rather a plurality of processes, or “a cognitive confluence of crazy spaghetti” (Schwitzgebel 2011, p. 19). According to Schwitzgebel, in our ordinary introspective judgments “pure introspection” does not exist, or it is entangled with non-introspective sources of knowledge to the extent that isolating it is impossible: a judgment like “I am feeling anxiety about next week’s exam” does not only involve directing my attention to my current phenomenology, but recruits proprioceptive bodily self-apprehension, knowledge of my social environment, mental simulation, self-shaping, inference, and other sources and processes. Such a scattered process of recruiting different sources of evidence to form judgments about one’s own mental states is unlikely to be underlain by a unique, separable cognitive process.

Schwitzgebel’s intuition does justice to the idea that most of our introspective judgments rely on much more than pure phenomenology and are therefore far from infallible. However, I do not believe that being pluralistic about introspective processes prevents us from advancing a general cognitive account with the aim of locating where, in this multifaceted process, errors are likely to happen. My proposal is to compare introspection to another multifaceted cognitive process by which we form judgments about the external world by recruiting different sources of evidence: belief formation.

In section 3.1, I will present a multistage model of regular belief formation advanced by Connors and Halligan (2015, 2020) that takes into consideration the variety of sources and processes that are likely to be employed when we form judgments about external objects or states of affairs, and posits some success and failure conditions. In section 3.2, I will argue that the same model can be applied to informative introspective belief formation. Such an account, I believe, can locate error factors in introspective reports more precisely that has been done so far in the literature, while maintaining the intuition that informative introspection is not an isolated cognitive process.

## 2.3 Informative Introspection and Belief Formation

### 2.3.1 Regular Belief Formation: Success and Failure Conditions

Despite the fundamental importance of the concept of belief in philosophy of mind, psychology and psychiatry, there is no well-accepted cognitive theory of belief (Coltheart 2007; Connors and Halligan 2015). It does not seem possible, for example, to locate the process of belief formation in the brain in a similar way as has been done for working memory, attention, or other cognitive processes. There is also a lot of controversy concerning whether beliefs are internal states (Fodor 1975), behavioral dispositions (Griffiths 1971), observable behavior (Dennett 1978), or whether they exist at all (Churchland 1981); however, some assumptions on how beliefs work are relatively uncontroversial. Beliefs are generally understood as attitudes[7] we have whenever we judge a certain proposition to be true: I believe that Berlin is the capital of Germany if I tend to assent to the statement "Berlin is the capital of Germany". Similarly, I believe that I am in a conscious mental state (for example, I am having a visual experience) if I believe that the proposition "I am having a visual experience" is true. Despite these

[7] Schwitzgebel (2002) has argued that beliefs are combinations of various kinds of dispositions.

controversies, various tentative accounts proposing candidate cognitive processes for belief formation have been proposed (David and Halligan 1996, 2000; Young 2000; Halligan and David 2001; Connors and Halligan 2015, 2020). These accounts are typically informed by research on delusional belief formation, and for this reason they pay particular attention to the conditions under which our beliefs are likely to go awry. In what follows, I will present a multi-stage cognitive account of regular belief formation advanced by Connors and Halligan (2015, 2020).

### The Five-Stages of Belief Formation

A recent hypothesis put forward by Connors and Halligan (2015, 2020) postulates a five-stage model of belief formation.

*The belief formation process is composed of a precursor (1), a search for meaning (2), an evaluation of candidate hypotheses (3), the acceptance of a belief (4) and the impact that the newly formed belief will have on new belief formation and lower-level processes (5).*

The precursor is a distal trigger that motivates the new belief and determines its content. For example, seeing some black clouds approaching can serve as a precursor for my belief that it is going to rain. Precursors can be perceptual inputs, social interactions, media, or memory traces; they can also involve more than one trigger. The main function of the precursor is to initiate the second stage of belief formation, namely the search for meaning: forming candidate hypotheses to explain the precursor. In this case, some possible proto-beliefs could be that it is going to rain, or that it is not. The third stage is an evaluation of those proto-beliefs, based on their capacity to explain the precursor and consistency with prior beliefs. If I believe that dark clouds bring rain, for instance, I might be inclined to choose the proto-belief that it is going to rain. Cognitive biases, emotions, and motivational factors also play an important part in this stage: it is likely that our brain evolved not only to favor beliefs with a high probability, but also beliefs that are useful for our survival and well-being (for the relation between utility and cognitive biases, see Galperin 2012; Haselton et al. 2015; Martin et al. 2021). The fourth stage is the belief itself, while the fifth stage is the impact that the new belief will have on lower-level processes, including perception, action and memory: if I believe that it's going to rain, I might be more aware of subtle raindrops starting to fall.

**Errors of Belief Formation**

The belief formation process can be understood as a way to make sense of the precursor. The precursor itself has very little doxastic content: the fact that I saw a dark cloud approaching conveys very little information if it is not backed up by the belief that meteorology is a science and that weather predictions on the basis of clouds are reliable. We need background beliefs to generate and compare hypotheses to explain the precursor; on the other hand, false background beliefs could easily lead to the formation of a new false belief, like in the case of the conspiracy theorist that believes that it is chemtrails, and not clouds, that bring rain. The process of belief formation involves giving up some protection from error in order to gain more informative beliefs that we can better use to act in our environment and communicate with others. This means that the process will produce false beliefs in the following cases:

*The five-stage belief formation process can produce false beliefs if I lack the background knowledge that would help me formulate the right proto-belief, if my background beliefs are false and lead me astray, or if my biases lead me to favor the wrong proto-belief.*

The interpretation of the precursor depends largely on our background beliefs and cognitive biases. The former have a fundamental role in stage two, where possible hypotheses are formulated, and in stage three, where they are evaluated on the basis of consistency with our belief system and capacity to explain the precursor. Our cognitive biases play a decisive part in stage three, where they can lead us to reject a high-probability hypothesis in favor of a high-utility one. Because our background beliefs and biases have such a strong influence on the formation and evaluation of candidate hypotheses to explain the precursor, there are different ways in which they can lead us astray and produce errors. Consider the following cases:

1. I see dark clouds approaching and I form the belief that it might rain soon. However, unbeknownst to me, the dark cloud is actually smoke from a fire a few

blocks away. In this example, lacking the appropriate background-beliefs leads to a failure of stage two: because I don't have the knowledge that would help me formulate the right proto-belief, the hypothesis that correctly explains the precursor is not even taken into consideration.

2. I am a conspiracy theorist and believe that it is chemtrails, and not clouds, that bring rain. So if I see dark clouds and no chemtrails, I will reject the hypothesis that it is going to rain and favor the proto-belief that it is going to be a sunny day. In this case, having the wrong background beliefs causes an error in stage three: I evaluate both the hypothesis of rain and the hypothesis of not-rain, but because I believe that rain is caused by chemtrails and not by clouds, I favor the hypothesis of not-rain.

3. I planned a picnic and I have strong motivational reasons not to want to believe it is going to rain. I reject the belief that it is going to rain despite its strong probability and consistency with prior beliefs. This is also a failure of stage three. However, in this case, it is motivational reasons and not beliefs that impair my probabilistic reasoning and lead me to reject the correct hypothesis.

It is also worth noting that the process of belief formation can be partially or completely unconscious: neither the precursor nor the background beliefs that play a role in the third stage are always transparent to us. Consider for example the following case (Lyons 2016; Senor 2008): I am looking at the sky and I form the judgment "This is a beautiful sunset." I cannot tell apart sunrises and sunsets just by looking at them, so the belief is epistemically and causally dependent on the prior beliefs that it is evening and not morning, and that the sun rises in the morning and sets in the evening. If I believed that it was morning, I would have formed the different judgment "This is a beautiful sunrise." However, I am not explicitly making these inferences in my conscious train of thought: the belief "This is a beautiful sunset" comes to mind in a seemingly immediate way. As famously argued by Nisbett and Wilson (1977), we often have to resort to confabulation and inferences when asked about the causes of our beliefs. Since the process of belief formation can be unconscious, I can be mistaken with regard to what triggered my beliefs or why I hold them: I can justify my beliefs with reasons that were irrelevant to my belief formation process and fail to identify factors that played an important role.

### 2.3.2 Informative Introspection: Success and Failure Conditions

In section 2, I argued that introspection, when informative, is not infallible. In what follows, I will compare introspective belief formation with regular belief formation and argue that the former can be understood as a subset of the latter. To this end, I will use Connors and Halligan's five-stage model of belief formation (2015; 2020) and argue that it can be applied not only to beliefs about the external world, but to introspective beliefs as well. It should be noted that my argument does not depend on accepting the five-stage model as correct: I only aim to use it as an example of a plausible, tentative account of how regular and introspective beliefs are formed and how they can fail.

#### Introspection and the Five-Stage Account

I argue that Connors and Halligan's account of belief formation can be applied to informative introspection. This can be phrased as follows:

*Informative introspective belief formation requires the same stages as regular belief formation, with the only difference being that in stage one, the process is triggered by a mental experience, and in stage four, the content of the belief is a proposition about the mental precursor.*

Not all beliefs with a mental precursor are introspective: the experience of seeing a green object, for instance, can trigger the introspective belief that I am having a visual experience of a green object but also the non-introspective belief that there is a green object in front of me. Similarly, not all beliefs about my mental experiences are introspective: if my only precursor for the belief that I am angry is my friend pointing out to me that I am exhibiting angry behavior, that belief will also not count as introspective. The process of belief formation counts as introspective only when a mental experience works as a precursor for a belief about that experience.

Let us think of a paradigmatic example of introspection in this light. I am feeling a sensation of discomfort. This sensation triggers a search for possible proto-beliefs that would explain it: it could be hunger, or it could be anxiety. After the search for meaning, comes the evaluation of proto-beliefs. I know that I have just eaten lunch and that I have a deadline coming up, so the proto-belief that I am feeling anxiety is the best one in terms of ability to explain the precursor, probability, and consistency with prior beliefs. In stage five, the belief "I am feeling anxiety" is accepted. In the final stage, the belief acts as a top-down influence to shape perception, evaluate new proto-beliefs, and so on: for example, I will be more likely to accept a future proto-belief whose content is consistent with the belief that I have just formed. It is important to note again that this process is not necessarily conscious: like the sunset example considered in the previous section, the belief "I am feeling anxiety" might seem psychologically immediate, but it is in fact casually and epistemically dependent on inferences and explicit or implicit prior beliefs like "I have a deadline", "I have just eaten lunch", "I usually am not hungry immediately after eating" and "I usually have anxiety before deadlines".

What is meant exactly by mental precursor, and how does it relate to regular precursors? A mental precursor can be understood as the raw access to a sensation, and it stands in relation to the formed belief "I have anxiety" as seeing a dark cloud stands in relation to the belief that it is going to rain: it precedes the search for meaning and hypothesis evaluation that are necessary to identify the sensation with the concept of anxiety, like seeing a dark cloud precedes the association of dark clouds with incoming rain. In this sense, the precursor can be understood as a judgment with high exclusiveness and low informativeness, like Gertler's introspective demonstrative (2012) or Chalmers' phenomenal belief (2003): because it is constituted by nothing more than the raw feeling, it is highly protected from error but contains very little information about the mental state in question. Through the search for meaning and the evaluation of candidate hypotheses, the precursor is interpreted and a new informative belief is created.

## Errors of Informative Introspection

Introspection has been traditionally regarded as a special way of obtaining knowledge about oneself: more direct, more reliable, less prone to error. However, as I have argued in the first section of this paper, introspection's protection from error stems from the exclusiveness of some introspective judgments, and thus loses its grip when it comes to more informative, non-exclusive judgments. The five-stage process serves the purpose of attributing meaning to the precursor by coming up with plausible hypotheses, connecting them with prior beliefs and concepts, and transforming the empty precursor into an informative introspective belief. As it happens with regular beliefs, this comes at a cost: together with exclusiveness, infallibility is also lost, exposing the belief to possible sources of error. I argue that informative introspective beliefs have very similar success and failure conditions as beliefs about the external world: lack of appropriate background beliefs, wrong beliefs or strong biases can contaminate the belief formation process, leading to introspective failures. Before exploring in detail what this means in the case of introspection, I will briefly reconstruct my argument. These are the premises that have been defended so far:

1. The five-stage theory: The belief formation process is composed of five-stages: a precursor (1), a search for meaning (2), an evaluation of candidate hypotheses (3), the acceptance of a belief (4) and the impact that the newly formed belief will have on new belief formation and lower-level processes (5).
2. Errors of the belief formation process: The five-stage belief formation process can produce false beliefs if I lack the background knowledge that would help me formulate the right proto-belief, if my background beliefs are false and lead me astray, or if my biases lead me to favor the wrong proto-belief.
3. Informative introspective belief formation: Informative introspective belief formation requires the same stages as regular belief formation, with the only difference being that in stage one, the process is triggered by a mental experience, and in stage four, the content of the belief is a proposition about the mental precursor.

If these premises are accepted, the conclusion must follow:

***Errors of Informative Introspection:*** *The belief formation process can also produce false informative introspective beliefs if I lack the background knowledge that would help me formulate the right proto-belief, if my background beliefs are false and lead me astray, or if my biases lead me to favor the wrong proto-belief.*

Consider the following cases and compare them with the ones presented in section 3.1:

1. I have intrusive thoughts. However, I lack the notion of intrusive thought and therefore I mistake my thoughts for desires. (Kind, ms)

In this example, lacking the appropriate background-beliefs leads to a failure of stage two: because I don't have the knowledge that would help me formulate the right proto-belief, the hypothesis that correctly explains the precursor is not even taken into consideration.

2. I am hungry, even though I have just eaten lunch: without my knowledge, I have a parasite in my body that causes me to remain hungry after having consumed a three-course meal. I also have a deadline tomorrow. I misinterpret the precursor and form the belief that I am experiencing anxiety for the deadline.

In this case, having the wrong background beliefs causes an error in stage three: I evaluate both the hypothesis of anxiety and the hypothesis of hunger, but because I believe that hunger after a full meal is improbable, I reject that hypothesis and favor the belief that I am experiencing anxiety.

3. I am angry at a friend for petty reasons. I know the reasons are petty and I don't want to be the kind of person who holds unmotivated grudges, so I form the belief that I am not experiencing anger even though I am.

This is also a failure of stage three. However, in this case, it's motivational reasons and not beliefs that impair my probabilistic reasoning and lead me to reject the correct hypothesis.

I have argued in section 3.1 that the regular belief formation process can be fully unconscious, and that neither the precursor nor the background beliefs that play a role in the third stage are always transparent to us. If my argument is sound, we would expect this to apply to introspective belief formation as well. This means that introspective errors, such as errors in regular belief formation, can go completely unnoticed; furthermore, it means that we might misidentify the precursor, and thus mistake a non-introspective belief (a belief triggered by a non-mental precursor) for an introspective one. Think of the following case:

4. I am participating in an EEG experiment. I distractedly look at the alpha waves on the screen and form the belief that I must be bored without realizing that my belief was triggered by the screen and not by a phenomenal experience. I still think I introspected, even though my belief was triggered by an external precursor.

By definition, this is not an introspective belief, as it is triggered by an external precursor and it is not first-person specific: a scientist looking at the same screen can easily come to the same conclusion in the same way. However, because the precursor is not transparent to us, the boundaries between introspective and non-introspective beliefs are difficult to assess.

Cases 1, 2 and 3 are all instances of introspective belief formation: they are beliefs about one's own experiences that are formed in a first-person specific way and triggered by a mental precursor, and so they satisfy the generally accepted conditions for a belief to qualify as introspective. Still, they are fallible; and the conditions under which they can fail are similar to the conditions under which beliefs about the world can fail. Furthermore, as Case 4 shows, we can never be sure whether someone's beliefs about their mental states are triggered by their own experience or by something else. I believe this undermines the psychological and epistemic difference thesis: introspective beliefs do not differ fundamentally from beliefs about the external world, neither in their psychological process or in their epistemic status. We can never be sure that people's reports about their mental states are really triggered by those mental states, even when they claim they are; even when beliefs are triggered by a mental precursor, we should be aware of potential influencing factors that might have introduced error in the process.

## 2.4 Objections and Future Directions

### 2.4.1 Objections

According to the five-stage model, precursor and accepted belief are separated by two intermediate steps: a search for meaning and an evaluation of candidate hypotheses. While this is plausible for a lot of the examples discussed in this paper, it might seem counterintuitive to apply it to those beliefs that seem to be formed spontaneously and more or less directly, making it hard to distinguish different phases. Carefully considering the weather forecast or assessing a complex emotional state seem very different from forming the belief “I see a pink car” or “there is a pink car here” based on a visual experience: while in the first two scenarios we might be consciously generating candidate hypotheses and assessing their probability or utility, in the latter it feels like we are jumping from precursor to belief without much space for generating or assessing candidate hypotheses.

This objection is easily resolved once we take a closer look at the five-stage model of belief formation. While all steps of the process can come to conscious awareness, they often do not, and it is likely that, at each level, a large number of automatic processes might be involved (Connors and Halligan 2015, 2020). In some cases, the path between precursor and belief might be automatized, or a proto-belief might be attributed an extremely high probability, making it superfluous to consciously entertain a search for meaning or an evaluation of proto-beliefs. The fact that stages two and three are not conscious, however, does not mean that this process is not happening in the background. As I have argued in section 2.2, informativeness derives from our capacity to conceptualize and interpret raw experiences in virtue of our background beliefs; in order to ascribe meaning to the precursor and produce an informative belief, stages two and three are always needed, even though they might be automatized or unconscious. Furthermore, because certain pathways are often automatized, it does not mean they are immune from error. A pink car passing by might automatically elicit the belief that there is a pink car to most people, but someone suffering from erotomania might interpret it as a secret love message, or someone with persecutory delusions might take it to reinforce their belief that the CIA is after them. Even without taking into account pathological cases, someone might mistake a painted car for a real one, or fall victim to a visual illusion. The same goes for introspective beliefs: for example, the unreflective belief that I am not angry at my friend can be influenced by motivational factors and unconscious biases. Thus, the fact that some beliefs feel immediate should not be taken to mean that they are psychologically direct or epistemically infallible.

A stronger objection comes from an idea often expressed in the literature on introspection (Moran 2001; McGeer 1996, 2008; Schwitzgebel 2011): namely, that of the self-shaping nature of introspection. These accounts emphasize our capacity to shape and determine our own states of mind. If introspection is self-shaping, its authority does not derive from an immediate, error-free detection of its object, but from our ability to regulate our mental states in accordance with the claims we make about them. This is a significant difference to non-introspective belief formation, whose objects are external and untouched by our capacity to self-regulate.

The self-shaping nature of introspection is particularly problematic as it plausibly interferes with the five-stage model, and specifically with stages three and five. In cases of high uncertainty, stage three (evaluation of proto-beliefs) is likely to involve taking a closer look at the precursor to tentatively test our hypotheses. In doing this while forming beliefs about external objects, these objects will not change: if I look back at the incoming clouds to test the hypothesis that they might be smoke coming from a close-by factory, this will not change the fact that they are black clouds. Doing this while introspecting, instead, might plausibly change the introspected state itself. When I examine my sensation of unpleasantness trying to figure out whether it is hunger or anxiety, I might try to think about my exam next week to test the hypothesis that I am feeling anxious about it, and this exercise is likely to trigger some anxiety. The self-shaping nature of introspection is also problematic for the last stage of belief formation: after evaluation and acceptance, the newly formed belief will have an impact on new beliefs and lower-level processes. In the case of introspective belief formation, these impacted mental states coincide with the object of the newly formed belief: thus, introspection might often turn out to be right, not thanks to an infallible capacity to deliver judgments in line with the pre-existing mental states, but by changing the mental states to be in line with its judgments.

I believe that the self-shaping of introspection is very plausible, and I agree that it might add back some immunity to error, at least in some cases. However, I have three observations in response. The first one is that this kind of first-person authority does not derive from an epistemic advantage, but from an agential one (McGeer 2008): as such, it is not a claim about how well we can know or detect our own mental states, but about how much control we can exercise over them. The epistemic difference thesis, instead, grounds first-person authority in a special or privileged way we come to know about our own mental states. The scope of this paper was not to debunk first-person authority in general but to argue, against the epistemic difference thesis, that the process of forming beliefs about our mental states is subject to similar errors as the process of forming beliefs about external objects and states of affair. It is still a valuable point that we do not always get it right about our mental states, even if we have, to some extent, the capacity to make it right.

The second observation is that the most intuitive version of self-shaping, as it has been defended by e.g., Schwitzgebel (2011), does not mean that first-person judgments are always right. Granted, in some cases introspective errors might be counterbalanced by self-shaping, but this does not mean that our mental states will always change in accordance with our judgments. If that was the case, introspection would merely come down to making decisions about what we want to experience, and any attempt to focus our attention to discover something about the contents of our mind would be trivial. Furthermore, if we believe that emotions or other mental states serve the evolutionary function of facilitating the organism's capacity to respond to threats and opportunities (Tracy 2014), always changing them as we please would be maladaptive. Instead, it is plausible that there is a relation of continuous adjustment between epistemic and agential power of introspection, one in which errors are still possible.

Lastly, the capacity to change the precursor in accordance with our beliefs might not be unique to introspection. According to the active inference hypothesis, action can be understood as a way to change sensory input to fit our predictions about it (Friston et al. 2006; Clark 2015). By actively testing our hypotheses or interacting with the environment as if they were true, we might to some extent be able to turn them into "self-fulfilling prophecies": for example, a teacher who believes that a student is exceptionally bright is likely to act towards them in a way that will maximize their chances of academic success (Rosenthal 2003). External objects, stimuli, or states of affairs can plausibly be shaped by our actions less radically than our own mental states and experiences; however, if the active inference framework is on the right track, this difference might be more superficial than it appears.

If we buy that the psychological process of introspection follows the same stages and is subject to the same errors as regular belief formation, some relevant implications follow. I will explore these in the next section.

### 2.4.2 Future Directions: Introspective Delusions?

In this paper, I have argued that the way introspective beliefs are formed does not differ fundamentally from the way regular beliefs are formed, and that it is susceptible to similar success and failure conditions. So far, I have mostly given an account of what this implies for daily instances of non-pathological introspection. However, I believe that this way of looking at introspection might bring fresh insight into pathological cases as well. While a full discussion of the implications for psychopathology goes beyond the scope of this paper, in this section I aim to introduce one of the questions that would benefit from following this line of investigation: namely, the question of whether introspective beliefs can be not only false, but delusional.

According to the DSM-IV, a delusion is a "false belief based on incorrect inference about external reality that is firmly sustained despite what almost everyone else believes and despite what constitutes incontrovertible and obvious proof or evidence to the contrary [...]" (American Psychiatric Association 1995). A lot of these criteria have been criticized: for example, there seem to be prima facie counterexamples of delusional beliefs that are not about external reality, but about mental and bodily states (Coltheart 2007). Among these counterexamples, some notable cases are blind patients who claim that they can see their doctors and hospital rooms (Carvajal et al. 2012; Chen et al. 2015; Goldenberg et al. 1995; Khalid et al. 2016; Martín Juan et al. 2018), schizophrenic patients who believe they can hear other people's thoughts (Hoerl 2001), patients who have lost the sense of smell that claim they are able to feel the scent of coffee (Sacks 2012), and patients who believe they can feel pain in limbs that are not anymore attached to their body (Halligan et al. 1993).

Because of controversial *prima facie* counter-examples, this definition has been changed in the DSM-V (2013) to exclude the reality condition. However, the question of whether such counterexamples actually entail false introspective beliefs remains. One reason to doubt it is that delusions are thought to be pathological failures in belief formation[8], and philosophers have long been skeptical of whether we could be as dramatically wrong in our beliefs about our own mental states as we are about external reality. The most widely accepted explanation of delusion formation postulates two factors responsible for the adoption and the maintenance of the delusion (Coltheart 2007; Davies et al. 2005; Langdon and Coltheart 2000). The first factor is an anomalous precursor, or a bizarre experience that causes an implausible proto-belief to be considered in the search for meaning. The second factor is a deficit in belief evaluation, like biases (e.g. jumping-to-conclusions or confirmation bias, Balzan et al. 2013; Corlett 2018) or motivational factors (Ramachandran 1996; Bortolotti 2015), that impairs stage three and leads to the endorsement of the wrong proto-belief despite its implausibility or the conflicting evidence.

This model has been applied to various cases of delusions about the external world: in Capgras delusion, for example, a damage to autonomic response in face processing causing a lack of affective response to familiar faces could trigger the implausible proto-belief that all one's friends have been replaced by identical impostors (Ellis et al. 1997), while a deficit in stage three could explain why this hypothesis is chosen among more plausible ones and maintained despite conflicting evidence and inconsistency with prior beliefs (Davies et al. 2001;

[8] It is not uncontroversial that delusions are beliefs (Jaspers 1963; Parnas 2004; Cermolacce et al. 2010). However, following Bortolotti (2009) and Connors and Halligan (2015 2020), I will build on the assumption that they are, and that they derive from failures of belief formation.

Coltheart 2010). If introspection differs psychologically and epistemically from regular belief formation, as many have claimed, it is not obvious how this model could be applied to beliefs about one's own mental states: introspective reports should be taken at face value, no matter how odd they sound. Patients might be wrong about the external presence of objects eliciting their experiences, but they are right about their experiences.

If my account is on the right track, however, introspection is not psychologically and epistemically different from regular belief formation. If introspection is susceptible to similar failure conditions as belief formation in non-pathological cases, there is no obvious reason why the same should not apply to pathological cases. It follows that the same failures that cause pathological beliefs about the external world could also cause pathological beliefs about one's internal world. I will call these "introspective delusions".

Consider Anton-Babinski Syndrome. Patients with this syndrome are cortically blind, and their stereotypical reports together with the lack of neural activations normally associated with hallucinations suggest that they are not having any kind of visual experience. However, they believe that they can see. If introspection works like regular belief formation, this can be accounted for by a two-factor theory, where an anomalous precursor triggers a bizarre proto-belief that is accepted because of a deficit in stage three. It has been suggested by Goldenberg et al. (1995) that at the heart of the Anton-Babinski experience there might be vivid acts of imagination, which could be the anomalous mental precursors triggering the implausible proto-belief "I can see". In addition to this, patients have strong motivational factors not to believe in their own illness, and to favor the proto-belief that they are seeing normally despite all evidence to the contrary (for example, their inability to interact normally with their environment, their doctors' advice, or the memories of what seeing felt like as opposed to imagining).

There are other reasons why someone might be skeptical about introspective delusions. Researchers widely agree that at the heart of delusional belief formation there is often a bizarre experience that determines its content; however, they disagree in how tight the link between experience and belief needs to be, or what role self-shaping plays in the maintenance of delusions. In-depth discussion of these issues goes beyond the scope of this paper; however, the argument that I presented here gives us a reason to reject at least one of the arguments against introspective delusions, namely the peculiarity of introspective belief formation.

## 2.5. Conclusion

In this paper, I have argued for an account that understands our daily instances of introspective belief formation as akin to regular belief formation. I think this has important advantages over theories that argue that introspection is fundamentally different from other ways of acquiring beliefs about the world. First, understanding introspection within a general theory of belief is a more parsimonious and efficient strategy, as it removes the need to postulate other cognitive mechanisms specific to introspection. Secondly, it targets and explains introspective beliefs that are informative, and that we can use to access and share information about our minds. Finally, understanding introspection as belief-like helps in locating possible limits and sources of error, and could give a plausible account of pathological delusions like Anton-Babinski Syndrome.

## 2.6. References

Alston, W. (1971). Varieties of privileged access. *American Philosophical Quarterly* 8 (3): 223–241.

American Psychiatric Association (1995). *Diagnostic and Statistical Manual of Mental Disorders (4th Ed.)*. Arlington, VA: Author.

American Psychiatric Association, (APA). (2013). *Diagnostic and statistical manual of mental disorders (5th ed.)*. Arlington, VA: Author.

Ayer, A.J. (1956). *The problem of knowledge.* Harmondsworth: Penguin books.

Balzan, R., Delfabbro, P., Galletly, C., and Woodward, T. (2013). Confirmation biases across the psychosis continuum: The contribution of hypersalient evidence-hypothesis matches. *The British Journal of Clinical Psychology/The British Psychological Society 52 (1): 53–69.*

Bortolotti, L. (2009). *Delusions and other irrational beliefs.* Oxford: Oxford University Press. https://doi.org/10.1093/med/9780199206162.001.1.

Bortolotti, L. (2015). The epistemic innocence of motivated delusions. *Consciousness and Cognition 33: 490–499.*

Carvajal, J.J.R., Cárdenas, A.A.A., Pazmiño, G.Z., and Herrera, P.A. (2012). Visual Anosognosia (Anton-Babinski Syndrome): Report of two cases associated with ischemic cerebrovascular disease. *Journal of Behavioral and Brain Science 02 (03): 394–398.*

Cassam, Q. (2010). Judging, believing and thinking. *Philosophical Issues 20: 80–95.*

Cermolacce, M., Sass, L. and Parnas, J. (2010). What is bizarre in bizarre delusions? A critical review. *Schizophrenia Bulletin 36: 667–679.* https://doi.org/10.1093/schbul/sbq001.

Chalmers, D. (2003). The content and epistemology of phenomenal belief. *Consciousness: New philosophical perspectives 220: 271.*

Chen, J. J, Chang, H. F., Hsu, Y. C., and Chen, D. L. (2015). Anton-Babinski syndrome in an old patient: A case report and literature review: Anton-Babinski syndrome. *Psychogeriatrics 15.1, 58–61.*

Churchland, P.M. (1981). Eliminative materialism and propositional attitudes. *The Journal of Philosophy 78 (2): 67–90.*

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind.* Oxford University Press.

Coltheart, M. (2007). Cognitive neuropsychiatry and delusional belief. *Quarterly Journal of Experimental Psychology 60 (8): 1041–1062.*

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

Connors, M.H. and Halligan, P.W. (2020). Delusions and theories of belief. *Consciousness and Cognition 81: 102935.*

Connors, M.H., and Halligan, P. W. (2015). A cognitive account of belief: A tentative road map. *Frontiers in Psychology 5: 1588.*

Corlett, P. (2018). Delusions and prediction error. In *Delusions in context*, ed. L. Bortolotti. Cham: Palgrave Macmillan.

David, A.S., and Halligan, P.W. (1996). Cognitive neuropsychiatry [editorial]. *Cognitive Neuropsychiatry 1: 1–3.* https://doi.org/10.1080/135468096396659.

David, A.S., and Halligan, P.W. (2000). Cognitive neuropsychiatry: Potential for progress. *Journal of Neuropsychiatry and Clinical Neurosciences 12: 506–510.*

Davies, M., Coltheart, M., Langdon, R., and Breen, N. (2001). Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, and Psychology 8.2–3: 133–158.*

Davies, M., Davies, A., and Coltheart, M. (2005). Anosognosia and the two-factor theory of delusions. *Mind and Language 20 (2): 209–236.*

Dennett, D.C. (1978). *Brainstorms.* MIT Press.

Descartes, R. (1641). Meditations on first philosophy. In *Descartes Philosophical Writings.* London: Thomas Nelson and Sons (1954).

Ellis, H., 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 (1384): 1085–1092.*

Fodor, J.A. (1975). *The language of thought.* Harvard University Press.

Friston, K., Kilner, J., and Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris 100 (1–3): 70–87.*

Galperin, A. (2012). Error management and the evolution of cognitive Bias. *Social Thinking and Interpersonal Behaviour 45 (63): 35.*

Galton, F. (1880). *Visualised Numerals.* Nature 21: 323.

Gertler, B. (2012). Renewed acquaintance. In *Introspection and consciousness,* ed. Declan Smithies and Daniel Stoljar, 89–123. Oxford University Press.

Griffiths, A.P. (1971). *Belief: The Gifford Lectures Delivered at the University of Aberdeen in 1960 by H. H. Price.* (The Muirhead Library: George Allen and Unwin 1969. Pp. 495.). Philosophy 46 (175): 63–68.

Goldenberg, G., Muellbacher, W. and Nowak, A. (1995). Imagery without perception—A case study of anosognosia for cortical blindness. *Neuropsychologia 33 (11): 1373-1382* https://doi.org/10.1016/0028-3932(95)00070-J.

Halligan, P.W., Marshall, J.C. and Wade, D. T. (1993). Three arms: a case study of supernumerary phantom limb after right hemisphere stroke. *Journal of Neurology Neurosurgery & Psychiatry 56 (2): 159-166* https://doi.org/10.1136/jnnp.56.2.159.

Halligan, P.W., and David, A.S. (2001). Cognitive neuropsychiatry: Towards a scientific psychopathology. *Nature Reviews Neuroscience 2: 209–215.* https://doi.org/10.1038/35058586.

Haselton, M.G., Nettle, D. and Murray, D.R. (2015). The evolution of cognitive Bias. In *The Handbook of Evolutionary Psychology, 1–20.* American Cancer Society.

Hohwy, J. (2013). *The predictive mind.* Oxford University Press.

Hoerl, C. (2001). On Thought Insertion. *Philosophy Psychiatry & Psychology* 8 (2): 189-200https://doi.org/10.1353/ppp.2001.0011.

Jaspers, K. (1963). *General psychopathology*. Chicago, IL: University of Chicago Press.

Juan, A. M., Madrigal, R., Etessam, J. P., San Baldomero, F. S. F., and Bueso, E. S. (2018). *Anton–Babinski syndrome, case report.* gercj93.11, 555–557.

Khalid, M., M. Hamdy, H. Singh, K. Kumar, and Basha, S. A. (2016). Anton Babinski syndrome - a rare complication of cortical blindness. *Galen Medical Journal 1 (1): 4.*

Kind, A. (ms). *The model based theory of psychiatric reasoning.*

Kriegel, U. (2013). A hesitant defense of introspection. *Philosophical Studies 165 (3): 1165–1176.*

Langdon, R., and Coltheart, M. (2000). The cognitive neuropsychology of delusions. *Mind and Language. 15 (1): 184–218.*

Locke, J. (1690). An essay concerning human understanding. *London: Thomas Bassett.*

Lyons, J. (2016). Unconscious Evidence. *Philosophical Issues 26 (1): 243–262.*

Martin, J. M., Solms, M.,  and Sterzer, P. (2021). Useful misrepresentation: perception as embodied proactive inference. *Trends in Neurosciences*, *44*(8), 619-628.

McGeer, V. (1996). Is "self-knowledge" an empirical problem? Renegotiating the space of philosophical explanation. *Journal of Philosophy 93: 483–515.*

McGeer, V. (2008). The moral development of first-person authority. *European Journal of Philosophy 16 (1): 81–108.*

Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge.* Princeton, NJ: Princeton University Press.

Nisbett, R., and Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review 84 (3): 231–259.*

Parnas, J. (2004). Belief and pathology of self-awareness: A phenomenological contribution to the classification of delusions. *Journal of Consciousness Studies 11: 148–161.*

Peels, R. (2016). The empirical case against introspection. *Philosophical Studies 173 (9): 2461–2485.*

Pronin, E. (2009). The introspection illusion. *Advances in Experimental Social Psychology 41: 1–67.*

Ramachandran, V.S. (1996). The evolutionary biology of self-deception, laughter, dreaming and depression: Some clues from anosognosia. *Medical Hypotheses 47 (5): 347–362.*

Ramachandran, V.S., and Hubbard, E.M. (2001). Psychophysical investigations into the neural basis of Synaesthesia. *Proceedings of the Royal Society of London. Series B: Biological Sciences 268 (1470): 979–983.*

Rosenthal, R. (2003). Covert Communication in Laboratories Classrooms and the Truly Real World. *Current Directions in Psychological Science 12 (5): 151-154https://doi.org/10.1111/1467-8721.t01-1-01250.*

Sacks, O. (2012). *Hallucinations.* Alfred A. Knopf.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs 36 (2): 249–275.*

Schwitzgebel, E. (2008). The unreliability of naive introspection. *Philosophical Review 117 (2): 245–273.*

Schwitzgebel, E. (2010). Introspection, in the *Stanford encyclopedia of philosophy (winter 2019 edition), Edward N. Zalta (ed.),* URL = <https://plato.stanford.edu/archives/win2019/entries/introspection/>.

Senor, T. D. (2008). Epistemological problems of memory. In the *Stanford encyclopedia of philosophy (winter 2019 edition), Edward N. Zalta (ed.),* URL = <https://plato.stanford.edu/archives/fall2019/entries/memory-episprob/>.

Schwitzgebel, E. (2011). Introspection, what? In *Introspection and consciousness, ed. D. Smithies and D. Stoljar, 29–48. Oxford University Press.*

Smithies, D. (2012). A simple theory of introspection. In *Introspection and consciousness, ed. Gerce and D. Stoljar, 259–294.* Oxford University Press.

Smithies, D., and D. Stoljar. (2012). Introspection and consciousness: An overview. In *Introspection and consciousness, ed. D. Smithies and D. Stoljar, 3–25.* Oxford University Press.

Tracy, J.L. (2014). An Evolutionary Approach to Understanding Distinct Emotions. *Emotion Review 6 (4): 308*-312 https://doi.org/10.1177/1754073914534478.

Young, A.W. (2000). Wondrous strange: The neuropsychology of abnormal beliefs. *Mind and Language 15: 47–73*. https://doi.org/10.1111/1468-0017.00123.

## Chapter 3
## Introspective Delusions

This work is currently being revised for publication in the volume “Routledge Handbook of the Philosophy of Delusion”, edited by Ema Sullivan-Bissett.

It was supported by funding from the Deutsche Forschungsgemeinschaft (RTG2386)

### 3.0. Abstract

Delusions are defined by the DSM-IV as false beliefs about external reality. However, it is unclear whether introspective delusions, namely delusional beliefs that are wrong about one’s own experience, are also possible. One reason to doubt this comes from the fact that delusion and hallucinatory experience seem to go hand in hand, suggesting a strong relationship between the two. In this chapter, I will review the existent literature on the relationship between experience and belief, and spell out the consequences for the possibility of introspective delusions. I argue that the current understanding of the relationship between experience and delusional belief leaves space for the possibility of introspective delusions; however, as our methods to access experience independently from introspective reports are still imperfect, it is still controversial whether specific candidates like Anton-Babinski Syndrome might be introspective delusions.

### 3.1. Introduction

A widely studied symptom in the psychiatric population is the presence of delusions, namely “false beliefs based on incorrect inference about external reality that are firmly sustained despite [...] what constitutes incontrovertible and obvious proof or evidence to the contrary. [...]” (American Psychiatric Association, 1995)[9]. Delusions vary in content and origin: they can be caused by localized brain damage, typically resulting in monothematic, insulated beliefs (“my father is an impostor”, “mirrors are windows to another reality”, “I am dead”), or they can emerge as symptoms of an organic condition, like schizophrenia or bipolar disorder. Organic delusions are often polythematic, involving more than one belief and gradually spreading into an interconnected web of false convictions.

The DSM-IV definition seems to suggest that delusions can only be about external reality; however, as Coltheart (2007) and Langdon (2011) point out, some cases call this assumption into question. Sometimes, delusional agents have beliefs about their own experiences (and not just about external reality) that appear odd at the very least: blind subjects that claim that they are seeing (Carvajal et al., 2012; Chen et al., 2015; Khalid et al., 2016; Martín Juan et al., 2018), schizophrenic subjects that believe they can hear other people’s thoughts (Fernández, 2010; Hoerl, 2001; Pickard, 2010; Sollberger, 2014), people who have lost the sense of smell that claim they are able to feel the scent of coffee (Sacks, 2012), and survivors of traumatic injuries that believe they can feel pain in limbs that are not any longer attached to their body (Halligan et al., 1993; Lotze et al. 2001). These cases lend themselves to two possible interpretations: (1). subjects could be having extremely peculiar experiences that go undetected by third-person perspective but are accessible to introspection, or (2). they could be wrong about their own experience. The first interpretation preserves the DSM-IV assumption that subjects jump to inaccurate conclusions about the external world, but they are not wrong about the internal world of their own experience. If the second interpretation is correct for at least some of these cases, instead, these

[9] See Footnote 2

should be treated as introspective delusions (henceforth ID): false pathological beliefs whose content includes one or more introspective mistakes.

In this chapter, I will review the debate concerning delusions and introspection. The possibility of IDs depends on how the relationship between hallucinatory experience and delusional belief is spelled out. First, I will introduce the relevant debates concerning delusions and introspection and clarify the assumptions that make the concept of being deluded about experience possible. In the second part of this chapter, I will discuss the entangled relationship between experience and delusion and focus on three case studies: Anton-Babinski Syndrome, thought insertion and supernumerary limb delusion. I conclude that only endorsement models of delusional belief formations are incompatible with delusional beliefs that are not in line with experience, and it is unlikely that all delusions can be explained by a purely endorsement account. However, as our methods to access experience independently from introspective reports are still imperfect, it is still controversial whether specific delusions might come with introspective mistakes.

For the purposes of this paper, I will assume throughout that delusions are pathological beliefs, meaning that they are still somewhat responsive to evidence, action-guidance and integration in one's belief system, although they might fail to respond appropriately more often or more drastically than the norm.

### 3.2. Introspective mistakes

Introspection is our capacity to access our own conscious mental states and experiences from a first-person perspective (Schwitzgebel 2019). I refer to introspection as the process that starts with having an experience with a phenomenal quality and ends with the formation of a belief about that experience.

Introspection has been traditionally assigned a privileged epistemic status (Alston 1971; Chalmers 2003; Descartes 1641; Gertler 2012; Locke 1690; Smithies 2012). The extreme version of this claim goes as far as saying that introspection is infallible: no error can be possible in a genuine introspective judgement. On the other side, philosophers and psychologists like Schwitzgebel (2008) and Pronin (2009) have pointed out the shortcomings of introspection as an imprecise or inconsistent measure of conscious experience.

I have argued in a recent paper (Caporuscio 2021) that the disagreement here is based on different conceptions of what an introspective belief is and, consequently, what counts as an introspective error. According to some defenders of the incorrigibility thesis, pure introspective judgements are exclusively determined by their target mental state, therefore their truth does not depend on anything other than the experience itself. For example, the judgement "I am feeling *this*" cannot be wrong when *this* is a demonstrative that directly refers to the phenomenal quality of my experience (Gertler 2012). It is hard to doubt that these kinds of judgements are somehow privileged.

However, even if we might have uniquely infallible access to the raw what-it-is-like of our experience, this does not mean we always have the capacity to correctly translate that experience into a full-fledged introspective belief. That requires us to encapsule our mental states into concepts, relate them to each other and to the external world. Beliefs like "I am feeling happy", "I am having an experience of geometric visuals in my periphery" or "I am feeling a throbbing pain" are examples of this kind of introspective beliefs.

For the possibility of introspective delusions to get off the ground, we need to assume that full-fledged introspective beliefs are not fundamentally different from our beliefs about the external world - both start with an experience, followed by a search for meaning and a selection of candidate hypothesis that culminates in a new belief (for a full discussion of this, see Connors and Halligan 2015 2020; Caporuscio 2021). This means that the same failure conditions that cause mistakes and inaccuracies in our beliefs about the external world are threatening introspective beliefs as well: we can be mistaken when we choose the wrong concept to communicate or express our phenomenal experience, for example when we are led astray by our background beliefs or motivational factors. If I mistake hunger for anxiety, or furiously state that I am perfectly calm, I am committing an introspective error.

### 3.3. What is an Introspective Delusion?

So far, we have established four assumptions:

1. Delusions are beliefs that fail more drastically than the norm at the test of rationality (responsiveness to evidence, action-guidance and integration)
2. Pathological deviations from the norm in the process of belief formation can give rise to delusions.
3. Introspection is our capacity to form beliefs about our conscious mental states from a first-person perspective.
4. At least some introspective judgements are susceptible to errors in a similar way to regular belief formation.

Note that all assumptions can be resisted or argued against - but if all these premises are accepted, it follows that there can be beliefs about our conscious mental states formed from a first-person perspective that contain introspective errors, and that fail more drastically than the norm at the test of rationality.

5. Pathological deviations from the norm in the process of introspective belief formation can give rise to IDs

Even if this basic premise is accepted, the matter of IDs is still controversial. Delusions about the external world are easily spotted because psychiatrists can independently check the external world and notice inconsistencies with a patient's beliefs. To use a straightforwad example, if a patient is convinced that the earth is in the middle of an alien invasion but there are no flying spaceships in sight, the psychiatrist can safely conclude that the patient is holding false beliefs about the external world. This type of access is more problematic when it comes to accessing subjects' private mental states and experiences – especially in the context of delusions, that often come together with bizarre hallucinations. When diagnosing delusions about one's own mental state, experience and belief seem irreversebly entangled: delusions often come together with bizarre experiences, and methods to access one's experience independently from the reported belief are lacking. In the following sections, I will review the literature on the relation between experience and belief (3) and argue, with the help of case studies (4) that most of these accounts allow for a dissociation between experience and belief that would make IDs possible. Finally, I will address the question of how to disentangle experience and belief on a case-by-case basis (5).

### 3.4. From experience to belief and from belief to experience

Delusion and hallucination are as conceptually distinct as belief and perception: delusions are normally described in terms of pathologically irrational and false beliefs, while hallucinations are experiences that do not reflect reality. However, this conceptual distinction can become blurry in practice. Belief and perception can mutually influence each other: our beliefs about the world are often grounded in what we can perceive from our senses, and a long-standing debate in the philosophy of perception regards to what extent our percepts are cognitively penetrable, i.e. they can be influenced by beliefs and other higher-level mental states (Macpherson 2017; Marchi 2017; Newen and Vetter 2017).

There are good reasons to think that the link between delusions and experience might be even tighter than the one between belief and experience. It is easily observable in clinical cases that delusions often come together with hallucinatory experiences. Capgras delusion, with the paradigmatic content "my loved one has been replaced by identical impostors", comes hand in hand with the experience of having an impaired affective reaction to a familiar stimulus (Coltheart and Davies 2022; Nuara et al 2020). subjects claiming that mirrors are windows to a different reality, or that their doppelganger is following them wherever there is a reflective surface, are usually diagnosed with mirror anosognosia, or the inability to recognize reflected images. The belief "there are insects crawling under my skin" comes with the experience of tickling and itching, and subjects suffering from persecutory delusions typically experience paranoia and discomfort.

If strange experiences and irrational beliefs are observed together, two outstanding questions arise: which one came first and what is the causal relationship between them? This question is of particular importance in the regards to ID: if the content of the experience and the content of the delusion are always identical, or are completely shaped by one another, it is difficult to see how someone could be deluded about their own experience.

Two macro categories of answers have been given to these questions. Bottom-up or empiricist accounts ground the delusion in an abnormal experience (Bayne  and Pacherie 2004b). According to top-down or rationalist accounts, abnormal experiences are grounded in deluded beliefs and not vice versa (Campbell 2001).

### 3.4.1. Bottom-up accounts

Bottom-up accounts of delusional belief-formation, instead, explain the link between delusion and hallucination by arguing that the abnormal experience has a prominent causal role in triggering the false belief and determining its content (Bayne  and Pacherie 2004a). In other words, subjects are deluded by experience (Sullivan-Bissett 2020, Nordhoof  and Sullivan-Bissett 2021): the reason why they adopt and maintain such bizarre beliefs in spite of strong counterevidence and inconsistency with prior beliefs is their capacity to make sense of their bizarre, private internal reality. The hallucinatory experience comes first, and shapes belief. A similar intuition is shared by some versions of the predictive processing account of delusional belief-formation, according to which delusions are formed and maintained because the costs of leaving a salient experience unexplained are higher than the costs of making highly significant revisions to the rest of the belief system. In Clark's words, the deluded brain forms 'increasingly bizarre hypotheses so as to accommodate the unrelenting waves of (apparently) reliable and salient yet persistently unexplained information' (Clark 2016, p. 206).

Bottom-up accounts are committed to the idea that subjects are deluded *because* of the anomalous experience, which prompts the bizarre hypothesis to be considered and determines its content. This causes a prima facie reason to resist IDs: if subjects are deluded *by* experience, can they also be deluded *about* experience? If we concede that delusions emerge from powerful hallucinations, is there room for a dissociation not only between reality and experience, but between experience and belief as well? Can someone be deluded by experience and about experience at the same time?

The answer to this question will vary in different versions of the empiricist claim. Empiricist theories of delusion formation agree that the aetiology of delusions can be understood as bottom-up: irrational beliefs are caused by unusual experiences. However, there is one important respect where disagreement still stands, namely in how tight the link between experiential and doxastic content needs to be. Is the bizarre belief just formed by taking the experience at face value, or is there more to the delusional content?

According to endorsement theories of delusion formation, delusions are the result of the patient doxastically endorsing the content of their unusual experience. If delusions are simply an endorsement of experience, then the subjects' beliefs about their own experiences are always correct: their mistake resides only in the generalization from "I experience the world as if P were true" to "P is actually true". This suggests that delusions arise from a dissociation between reality and experience, and not between experience and belief: a delusion is a hallucination so believable that it drives belief-formation astray. Endorsement theories are therefore committed to the claim that there can be no dissociation between experience and belief: delusional subjects believe what they experience.

Explanationist theorists, instead, believe that the link between experiential and doxastic content is much less tight than what is defended by endorsement models. According to these accounts, the delusional belief serves to provide an explanation for an unusual experience. The content of the delusional belief does not need to match the content of the unusual experience, because the experience only works as an initial precursor that triggers a search for meaning (often unconscious: see Bongiorno  and Bortolotti 2019) and urges the patient to come up with an interpretation, which can be very far from the original content of the experience (Connors  and Halligan 2015 2020).[10]

### 3.4.2. Top-down accounts

In top-down accounts, delusions are not caused by unusual experiences, but involve a "top-down disturbance in some fundamental beliefs of the subject,

[10] It should be noted that "endorsement/explanationist" is not the same as "one factor/two factor accounts", because the first taxonomy refers to how a delusional belief acquires its content, while the second factor is supposed to explain why the delusional belief is mantained despite counterevidence. Thus, nothing stops explanationist theorists from endorsing one-factor accounts, or endorsement theorists from endorsing two-factor accounts.

which may consequently affect experiences and actions" (Campbell 2001, p. 89). Distorted belief comes first: because of motivational factors (Gunn and Bortolotti 2018), cognitive disturbances or organic malfunction, the subject comes to be convinced of something bizarre. It is only as a consequence of the delusional belief that hallucinations start to happen. This process is known as cognitive penetration, namely, the influence of higher-order cognitive states on perceptual experience (Macpherson 2012, 2017). For example, the belief of being followed can turn a perfectly normal experience of a day at the beach into a paranoid nightmare of shadows and creepy noises, and the phobia of insects can create the hallucinatory experience of itching.

This is not incompatible with IDs. Top-down disturbances affecting experience and belief do not necessitate that the two must always be aligned: while it is likely that some elements of the belief will trickle down to experience, this does not mean that the entirity of the subject's experience will be shaped by their beliefs (Macpherson 2017).

Summing up: ID are not compatible with endorsement models of delusional belief-formation, but they are compatible with explanationist and top-down models. Note that I am not arguing that adopting an explanationist or top-down account necessitates IDs, but merely that these account make it plausible that there might be delusions where agents come to endorse a false belief about their experience.

### 3.5. Case studies

In this section, I will illustrate the point made in section 4 by developing competing interpretations of specific delusions. Depending on which account is preferred (top-down, explanationist or endorsement) it is possible to interpret these cases both as an ID as as not an ID. The three candidate IDs are thought insertion, namely the false belief that one's thoughts belong to someone else, supernumerary limb delusion, namely the false belief of possessing and experiencing a limb that no longer exists, and Anton-Babinski Syndrome, namely the false belief of having visual experiences after becoming blind.

### 3.5.1. Thought insertion

“Thoughts are put into my mind like “Kill God”. It is just like my mind working, but it isn’t. They come from this chap, Chris. They are his thoughts.”
(Frith, 1992, p. 66)

This is an example of thought insertion, a common delusion in schizophrenic subjects that involves experiencing one´s thoughts as someone else´s. Prima facie, thought insertion may look like a delusion about one´s own experience: subjects think that they are hearing someone else´s thoughts, while their true experience is one of thinking, or hearing voices. However, different accounts of thought insertion have something different to say about whether or not this delusion involves a belief that is mistaken about one’s experience.

Sollberger (2014) put forward an endorsement account of thought insertion. According to this model, subjects do not only believe that their thoughts belong to someone else, but they also experience them as if they belonged to someone else: the attribution of their thoughts to an external entity is not a doxastic stance on their experience, but it is part of their experience itself. Their mistake, then, is not introspective: their access to their own experience is flawless, but they incorrectly take the experience at face value and generalize it to the external world. They are right in believing that they are having an experience as of inserted thoughts, but they are wrong in believing that someone is actually inserting thoughts in their mind. In this account, thought insertion is not an ID: it´s a false belief about the external world (someone else is thinking these thoughts), but not about their experience (thoughts without a sense of ownership).

According to the explanationist model put forward by Pickard (2010), instead, schizophrenics disown mental events if they are manifestations of mental states they do not endorse: the impulse of laughing without a background state of happiness, or intrusive thoughts that express beliefs they do not recognize as their own. Despite this, they are unable to suppress the mental states in question. In this account, the experience of thought insertion subjects is extremely minimal: all there is in the experiential state is a very salient mental event that they do not endorse or identify with. It is the (faulty) interpretation of this experience that leads them to form the delusional introspective belief that the disowned thoughts are not theirs, but are inserted from an external agent.

A possible top-down interpretation of thought insertion could hold that the thought insertion patient has severe paranoia, which causes them to believe that they can hear other people's thoughts (for example, negative things about the delusional person). Eventually, this belief causes auditory hallucinations and the experience of thought insertion. Paranoia is a known symptom of schizophrenia, that is often associated with thought insertion.

In the last two interpretations of thought insertion, there is a dissociation between the delusional person's experience and their belief about the experience. Their belief is a mistaken interpretation of their phenomenal state (ID).

### 3.5.2. Anton-Babinski Syndrome

Here, a doctor (G.G.) is asking a blind patient (H.S.) about her vision.

"G.G.: What can you see of me?
H.S. The head and... you are wearing a white coat
G.G. (covers his face with a black fan): Do you see my eyes?
H.S.: Yes.
G.G.: Do I wear glasses?
H.S: I think not."

(Goldenberg et al., 1995)

This is an example of Anton-Babinski Syndrome, namely the false belief[11] held by blind subjects that they are still able to see. Unlike Charles-Bonnet Syndrome, a similar condition where blindness is caused by peripheral damages, leaving the visual cortex intact and free to conjure up dream-like hallucinations (Kazui et al. 2009), in Anton-Babinski syndrome blindness is caused by severe damage to the

[11] It is controversial whether Anton-Babinski syndrome subjects really believe that they are seeing, as their reports are sometimes treated as confabulatory. A detailed discussion on the distinction between confabulation and delusion goes beyond the scope of this chapter. For the sake of this example, I will assume that subjects believe what their reports suggest.

visual cortex itself, making it doubtful whether visual experiences can be possible at all.

Endorsement models of ABS claim that subjects are undergoing experiences that are are phenomenally indistinguishable from perception: hallucinations (Allen-Hermanson 2017) or vivid imagination that is phenomenally indistinguishable from perception (Goldenberg 1995). The Anton-Babinski patient takes that experience at face value to endorse the belief that they are not blind. Thus, ABS is not an ID in this account.

Churchland (2002) and Macpherson (2010) disagree with this account, and defend the view that ABS subjects have an impaired access to their own experience. According to their accounts, ABS is an ID. The difference-maker between a regular person experiencing a vivid act of imagination and a delusional patient suffering from Anton-Babinski syndrome is not to be found in the experience itself, but in motivational factors and biases that affect the subjects' belief but not (or to a lesser extent) the experience. This means either that ABS subjects have a special kind of experience that they interpret as vision (explanationist account), or that there is no kind of experience present, and the delusion is entirely caused by motivational factors and difficulty in accepting one's own loss of sight, leading to anosognosia, namely cognitive anawareness of one's condition.

### 3.5.3. Supernumerary limb delusion

Here, the participant (P) is talking to the experimenter (E) about his non-existent third hand.

"E. Does it get cold?
P. Yes, it does get cold.
E. Can you feel it?
P. Yes, I do!
E. So sometimes this third hand gets cold?
P. Yes, it does."

(Halligan et al. 1993)

This is a report from a case of supernumerary limb delusion. According to Halligan and colleagues (1993), destruction of the sensory roots leads to the phenomenological experience of a supernumerary limb. Phantom pain has also been widely studied as a relatively common phenomenon following the loss of a limb (Di Pino et al. 2021; Lotze et al. 2001; Melzack, 1990). If subjects can feel a limb that is not there, it would simply take an endorsement of that experience to come to believe that they possess such limb. Supernumerary limb delusion could only count as an ID if experience and belief were further apart: for example, if the experience of deluded subjects amounted to purely imagined pain or other sensations in the missing limb (explanationist account) or if their delusion were purely driven by motivational factors and refusal to accept one's illness, without any background experience in the missing limb (top-down account).

The cases I've discussed in this chapter show that the close relationship between experience and belief does not rule out the possibility of IDs – at least not unless we claim that accounts of delusional belief-formation should be exclusively limited to endorsement models. In the next section, I will argue that this is implausible: delusions are more likely to differ in whether and how much belief is dissociated from experience. Whether a specific delusion entails false introspective beliefs or not is better assessed on a case-by-case basis.

### 3.6. How can we know?

The literature on delusional belief formation widely agrees that it is unlikely that all delusions might be explained in the same way. Langdon and Bayne talk about a received-reflective spectrum (Langdon 2011; Langdon and Bayne 2010): on one end lie delusions that arise directly from experience, while on the other lie delusions whose content is elaborated and therefore distant from experience.

Consider the following reports:

1. "FE believed that his own reflection was another person who was following him around, not only in his home, but anywhere that there was a reflecting surface" (Breen et al. 2001, p. 240).

2. "As I walked along, I began to notice that the colors and shapes of

everything around me were becoming very intense. And at some point, I began to realize that the houses I was passing were sending messages to me: Look closely. You are special. You are especially bad. Look closely and ye shall find. There are many things you must see. See. See. I didn't hear these words as literal sounds, as though the houses were talking and I were hearing them; instead, the words just came into my head – they were ideas I was having. Yet I instinctively knew they were not my ideas. They belonged to the houses, and the houses had put them in my head" (Saks 2007, p. 27).

3. "A woman [reports that she] is plagued by wireless phones, the blue is put upon her. She has under hypnosis had many children with 'astronomas'. Astronomas are different people, who are mutually identical ... astronomas speak to her and can perform 'indications'. That is what one sees. If the doctor kills you, they can perform an indication."(Strømgren, 1956)

The first report seems easily explainable with an endorsement-like model of delusion. FE had significant face processing deficits and deficits in his ability to interpret reflected space (Breen et al. 2001): his experience of looking at the mirror was plausibly the same experience as if he was looking through a window and seeing a face he could not recognize. The second report is harder to analyze in endorsement terms. While it is still plausible to imagine an experience with the content "these thoughts are not mine", the report suggests otherwise: Saks talks about her experience as extremely similar to her regular experience of thinking ("the words just came into my head – they were ideas I was having") but with an added component that she describes in cognitive terms ("I instinctively knew they were not my ideas"). Furthermore, there are elements to her experience that do not seem to be mirrored in the delusion ("The colors and shapes around me were becoming very intense"). In the third report, an experience in line with the delusion seems downright impossible: the woman had a number of apparently unrelated beliefs (some about wireless phones, some about astromas, and some about the doctors), her delusions are extremely conceptual and difficult to imagine as experiences. The hypothesis that she formed these beliefs in an attempt to interpret a vague but persistent experience of discomfort and paranoia seems more plausible.

How to distinguish received from reflective delusions? I believe this is better assessed on a case-by-case basis. Delusions where the experience is understood as being highly specific and in line with the delusional belief are more likely situated at the received end of the spectrum. Delusions stemming from vague discomforting experiences are instead likely to be more elaborated. At the current state of research, our methologies to access experience independently from introspective reports are far from perfect, both in terms of spatial and temporal resolution and in terms of our understanding of how brain processes map into experience. Thus, disentangling experience and belief will be easier for some delusions than for others.

There are good reasons to think that supernumerary limb delusion is at the received end of the spectrum. Research with fMRI and other neural imaging methods has identified neural correlates of phantom limb sensations and phantom limb pain (Lotze et al. 2001) and subjects' reports and behaviour offer ulterior support to the hypothesis that phantom pain is really being experienced, and not only imagined. The origin of these sensations has also been studied: according to a popular theory, phantom pain emerges from a mismatch between the lack of sensory signals from the amputated limb and its preserved representation and movement attempts by the neuromatrix, a widespread neural network representing the bodily self (Melzack 1990; Di Pino et al. 2021). This suggests that the precursor triggering the delusion is the phenomenological experience of pain and other sensations in a limb that does not exist. The hypothesis that is selected to explain the precursor ("I have a supernumerary limb") is excessively driven by experience, while previous knowledge and other sources of evidence are overruled. Motivational factors might play a role in accepting experience as correctly representing reality despite counterevidence. In this model, supernumerary limb delusion does not entail any introspective error, but only a false inference from experience to external reality.

The Anton-Babinski case seems different for a number of reasons. First, the extent of the damage to the visual cortex present in Anton-Babinski subjects raise doubts regarding whether these subjects can have visual experiences at all. Furthermore, the reported visual images seem to derive from confabulation, memory and synesthetic imagination: subjects are likely to report that their doctor is wearing a white robe because they know they're in a hospital, or that they can see a match being lit when they hear the sound or feel the warmth (Redlich  and Bonvincini, 1911; MacPherson 2010). subjects have strong motivational reasons not to want to believe they've gone blind: thus, they might mistake the experience of vividly imagining something for a visual experience. Neither of these elements is proof that the subjects' experience is not one of vision, and more investigation is needed as we shed light on the neural underpinnings of perception, hallucination and imagination. However, this is a case where there are at least have good prima facie grounds to doubt the patient's report of their introspected experience. The possibility that this is a delusion involving an introspective error should be taken seriously.

The absence of a specific physical or neurological deficit makes it difficult to investigate the experience of thought insertion subjects. Thought insertion is often a symptom of schizophenia, whose complex phenomenology and unclear neural underpinnings make it difficult to isolate a specific delusion from its surroundings. However, as the mechanism and origins of schizophrenic delusions become more clear, it might be possible to shed light on this case.

### 3.7. Concluding Remarks and Future Directions

The possibility of IDs opens the door for rejecting introspective incorrigibility, namely the claim that in case of disagreements between the subjects' introspective access to their experience and any kind of external access, introspection will always be right. The example of Anton-Babinski syndrome suggests that, as our third-person methods of accessing conscious experience get refined, introspective reports might be corrected by external observations. This raises an important question for psychiatric practice: Is it ever ethically acceptable to question a patient's introspective report based on second or third-person evidence?

Further work is needed to explore the ethical dimension of rejecting introspective incorrigibility. However, it is crucial to point out that critically evaluating introspective reports should never mean overwriting the patient's perspective. On the contrary, psychiatrists should focus their interpretative efforts on understanding why subjects come to believe what they believe given what they are experiencing: when evaluating disputes, dialogue and empathy are necessary for successfully considering first, second and third-person perspectives and come to a conclusion together with the patient.

In this chapter, I have reviewed the literature on the relationship between experience and belief in delusional belief formation from the perspective of introspective delusions, namely, false pathological beliefs about one's experience. I have argued that the current understanding of the relationship between experience and delusional belief leaves space for the possibility of introspective delusions: only endorsement models of delusional belief formations are incompatible with delusional beliefs that are not in line with experience, and it is unlikely that all delusions can be explained by a purely endorsement account. However, as our methods to access experience independently from introspective reports are still imperfect, it is still controversial whether specific delusions like Anton-Babinski Syndrome might be introspective delusions.

### 3.8. References

Allen‐Hermanson, S. (2017). Introspection, Anton's Syndrome, and Human Echolocation. *Pacific Philosophical Quarterly*, *98*(2), 171-192.

Alston, W. (1971). Varieties of Privileged Access. *American Philosophical Quarterly*, *8*(3), 223–241.

American Psychiatric Association, (APA). (1995). *Diagnostic and statistical manual of mental disorders (4th ed.)*. Arlington, VA: Author.

Bayne, T., and Pacherie, E. (2004a). Bottom-Up or Top-Down: Campbell's Rationalist Account of Monothematic Delusions. *Philosophy, Psychiatry, & Psychology*, *11*(1), 1–11. https://doi.org/10.1353/ppp.2004.0033

Bayne, T., & Pacherie, E. (2004b). Experience, Belief, and the Interpretive Fold. *Philosophy, Psychiatry, &amp; Psychology*, *11*(1), 81–86. https://doi.org/10.1353/ppp.2004.0034

Berrios, G. 1991. Delusions as “wrong beliefs”: A conceptual history. British Journal of Psychiatry 159:6-13.

Bongiorno, F., & Bortolotti, L. (2019). The Role of Unconscious Inference in Models of Delusion Formation. In *Inference and Consciousness* (pagg. 74–96). Routledge.

Bortolotti, L. (2009). *Delusions and Other Irrational Beliefs*. OUP Oxford.

Campbell, J. (2001). Rationality, meaning, and the analysis of delusion. *Philosophy, Psychiatry, & Psychology*, *8*(2–3), 89–100. https://doi.org/10.1353/ppp.2001.0004

Caporuscio, C. (2021). Introspection and Belief: Failures of Introspective Belief Formation. *Review of Philosophy and Psychology*, 1-20.

Carvajal, J. J. R., Cárdenas, A. A. A., Pazmiño, G. Z., & Herrera, P. A. (2012). Visual Anosognosia (Anton-Babinski Syndrome): Report of Two Cases Associated with Ischemic Cerebrovascular Disease. *Journal of Behavioral and Brain Science*, *02*(03), 394–398.

Chalmers, D. (2003). The Content and Epistemology of Phenomenal Belief. *Consciousness: New philosophical perspectives*, *220*, 271.

Chen, J. J., Chang, H.-F., Hsu, Y.-C., & Chen, D.-L. (2015). Anton-Babinski syndrome in an old patient: A case report and literature review: Anton-Babinski syndrome. *Psychogeriatrics*, *15*(1), 58–61.

Churchland, P. S. (2002). *Brain-wise: Studies in Neurophilosophy*. Cambridge, MA: MIT Press.

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190217013.001.0001

Coltheart, M. (2007). Cognitive neuropsychiatry and delusional belief. *Quarterly Journal of Experimental Psychology (2006)*, *60*(8), 1041–1062.

Coltheart, M., & Davies, M. (2022). What is Capgras delusion? *Cognitive*

*Neuropsychiatry, 27(1), 69-82.*

Connors, M. H., & Halligan, P. W. (2015). A cognitive account of belief: A tentative road map. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01588

Connors, M. H., & Halligan, P. W. (2020). Delusions and theories of belief. *Consciousness and Cognition*, *81*, 102935. https://doi.org/10.1016/j.concog.2020.102935

Currie, G., & Ravenscroft, I. (2002). *Recreative Minds: Imagination in Philosophy and Psychology*. Clarendon Press.

Descartes, R. (1641). Meditations on First Philosophy. In *Descartes Philosophical Writings.* Thomas Nelson & Sons (1954).

DI Pino, G., Piombino, V., Carassiti, M., & Ortiz-Catalan, M. (2021). Neurophysiological models of phantom limb pain: What can be learnt. *Minerva Anestesiologica*, *87*(4), 481–487. https://doi.org/10.23736/S0375-9393.20.15067-3

Fernández, J. (2010). Thought Insertion and Self-Knowledge. *Mind & Language*, *25*(1), 66–88. https://doi.org/10.1111/j.1468-0017.2009.01381.x

Frith, C. (1992). *The Cognitive Psychology of Schizophrenia*, Hillsdale NJ, Erlbaum

Gertler, B. (2012). Renewed Acquaintance. In D. Smithies & D. Stoljar (A c. Di), *Introspection and Consciousness* (pagg. 89–123). Oxford University Press.

Goldenberg, G., Muellbacher, W., & Nowak, A. (1995). Imagery Without Perception—A Case Study of Anosognosia for Cortical Blindness. *Neuropsychologia*, *33*, 1373–1382. https://doi.org/10.1016/0028-3932(95)00070-J

Gunn, Rachel & Bortolotti, Lisa (2018). Can delusions play a protective role? *Phenomenology and the Cognitive Sciences 17 (4):813-833.*

Halligan, P. W., Marshall, J. C., & Wade, D. T. (1993). Three arms: A case study of supernumerary phantom limb after right hemisphere stroke. *Journal of*

*Neurology, Neurosurgery & Psychiatry*, *56*(2), 159–166. https://doi.org/10.1136/jnnp.56.2.159

Hoerl, C. (2001). On Thought Insertion. *Philosophy, Psychiatry, & Psychology*, *8*(2), 189–200. https://doi.org/10.1353/ppp.2001.0011

Kazui, H., Ishii, R., Yoshida, T., Ikezawa, K., Takaya, M., Tokunaga, H., Tanaka, T., & Takeda, M. (2009). Neuroimaging studies in subjects with Charles Bonnet Syndrome. *Psychogeriatrics*, *9*(2), 77–84. https://doi.org/10.1111/j.1479-8301.2009.00288.x

Khalid, M., Hamdy, M., Singh, H., Kumar, K., & Basha, S. A. (2016). Anton Babinski Syndrome—A Rare Complication of Cortical Blindness. *GMJ*, *1*(1), 4.

Langdon, R. (2011). The cognitive neuropsychiatry of delusional belief. *WIREs Cognitive Science*, *2*(5), 449–460. https://doi.org/10.1002/wcs.121

Langdon, R., & Bayne, T. (2010). Delusion and confabulation: Mistakes of perceiving, remembering and believing. *Cognitive Neuropsychiatry*, *15*(1–3), 319–345. https://doi.org/10.1080/13546800903000229

Locke, J. (1690). *An Essay Concerning Human Understanding*. Thomas Bassett.

Lotze, M., Flor, H., Grodd, W., Larbig, W., & Birbaumer, N. (2001). Phantom movements and pain. An fMRI study in upper limb amputees. *Brain: A Journal of Neurology*, *124*(Pt 11), 2268–2277. https://doi.org/10.1093/brain/124.11.2268

Macpherson, F. (2012). Cognitive penetration of colour experience: Rethinking the issue in light of an indirect mechanism. *Philosophy and Phenomenological Research, 24-62.*

Macpherson, F. (2017). The relationship between cognitive penetration and predictive coding. *Consciousness and Cognition*, *47*, 6–16. https://doi.org/10.1016/j.concog.2016.04.001

Marchi, F. (2017). Attention and cognitive penetrability: The epistemic consequences of attention as a form of metacognitive regulation. *Consciousness and Cognition*, *47*, 48–62. https://doi.org/10.1016/j.concog.2016.06.014

Martín Juan, A., Madrigal, R., Porta Etessam, J., Sáenz-Francés San Baldomero, F., & Santos Bueso, E. (2018). Anton–Babinski syndrome, case report. *Archivos de La Sociedad Española de Oftalmología (English Edition)*, *93*(11), 555–557.

Melzack, R. (1990). Phantom limbs and the concept of a neuromatrix. *Trends in Neurosciences*, *13*(3), 88–92. https://doi.org/10.1016/0166-2236(90)90179-e

Miyazono, K., & Bortolotti, L. (2014). The Causal Role Argument Against Doxasticism About Delusions. *Avant: Trends in Interdisciplinary Studies*, *3*, 30–50. https://doi.org/10.26913/50302014.0112.0003

Newen, A., & Vetter, P. (2017). Why cognitive penetration of our perceptual experience is still the most plausible account. *Consciousness and Cognition*, *47*, 26–37. https://doi.org/10.1016/j.concog.2016.09.005

Nuara, A., Nicolini, Y., D'Orio, P., Cardinale, F., Rizzolatti, G., Avanzini, P., Fabbri-Destro, M. & De Marco, D. (2020). Catching the imposter in the brain: the case of Capgras delusion. *Cortex, 131, 295-304.*

Pickard, H. (2010). Schizophrenia and the epistemology of self-knowledge. *European Journal of Analytic Philosophy*, *6*(1), 55–74.

Pronin, E. (2009). Chapter 1 The Introspection Illusion. *Advances in Experimental Social Psychology*, 1–67. https://doi.org/10.1016/S0065-2601(08)00401-2

Redlich, E., & Bonvincini, G. (1911). Weitere klinische und anatomische Mitteilungen fiber das Fehlen der Wahrnehmung der eigenen Blindheit bei Hirnkrankheiten. *Neurol. Centralbl.*, *30*, 227–235.

Saks, E. R. (2007). *The Center Cannot Hold: My Journey Through Madness.* Hachette UK.

Sacks, O. (2012). *Hallucinations* (pagg. xiv, 326). Alfred A. Knopf.

Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. *Philosophical Review*, *117*(2), 245–273. https://doi.org/10.1215/00318108-2007-037

Schwitzgebel, E. 2010. *Introspection,* in the Stanford encyclopedia of philosophy (winter 2019 edition), Edward N. Zalta (ed.), URL =

<https://plato.stanford.edu/archives/win2019/entries/introspection/>.

Smithies, D. (2012). A Simple Theory of Introspection. In D. Smithies & D. Stoljar (A c. Di), *Introspection and Consciousness* (pagg. 259–294). Oxford University Press.

Sollberger, M. (2014). Making Sense of an Endorsement Model of Thought-Insertion. *Mind & Language*, *29*(5), 590–612. https://doi.org/10.1111/mila.12067

Strømgren, E. (1956). *Psykiatri*. Munskgaard.

Sullivan-Bissett, E. (2018). Monothematic delusion: A case of innocence from experience. *Philosophical Psychology*, *31*(6), 920–947. https://doi.org/10.1080/09515089.2018.1468024

Sullivan-Bissett, E. (2020). Unimpaired abduction to alien abduction: Lessons on delusion formation. *Philosophical Psychology*, *33*(5), 679–704. https://doi.org/10.1080/09515089.2020.1765324

## Chapter 4

## When Seeing Is Not Believing: A Mechanistic Basis for Predictive Divergence

This work was published as “Caporuscio, C., Fink, S. B., Sterzer, P., & Martin, J. M.[12] (2022). When seeing is not believing: A mechanistic basis for predictive divergence. *Consciousness and Cognition*, *102*, 103334.”

It was supported by funding from the Deutscher Akademischer Austauschdienst (DAAD), the Berlin School of Mind and Brain, the Deutsche Forschungsgemeinschaft (337619223/RTG2386, STE 1430/8-1, STE 1430/9-1), and the Berlin Institute of Health for a Clinical Fellowship.

**CRediT authorship contribution statement**

Chiara Caporuscio: Conceptualization, Writing – original draft, Writing – review & editing. Sascha Benjamin Fink: Writing – review & editing, Supervision, Visualization. Philipp Sterzer: Writing – review & editing, Supervision. Joshua M. Martin: Conceptualization, Writing – original draft, Writing – review & editing.

### 4.0. Abstract

Visual illusions provide a compelling case for the idea that perception and belief may remain incongruent. This can be explained by modular theories of mind, but it is not straightforwardly accommodated by the Predictive Processing framework, which takes perceptual and cognitive predictions to derive from the same underlying inferential hierarchy. Recent insights concerning the neural implementation of Predictive Processing may help elucidate this. Specifically, prior information is proposed to be approximated by mechanisms in both the top-down and bottom-up streams of information processing. While the former is context-dependent and flexible in updating, the latter is context-independent and difficult to revise. We propose that a stable divergence between perception and belief may emerge when flexible prior information at higher hierarchical levels contradicts inflexible prior information at lower ones. This allows Predictive

[12] Shared first authorship between Caporuscio, C. and Martin, J.M.

Processing to account for conflicting percepts and beliefs while still maintaining a hierarchical and unitary conception of cognition.

### 4.1. Introduction

Perception and belief[13] are both important for agents to successfully navigate the world around them. On the surface, perception appears to be concrete, fast, and concerned with the immediately available environment. Beliefs, in contrast, can be more abstract and are not necessarily as grounded in the here-and-now. This can be illustrated by a simple example. Imagine you are walking past a funny-looking dog on the street. While you can only keep perceiving it for as long as it is close enough to be picked up by your senses, you can keep thinking of it and continue to maintain beliefs about it for hours, days, or (if it was extremely funny) even years after it has disappeared from your perception. Perception and belief also feel phenomenally different: visual perception is iconic in format, while beliefs have been argued to have a sentiential or propositional form (Block 2014, Burge 2010). Additionally, it is not clear whether believing feels like anything while this is hardly controversial when it comes to perceiving (Montague and Bayne 2017). Despite these differences, both belief and perception can be about the same object and, largely, they will be consistent with each other in such cases. If I'm sitting at a bus stop and I *see* a bus approaching, I will most likely also *believe* that the bus is approaching.

However, this is not always the case. Consider this image (Fig. 1). Are the upper and lower surfaces the same color? If you are not familiar with this illusion, your will probably think that they are not. The upper surface will appear considerably darker than the lower surface. However, if you are already familiar with the Cornsweet effect, your answer may be that they are in fact the same. And indeed they are: if you are still unsure, try occluding the middle with your finger to compare the shading. Yet upon learning this fact, your brain does not adjust your perception in line with this newly formed belief or with the incongruent perception you had when occluding the middle line. That is, despite changing

[13] Our use of the term "belief" is inherited from folk psychology and refers to high-level, conscious attitudes. This is different from a lot of the Predictive Processing literature in which the term "belief" is used interchangeably with "prior" or "prior belief", that need not be high level or conscious (see Adams et al. 2013). For terminological clarity, we differentiate beliefs in the folk psychological sense from priors.

your beliefs about the darkness of these patches, your perception remains unchanged. Most of the time, seeing is believing. Visual illusions like the Cornsweet effect provide a compelling exception to this rule.

Here, we aim to provide a mechanistic basis for the persistent incongruence between perceiving and believing in visual illusions from a Predictive Processing perspective. We call this phenomenon *predictive divergence* – cases in which the brain appears to settle upon conflicting predictions of a sensory cause. In section two, we will compare theories defending that the mind is modular, meaning that perception and belief are informationally encapsulated, with Predictive Processing, according to which information flows openly among different hierarchical levels of the cognitive system (Gallagher, Hutto  and Hipólito 2021). While the Modular Mind Hypothesis has a simple explanation for why some illusory percepts remain unchanged despite our updated beliefs about them, this stubborn resistance of information flow between cognitive and perceptual levels poses a challenge for Predictive Processing. We will argue that standard accounts of Predictive Processing, according to which predictive information is approximated by the brain exclusively in a top-down fashion, leave the problem largely unsolved. In section three, we will build on a recent hypothesis put forward by Teufel and Fletcher (2020). They propose that information derived from the adaptation of the neural system to context-independent features of the environment over longer time scales is implemented in the brain as bottom-up constraints on sensory processing. Unlike top-down priors as standardly conceived by the Predictive Processing story, such bottom-up constraints are highly resistant to updating due to short-term changes in context. In section four, we will argue that introducing the notion of constraints offers a solution to the problem of predictive divergence found in cases such as visual illusions.

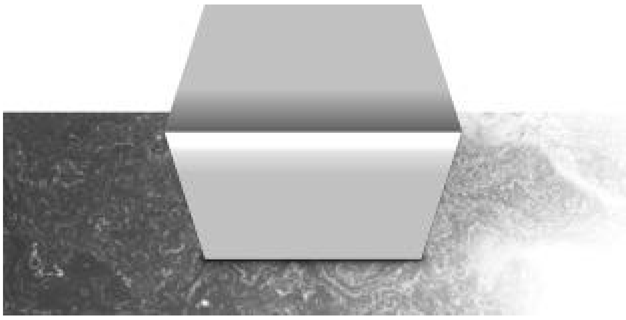

Fig. 1. An illustration of the Cornsweet illusion. People typically perceive the top surface of the central stimulus to be darker than the bottom surface, despite the fact that they are objectively the same.

### 4.2. Predictive divergence: A challenge for predictive processing

#### 4.2.1. Modular and hierarchical minds

A promising explanation for how perception and belief can diverge is found in modular theories of mind. According to the Modular Mind Hypothesis, perceptual systems are informationally encapsulated, meaning that “the data that can bear on the confirmation of perceptual hypotheses includes, in the general case, considerably less than the organism may know” (Fodor, 1983, p. 69). The reason perception and belief diverge is that they do not have access to the same evidence. Because perception needs to be tractable and fast, it simply cannot take all of the vast amounts of information available to one’s mind into consideration. Belief, on the other hand, is not as time-sensitive, and therefore it can draw on a considerably larger amount of evidence, evaluating the input it receives from visual perception together with prior knowledge or information coming from other senses or from theoretical or abstract reasoning. The picture of the brain resulting from these theories is that of a multitude of modular systems responsible for sensory processing and a modality independent belief system. The former captures quick and immediate information and passes it to the latter, which draws on different sources of evidence to create a more integrated picture of the external world and guide decision-making and planning accordingly.

In recent years, however, an alternative conception for the architecture of the mind has emerged – the idea that the mind is hierarchical (Williams 2019). The most famous example of this approach is the Predictive Processing framework (henceforth PP; Clark 2016, Friston 2005, Hohwy 2013, Wiese and Metzinger 2017). According to PP, the core function of the brain is prediction error minimization – to reduce the discrepancy between what a system predicts and incoming sensory input from the body and the external environment. In this article, we describe PP in terms of "hierarchical predictive coding" (see Huang and Rao 2011, Spratling 2017 for reviews), which is a more specific proposal concerning how prediction error minimization may be implemented. According to this perspective, the brain can be seen as entertaining a hierarchical generative model of the world, which formulates predictions at multiple levels of abstraction and spatio-temporal scale. Levels at the bottom of the hierarchy capture relatively concrete and fast-changing features of the environment, while higher-layers encode ever-more abstract and temporally extended properties. Within this inferential hierarchy, descending top-down signals are thought to carry model-based predictions concerning the system's current best explanation of incoming sensory input, while ascending bottom-up signals carry discrepancies between these predictions and sensory input ('prediction errors'). When predictions do not adequately capture sensory causes in the world, unexpected sensory input generates prediction errors, which propagate up the hierarchy and act as a learning signal to revise predictions at higher levels. This prediction error may be minimized by changing one's predictions in order to better accommodate the unexplained sensory input[14]. In this process, different hierarchical levels will interact with one another via iterative message passing allowing the system "to settle into a mutually consistent whole, in which each 'hypothesis' is used to help tune the rest" (Clark 2013, p. 183). The overall prediction on which the system finally settles will be that which is estimated to minimize global prediction error at multiple levels of spatiotemporal scale.

From a PP perspective, all mental occurrences (including perception and belief) are thought to derive from an approximate form of Bayesian inference, where

[14] In this paper, we focus on perception. However, according to the Active Inference proposal, prediction error may also be minimized through action – directly changing sensory input to fulfill one's predictions (see Parr, Pezzulo, and Friston 2022). From this perspective, perception and action are optimized towards the common objective/goal of minimizing prediction error, although they achieve this via distinct strategies in their direction of fit from mind to world.

sensory evidence ('likelihood') and background knowledge ('prior') are combined. What determines our perception or beliefs is the outcome of this process – the prediction which minimizes global prediction error or, in Bayesian terms, the overall hypothesis that optimally combines the likelihood and the prior (carrying the highest posterior probability, Hohwy 2013). Importantly, in order to minimize statistical overfitting, the brain will take into account the expected precision of an information source (mathematically, its inverse variance) when formulating a prediction. This is achieved by assigning a precision-weighting to the priors and prediction errors at each hierarchical level, based upon their estimated reliability or uncertainty in a given context. For example, prediction errors deriving from imprecise sensory evidence (e.g. visual signals at night) will be assigned less weight in updating predictions and may be more likely to be explained away as noise by the system; conversely, prediction errors that are expected to be precise (e.g. visual signals in broad daylight) are weighted more highly and have a greater influence on revising predictions further up the predictive hierarchy. In a similar way, predictions that are based upon priors that are estimated to be precise are less likely to be updated in the face of conflicting prediction errors and vice versa, due to their estimated reliability for efficiently explaining away sensory input in the past.

Unlike modular accounts, the hierarchical and unitary conception of cognition proposed by PP implies that perception and belief are not independent of each other, and that there is no clear-cut boundary between the two. It depicts perceptual and cognitive states as unified in terms of their goals and their information-processing architecture, and only distinct in terms of their spatio-temporal scale (Badcock et al. 2019, Friston and Buzsáki 2016, Kiebel et al. 2008) and the level of abstraction (Williams 2019, Williams 2020). Percepts and beliefs can thus be said to differ along a continuum, where mental states deriving from predictions at lower levels of the hierarchy are 'more percept-like', while predictive states deriving from levels higher up the hierarchy are 'more belief-like' (Clark 2016, Lupyan 2015, Macpherson 2017). One may then assume that the very highest levels reflect higher-order doxastic states, while the very bottom layers reflect detailed percepts (Macpherson 2017). From this perspective, percepts and beliefs are both products of inferences based on posterior estimates of worldly causes that derive from the same underlying predictive hierarchy, where a relatively open exchange of information occurs between different hierarchical levels in order to perform the same unifying function of minimizing global prediction error.

If there are no clear boundaries between higher and lower levels of the hierarchy, does it mean that all information is available to all levels in a "free for all" fashion? Hohwy 2013, Drayson 2017 argue that it is not. While Hohwy grounds this claim in the epistemic advantages of keeping different sources of evidence segregated, Drayson argues that it follows from a purely practical limit of Bayesian computation. In probabilistic causal networks, the causal influence between states is weak: this means that causal influence between any two levels is not transitive, and thus it won't be preserved over long causal chains. This is analogous to how meteorological models work: yesterday's weather influences today's weather, but this does not mean that it will influence weather in a hundred years in any discernible way (Spohn 2008). The further apart two levels in the hierarchy are, the less likely it is that there is any discernible causal influence between them. This means that priors at higher levels will more strongly inform inferences at higher levels, and priors at lower levels will more strongly inform inferences at lower levels. For example, a prior encoded at a high level will inform inferences at that level and at the levels immediately below, but its discernible influence will slowly cease with each intermediate step until it reaches levels much lower down in the hierarchy. Instead of a "free for all" exchange of information between hierarchical levels, we can rather think of a slow trickling of information across the inferential hierarchy, an open but restricted flow of information between levels.

### 4.2.2. Predictive processing and visual illusions

With this framework in mind, let us now turn back to visual illusions. From a PP standpoint, there are two things that call for an explanation in visual illusions like the Cornsweet effect. The first one is the divergence between reality and experience: why do we perceive the upper surface as darker than the lower one if they are actually the same? The second one is the apparent persistent divergence between perception and belief: why do we still perceive the upper surface as darker than the lower surface, even though we have now formed the belief that they are the same?

Illusory percepts have been explained in PP as arising from sensory input that contradicts highly precise priors of our perceptual system. For example, the Cornsweet illusion is thought to arise from precise prior predictive information encoded in the visual system, reflecting the assumption that light comes from above (Lupyan  and Clark 2015). In this sense, visual illusions reflect "optimal percepts" (Clark 2018) or "globally optimal solutions" (Lupyan 2015): if a prior is attributed a high degree of precision, dismissing conflicting sensory evidence as noise is the best strategy to adaptively infer causes in the world. In order to minimize global prediction error on average, our brain will not always employ the best world models for every instance of sensory data, but the most generalizable ones—the ones that will apply to most of the cases we are likely to encounter in real life. "Given the structure and statistics of the world we inhabit, the optimal estimate of the worldly state [...] will be the one that, on some occasions, gets things wrong" (Clark 2016, p. 51).

But what about the second explanandum, namely the *stability* of this divergence between perception and belief? In the literature, the stability of visual illusions has been mainly discussed in terms of cognitive impenetrability: why higher-level states, such as the newly formed belief that the upper and the lower surface are of the same color, fail to influence perception. For example, according to Hohwy, whether higher-order beliefs penetrate perception will depend on how early in the hierarchy prediction error is sufficiently suppressed (Hohwy 2013, pp. 124–128).

However, this conception leaves something unexplained. If precise priors restrict our perceptual predictions, then why do they not equally restrict our doxastic predictions (i.e. beliefs)? Given that perception and belief are both derived from the same underlying inferential hierarchy where a relatively open exchange of information occurs between different hierarchical levels, we should expect that predictive information that inflexibly determines one level should also inflexibly determine the other. Instead, when it comes to visual illusions, different levels of the generative model seem to settle on divergent predictions, as high-level beliefs conflict with low-level percepts. This seemingly contradicts a principle of prediction error minimization through hierarchical message passing and the notion that predictions should interact with one another in order "to settle upon a mutually consistent whole" (Clark 2013, p. 183). A similar worry has been raised by Gallagher et al. (2021, p. 9) in exploring how PP can account for the cognitive impenetrability of perceptual illusions: "Why don't perception and cognition engage in effective exchange in order to minimize error, rather than preserving the sensory error and allowing it to rule?". The challenge from a PP perspective is to explain how such predictive divergences may occur and persist in visual illusions, while still maintaining its core assumptions regarding the brain's underlying predictive architecture.

In a recent article, Gallagher et al. (2021) attempt to provide a 4E (embodied, embedded, extended, enactive) solution to this problem. They claim that an internalist perspective of PP fails to take into account the limitations that are imposed by the factors involved in the broader cognitive environment:

"There is a kind of structural resistance introduced by experimental factors, or more generally, ecological factors. In that case, however, it's not brain architecture, or a conditional independence of different sensory systems, or a conceptually implausible model, or the ignoring of prediction error that is doing the work; it's the environment […] and the non-ecological circumstances of the experimental situation that place constraints on the system and prevent the agent from taking action." (p. 14)

From this perspective, a PP account is not at odds with visual illusions when we consider the involvement of broader influences from the body and the environment. For example, participants can adapt relatively quickly to some visual illusions once they are allowed to interact with them (Gallagher et al.

2021). These adaptations are thought to depend strongly on the involvement of bodily and environmental factors (e.g. Cesanek  and Domini 2017).

In the next section, we present a brain-based, but potentially compatible alternative explanation. Unlike Gallagher et al. (2021) who appeal to the structural limitations imposed by broader ecological factors, we focus our attention on the structural limitations imposed by mechanisms in neural architecture. To do this, we draw upon a recent proposal by Teufel and Fletcher (2020) who argue that the implementation of prior predictive information in the brain is differentially structured according to the type of regularities in the environment it is optimized to capture. This results in (at least) two distinct ways that a 'prior' may be approximated at a neural level: there are top-down hierarchical mechanisms that are optimized to capture context-dependent regularities in the world; and there are also inflexible bottom-up mechanisms that are optimized to capture context-independent regularities in the world. Our suggestion is that the dissociation of percepts and beliefs, such as in visual illusions, may be due to conflicting sources of prior information that are hardwired in these respective top-down and bottom-up streams of information processing. Because the neural mechanisms underlying context-independent prior information are hard-wired into the system, they are relatively resistant to short-term changes in the activity of the system. This can lead to predictive divergence in cases where the brain's implicit context-independent assumptions about the world (e.g. for vision that light comes from above) are violated, such as in the case of some visual illusions.

### 4.3. Different types of predictive information: Expectations and constraints

#### 4.3.1. Prior information in the bottom-up stream

In the Bayesian literature, the term 'prior' describes the assumed probability distribution of possible causes of sensory data *before* current evidence is assessed. Importantly, this definition is agnostic as to the mechanistic implementation of this predictive information (Körding 2007): there may be different ways in which a prior may be neurally implemented, while still achieving the same computational goal.

Despite this, the PP literature often assumes that predictive information in the brain is exclusively implemented in the top-down stream, where “the feedback from higher inference layers provides the priors to shape inferences at earlier levels” (Lee  and Mumford 2003, p. 9). Following Teufel and Fletcher (2020, p. 235), we will refer to priors exerting a top-down influence on perception as *expectations*: “higher-level processes extract contextual information, derive predictions and feed them back to modulate earlier aspects of perceptual processing”. Traditional accounts of PP have assumed that priors are always implemented in a top-down fashion as expectations, while all that is carried by the bottom-up stream are prediction errors – the discrepancy between top-down predictions and sensory input.

However, recent developments suggest that this might not be the whole story. According to Teufel and Fletcher (2020), predictive information may also be approximated in the bottom-up stream in the form of *constraints*. A constraint is a property of the nervous system (or part of it) that imposes a restriction on bottom-up information processing. For example, there is an overrepresentation of neurons in the primary visual cortex that are tuned to vertical and horizontal orientations, which is thought to underlie a perceptual preference for lines at cardinal orientations (0°/90°) over oblique orientations (45°/135°) (Furmanski and Engel 2000, Li et al. 2003). This perceptual bias is thought to reflect a structural adaptation to natural scene statistics, where lines at cardinal orientations are similarly overrepresented (Girshick, Landy  and Simoncelli 2011). From the definition presented previously, the overrepresentation of neurons tuned to the cardinal axes can be interpreted as the hard-wired basis of a prior, since it reflects a deeply ingrained hypothesis about the agent's environment, one that shapes sensory evidence and exists before any evaluation of current sensory evidence takes place.

Teufel and Fletcher’s proposal draws upon early thinking in cybernetics, which claimed that for an agent to successfully control the impact of its environment, there should be a structural mirroring between regularities in the environment and the controllers or regulators of such features in the agent (Conant  and Ashby, 1991). Context-dependent regularities in the environment (i.e. relevant for only specific encounters between the agent and her environment) will be regulated by context-dependent features of the agent, while relatively context-independent regularities (i.e. relevant for every encounter between an agent and her environment) will be regulated by relatively context-independent features of the agent.[15] A helpful analogy provided by Teufel and Fletcher compares the agent to a steersman who wishes to control for different types of regularities while out at sea: A good steersman needs to continuously change flexible parameters of his boat, like the mast and sails, to deal with changing tides, wind and currents; but a boat must also possess stable and unchanging features, like its material and shape, to deal with the constant, unvarying properties of the sea. For this reason, we should expect the agent to have "unchanging features, which regulate the unchanging influences of its world but also context-dependent features, which mirror and regulate the context-dependent features of its world." (Teufel and Fletcher 2020, p. 232). The latter are expectations, flexibly implemented in the top-down stream, while the first are constraints, relatively inflexibly implemented in the bottom-up stream.[16]

### 4.3.2. Constraints are asymmetrically implemented

We believe that the extent to which predictive information captures context-independent regularities can be expected to vary along the cortical hierarchy: constraints are more likely to be embedded at lower levels as opposed to higher levels. While this claim was not explicitly stated by Teufel and Fletcher (2020), the respective examples they highlight are suggestive of this trend. While

[15] While this notion of structural mirroring was proposed by Teufel and Fletcher (2020) in relation spatiotemporal regularities in the environment (i.e. what is probable), a similar structural correspondence may apply to environmental regularities in terms of their biological importance or estimated 'utility' in regulating physiological needs (see Martin, Solms and Sterzer 2021).

[16] At the same time, not all inflexible features of the boat are reflecting features of the environment: there may be shapes or materials of the boat that would work even better than those typically used by the shipbuilders of a given time and age, but they may just be impossible to realize, given the limitations of materials and craftsmanship available to those shipbuilders. Similarly, it should be noted that not all structural constraints of the bottom-up stream are representing information regarding the probabilistic features of the environment: some are simply imposed by what is biologically or physically possible or not. For example, an upper bound on processing the speed of visual motion might exist just because it is not physically possible for any biological system to parse objects moving at the speed of light. Constraints of this kind should not be taken to be the neural basis of priors.

they indicate that constraints may be involved in higher-order aspects of learning and cognition, the rest of the evidence for constraints derives from predictive information that is structurally embedded in fairly low levels of the visual system.[17]

At first glance, the claim that constraints are more likely to be embedded at lower hierarchical levels may seem paradoxical: predictions formed at higher levels are often contrasted with lower levels in terms of their relative invariance to changes in sensory input (Kiebel et al. 2008, Rossi-Pool et al. 2021), and constraints are proposed to regulate relatively unchanging features of the environment. However, this may be misleading, since the notion of 'variance' will depend upon the timescale and information source in question. While higher-levels may be less variant with respect to a specific source of sensory input and over shorter periods, they can be described as relatively more variant if we consider how their activity is modulated over longer periods by wider changes occurring elsewhere in the system. For example, lower levels such as V1 may be highly variant to short-term changes in visual information, but relatively shielded off activity occurring in other sensory modalities or changes in information over longer timescales. This is in contrast to higher hierarchical levels, such as the midline default mode network core, which are "well-suited to processing transmodal information unrelated to immediate sensory input" (Stawarczyk, Bezdek  and Zacks 2021, p. 168). If we conceptualize context-dependency in terms of how the brain integrates and modulates activity according to surrounding factors, then one may expect a functional gradient to exist whereby higher hierarchical levels are likely to reflect increasingly context-dependent features of the environment due to their integrative nature and far-reaching connections. If Teufel and Fletcher's account is on track, then this functional gradient should be mirrored by underlying differences in mechanism, whereby constraints are more likely to be implemented at lower hierarchical levels.

[17] We admit that this may be, at least in part, due to the fact that it becomes increasingly unclear what would constitute a clear example of a constraint at a higher-cognitive level. For example, cognitive biases and heuristics can be compared to visual illusions, since they reflect cognitive inferences that are automatically deployed in a relatively context-independent manner (Kahneman 2011). However, unlike visual illusions, they seem much more flexible to revision. The ambiguities in such cases may also reflect why expectations and constraints are proposed to signify two end points of a continuum, as many neural approximations of priors may lie somewhere in between.

A second reason for endorsing an asymmetry in the distribution of constraints concerns the relative computational benefits they confer. By taking a given parameter as fixed (e.g. that light comes from above), constraints avoid the need to constantly estimate parameters in each and every predictive context. While this may sacrifice predictive flexibility and lead to some misguided assumptions for statistical outliers, it will also minimize the metabolic resources and processing speed involved in updating predictions, due to fewer processing steps being needed, each of which will burn valuable energy. Such improvements in processing speed are arguably more beneficial for lower levels of the processing hierarchy where predictions concerning fast-changing regularities need to be sufficiently rapid. In this way, errors deriving from mechanistic inflexibility at lower levels may be more tolerable due to their ability to reduce the computational burden in perceptual performance.

### 4.3.3. Constraints are relatively resistant to updating

One of the central tenets of PP states that when our model-based predictions do not adequately explain sensory causes, a prediction error results in an updating of our prior models. This is unproblematic in the case of expectations, which are relatively flexible due to their ability to change according to contextual demands. However, one of the core features of constraints is their relative immutability or resistance to revision: There will be inherent restrictions on the extent to which implicit priors, such as the overrepresentation of neurons tuned to the cardinal axes in V1, may be updated in relation to experience. This raises a potential worry: if constraints are limited in their capacity to update, to what extent can they be considered to act as 'priors' in cognitive inference and in what way can we say that the brain is still operating under a 'Bayes-optimal' principle? A related line of argument is put forth by Orlandi (2018)[18]: "The thought seems to be that perception could be Bayesian simply in virtue of percepts being derived as a function of the posterior probability of a perceptual hypothesis, with no need to think that the perceptual system also updates in accordance with Bayes' rule" (p. 2378). The worry then is that by giving up this notion, the PP story on offer

[18] Although Orlandi (2018) mentions this point in relation to the problem of "backwards blocking", the same principle may apply here since both cases are concerned with how prior probabilities are flexibly updated in light of new evidence.

"would give up a distinctive feature of Bayesianism" (p. 2378). From this perspective, rather than classifying constraints as acting as prior inputs to cognitive inference, they may be perhaps more parsimoniously described as biases that are embedded in the perceptual apparatus of the cognitive system (Orlandi 2016).

While a thorough discussion of this point is beyond the scope of the present account, there are at least three points that may help alleviate this worry. Firstly, the claim of PP is not that the brain is literally implementing Bayesian inference, but rather that it may be an approximation of some form (Clark 2016, Jacobs and Kruschke 2011, Rescorla 2015, Zednik  and Jäkel 2016). The model where the posterior is the new prior may be a simplified or idealized version of Bayesian inference that works for some standard learning scenarios, but that covers only part of what is going on in reality. Just because the system will update its world model in light of new evidence, does not necessarily imply that all contributing priors will be equally changed according to the new posterior. Indeed, it is probably often the case that various sources of information need to be integrated into a common posterior, as it happens in multisensory integration (Knill and Pouget 2004, Ernst and Banks 2002). The neural representation of this posterior will be different from the neural representation of all the contributing priors, meaning that not all the contributing priors will be updated equally. That does not mean that perception is not Bayesian; it just means that not all the neural processes (or structures) that give rise to priors are necessarily updated according to the posterior. For a similar reason, the fact that some neural properties can effectively act as priors in perceptual inference does not necessarily imply that this neural property will be changed according to the new posterior. Constraints may not be updated as easily as expectations because (1) they are hard-wired and (2) their neural implementation is different from the representation of the posterior. Thus, constraints may still effectively act as priors, but may not be automatically updated in the way we would expect given an unconstrained application of the Bayes rule.

Secondly, even if there are inherent limitations on how constraints may be updated, this does not mean that the system will be unable to learn or adapt to conflicting evidence when it arises. Rather, the system might more easily update context-dependent, short-term expectations than the more rigid constraints (Teufel and Fletcher 2020, p. 238). In other words, prediction errors may update the brain's prior models but this will be more likely to result in changes to expectations, while constraints remain relatively fixed. This is nicely illustrated by the example of haptic training, intended to reverse the light-from-above prior, which is thought to be neurally implemented in the form of a constraint (Teufel and Fletcher 2020). While haptic training may temporarily shift predictions specific to the laboratory context (Adams, Graf  and Ernst 2004), predictions external to that specific environment will keep working under the assumption that light comes from above (Adams, Kerrigan  and Graf 2010). This suggests that the training produced a new context-dependent expectation for a change in light source that differentially affected the newly formed prediction, while the underlying constraint remained unchanged. Just because the constraint was not updated does not mean that the system was static or unable to learn; rather, this change in prior information was reflected by mechanistic changes elsewhere in the system.

Lastly, the claim that constraints are more rigid than expectations should not be taken to imply that constraints are completely immutable. Rather, we can think of constraints at lower levels of the system as being *relatively* static, in the sense that they are highly resistant to revision. It should be noted that this is not unique to predictive information that is approximated by the bottom-up stream: some prior information implemented as expectations might also be highly resistant to revision, like highly precise ones derived from learnt traumatic experiences (Lyndon  and Corlett 2020). This relative rigidity does not mean that prior information is impossible to update, but rather that it may require reliable conflicting evidence over extended periods. Some constraints could in principle be modified by experience over longer time scales during adulthood (Teufel  and Fletcher 2020) – for example, by constant exposure for years to a different environment, like a planet where light comes from below or where faces are concave. Some others might be updated in sensitive periods of the development of the nervous system. Or they could be updatable over even longer time scales, not by the individuals but by the species, through evolution and natural selection. In this last case, the information implicitly carried by constraints would not qualify as a prior in an internalist sense, as it would not be a hypothesis formed and updated by the individual through prior experience (Hohwy 2016); however, it could be considered as a prior in a broader sense by externalist accounts that argue that extra-cranial factors, such as the body, the environment and evolution, are integral parts of the PP story (Clark 2016, Gallagher et al. 2021). Detailed discussion on this point goes beyond the scope of this paper, but we believe that the notion of evolved priors could be very relevant to the discussion between internalism and externalism.

### 4.4. Accounting for divergence

#### 4.4.1. How does predictive divergence happen?

Teufel and Fletcher (2020) stress that the predictive information involved in some cases of visual illusions might be best explained by constraints in the bottom-up stream. For example, the light-from-above prior that underlies a number of common visual effects, including the Cornsweet illusion, appears to be implemented in the form of a context-independent constraint on visual perception, embedded in early subcortical or even retinal processes (Anderson et al. 2009, Dakin and Bex 2003, Ramachandran, 1988). This can be explained by the fact that we inhabit an environment where light comes from above, which generalizes in a context-independent fashion and is therefore inflexibly implemented as a constraint in the feedforward stream of information processing. We believe that the distinction between expectations and constraints might shed light on the question of how PP can account for the stable divergence between percepts and beliefs in cases like the Cornsweet illusion. The aim of this section is to bring the puzzle pieces together.

As we previously argued, constraints are more likely to be implemented at lower levels, as opposed to higher levels of the cortical hierarchy. From the general PP account, it is proposed that beliefs arise from inferences at higher levels, while perceptual states arise from inferences at lower levels (Macpherson 2017). Drayson (2017) adds that updating at every level is more likely to be informed by priors at the same level or immediately close-by levels. This leads to the conclusion that, relative to beliefs, percepts are more likely to be influenced by constraints, because these are more likely to be found at lower levels of the predictive hierarchy.

Given that higher and lower hierarchical levels are likely to be influenced by different priors to varying degrees, it seems plausible that predictions made at different levels may settle upon different posterior estimates when they are driven by conflicting priors. Therefore, perception and belief may diverge under conditions where expectations and constraints at different levels are conflicting. But is it possible for expectations and constraints to conflict with one another? Yes, given the peculiar way that expectations and constraints interact, we should not always expect predictive information embedded in constraints and expectations to be consistent and aligned (Teufel and Fletcher 2020). Expectations are context-dependent while constraints, in contrast, are context-independent. Thus, a change in context that provides the system with new context-dependent information may lead to a change in expectations that is incongruent with an underlying constraint that is relatively resistant to updating. This should be most likely to occur when the system encounters reliable information that violates a context-independent assumption about the world. From this perspective, visual illusions like the Cornsweet effect elicit a mental state consisting of divergent predictions: a constraint-driven percept and an expectation-driven belief. Given that constraints are more rigid and resistant to revision than expectations, our perception will remain unchanged even after reliable contextual information causes us to update our beliefs.

We say predictions 'may' diverge under these conditions because conflicting constraints and expectations do not always lead to an incongruence between perception and belief. Consider, for example, the case of the light-from-above prior that we discussed before. While the effect is proposed to be driven by a constraint, haptic training may still change one’s perceptual prediction away from an assumption of an overhead light source (Adams et al. 2004). These and other similar changes in perceptual prediction are specific to the laboratory context (Adams et al. 2010), suggesting that such training formed a new context-specific expectation as opposed to revising the original constraint (Teufel  and Fletcher 2020). If this is true, then despite conflicting priors, the resulting percept was congruent with the newly formed expectation. But how does this example differ from the visual illusions discussed before? Possible reasons may involve the relative precision ascribed to the priors and prediction errors in question (Hohwy 2013), as well as the length of the probabilistic causal chain between the two levels (Drayson 2017). To this end, the tactile feedback in haptic training may be effective in inducing a more precise expectation at a relatively low perceptual level. If this is on track, then divergent predictions should be more likely to occur when lower-level constraints interact with higher-order expectations at far away ends of the cortical hierarchy.

### 4.4.2. Why does predictive divergence happen?

We have offered an account of how divergent percepts and beliefs may be implemented in a manner consistent with PP. If the picture we have sketched is on the right track, predictions at lower and higher levels of the hierarchy are sensitive to different types of information: specifically, constraints are more likely to be involved in perception than in belief formation. Because they are not informed by the same evidence, perception and belief can come to different conclusions about the hidden wordly causes of sensory input. Other accounts have suggested that a partial segregation of information is present in the predictive architecture, and have argued that this follows from the epistemic advantages of recruiting independent sources of evidence (Hohwy 2013) or from inevitable information loss across long chains of probabilistic inference (Drayson 2017). It might be questioned whether the distinction between expectations and constraints adds anything to these accounts. We believe that our explanation is not incompatible with previous accounts, but it might offer additional adaptive reasons why it is desirable to sometimes settle on different predictions at perceptual and cognitive levels.

The asymmetrical implementation of predictive information in the form of constraints at different levels of the hierarchy reflects a compromise between conservation of energy and predictive flexibility. Inflexible bottom-up constraints rigidly implemented in the visual cortex allow us to have mostly optimal percepts with a relatively little metabolic cost: having fixed parameters that we don't need to estimate is very computationally inexpensive, the same way relying on in-built features of the boat requires no additional effort from the steersman while at sea (Teufel and Fletcher 2020). This is extremely useful for perception, which requires the processing of sensory input to be fast and immediate. However, relying too heavily on constraints means running the risk of statistical underfitting whenever these fixed estimates are contradicted by other sources of evidence. For example, in most cases, we can't help but perceive a light source as coming from above, despite convincing alternative evidence suggesting otherwise.

Instead, top-down expectations require a higher computational effort but allow the integration of other sources of information that take context-dependent information into account. Cognition is slower than perception, so it can tolerate the computational burden of estimating new parameters via the integration of contextual sources of information and counterbalance the statistical underfitting of predictions at lower levels. It is important to remember that we may find constraints and expectations at different levels; but the asymmetrical distribution of constraints at different levels of the hierarchy allows us to prioritize computational efficiency at lower levels and contextual integration at higher ones.

### 4.5. Concluding remarks and future directions

If our account is on the right track, there is an explanation for the persistent divergence between perception and belief in visual illusions that can maintain some of PP's core assumptions, such as the hierarchical structure of perception and cognition, the idea that the brain approximates Bayesian inference and the lack of clear boundaries between perception and belief. However, some elements of the standard PP story are weakened, like the commitment to unrestricted communication between hierarchical levels, while others do not survive, like the assumption that the bottom-up stream only conveys model-neutral sensory data.

Furthermore, some constraints that implicitly represent predictive information might be updatable by evolution and natural selection over longer time scales than an individual's life. Therefore, a commitment to interpreting all predictive information in the bottom-up stream implicitly representing features of the environment as Bayesian priors might force us to abandon internalist and strictly brain-based versions of PP in favor of externalist accounts that include broader ecological factors.

Several implications for future research follow from our account. We should expect that many visual illusions characterized by predictive divergence may be traced back to conflicting expectations and constraints. Specifically, our account predicts that percepts in such cases will be driven by constraints embedded at fairly lower levels, reflecting context-independent assumptions about the world (e.g. that light comes from above). Second, we would expect that illusory perception in these cases is more likely to align with beliefs under experimental conditions similar to the haptic training example, where a highly precise expectation is induced at a relatively low level in the cortical hierarchy. Finally, it would be interesting to explore whether machine learning using hierarchical multilevel architectures can simulate forms of predictive divergence through imposing constraints at fairly lower levels, or whether mechanisms suggestive of constraints are more likely to arise at lower hierarchical levels when systems are allowed to implement prior information in a non-hierarchical fashion in order to gain computational benefits.

Our account may also be useful in characterizing distinct forms of misguided inference and their possible causes in psychopathology. As we have stated in the previous section, predictive divergence ultimately serves the purpose of mediating between the need for inflexibility at lower levels and the demand for flexibility at higher levels. When this balance is lost, dysfunctional inferences might arise. Too much flexibility at lower levels might make perception too penetrable by higher-level influences, causing hallucinations (Corlett et al. 2019). Too much inflexibility at higher-levels might cause beliefs to be too resistant to revision when new contextual information is acquired, thus shedding light onto the persistence of delusions despite contradictory evidence (Heinz et al. 2019, Schmack et al. 2013, Sterzer et al. 2018). We expect that more research into how expectations and constraints underlie dysfunctional inference may provide new insights into mental illness and neuropathology.

There are likely to be be other cases in which the consistency of perception and belief is violated: for example, people often incorrectly perceive patterns or objects in noisy stimuli ("pareidolia"), such as seeing faces in clouds, without believing that they are actually there. Similarly, a regular user of psychedelics may not assume that the hallucinations they experience under the influence of the drug correspond to reality. More trivially, we do not think that the room has gone dark every time we close our eyes. Our account does not necessarily apply to all these cases: there is evidence that pareidolia at least might be driven by a predominantly top-down circuit (Liu et al. 2014), which would indicate a different type of mechanism. An alternative explanation for some cases of predictive divergence might be found in the updating of a high-level expectation, such as the prior that predicts that perception is tracking system-external reality. Limanowski and Friston (2018) treat such "transparency" (see Metzinger 2014) as associated with precision estimation for active inference. If such a "transparency prior" is weakened, the perceptual signal will be treated as being too noisy to force an update on the associated belief. However, the noise is systematic and stable. This systematicity and stability is a signal. The system could predict it by a dissociation: the system dissociates the content of belief as providing information about external causes (made probably by pooling over previous knowledge as well as a range of percepts) while the content of the percept provides information about how internal causes (i.e. idiosyncrasies of one's nervous system) skew information about external causes. Either is valuable information to the system for action where it is beneficial to keep both percept and belief stable.

### 4.6. References

Adams, Stephan, K. E., Brown, H. R., Frith, C. D., and Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, *4*. https://doi.org/10.3389/fpsyt.2013.00047

Adams, W. J., Graf, E. W., and Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, *7*(10), 1057–1058. https://doi.org/10.1038/nn1312

Adams, W. J., Kerrigan, I. S., & Graf, E. W. (2010). Efficient visual recalibration from either visual or haptic feedback: the importance of being wrong. *Journal of Neuroscience*, *30*(44), 14745–14749. https://doi.org/10.1523/JNEUROSCI.2749-10.2010

Anderson, E. J., Dakin, S. C., & Rees, G. (2009). Monocular signals in human lateral geniculate nucleus reflect the Craik–Cornsweet–O'Brien effect. *Journal of Vision*, *9*(12), 14–14. https://doi.org/10.1167/9.12.14

Badcock, P. B., Friston, K. J., & Ramstead, M. J. D. (2019). The hierarchically mechanistic mind: A free-energy formulation of the human psyche. *Physics of Life Reviews*, *31*, 104–121. https://doi.org/10.1016/j.plrev.2018.10.002

Block, N. (2014). Seeing-as in the light of vision science. *Philosophy and Phenomenological Research*, *89*(3), 560–572. https://doi.org/10.1111/phpr.12135

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

Cesanek, E., & Domini, F. (2017). Error correction and spatial generalization in human grasp control. Neuropsychologia, 106, 112-122. https://doi.org/10.1016/j.neuropsychologia.2017.09.026

Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, *17*(3), 521–534. https://doi.org/10.1007/s11097-017-9525-z

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181-204. https://doi.org/10.1017/S0140525X12000477

Conant, R. C., & Ashby, W. R. (1991). Every good regulator of a system must be a model of that system. In G. J. Klir, *Facets of Systems Science* (pp. 511–519). Springer US. https://doi.org/10.1007/978-1-4899-0718-9_37

Corlett, P. R., Horga, G., Fletcher, P. C., Alderson-Day, B., Schmack, K., & Powers, A. R. (2019). Hallucinations and strong priors. *Trends in Cognitive Sciences* 2*3*(2), 114–127. https://doi.org/10.1016/j.tics.2018.12.001

Dakin, S. C., & Bex, P. J. (2003). Natural image statistics mediate brightness 'filling in'. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 2*70*(1531), 2341–2348. https://doi.org/10.1098/rspb.2003.2528

Drayson, Z. (2017). Modularity and the predictive mind. In T. Metzinger & W. Wiese (Eds.) *Philosophy and predictive processing*. MIND Group.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(*6870*), 429-433. https://doi.org/10.1038/415429a

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

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K., & Buzsáki, G. (2016). The functional anatomy of time: What and when in the brain. *Trends in Cognitive Sciences*, *20*(7), 500–511. https://doi.org/10.1016/j.tics.2016.05.001

Furmanski, C. S., & Engel, S. A. (2000). An oblique effect in human primary visual cortex. *Nature Neuroscience*, *3*(6), 535–536. https://doi.org/10.1038/75702

Gallagher, S., Hutto, D., & Hipólito, I. (2021). Predictive processing and some disillusions about illusions. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00588-9

Girshick, A., Landy, M. & Simoncelli, E. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*, 926–932. https://doi.org/10.1038/nn.2831

Heinz, A., Murray, G. K., Schlagenhauf, F., Sterzer, P., Grace, A. A., & Waltz, J. A. (2019). Towards a unifying cognitive, neurophysiological, and computational neuroscience account of schizophrenia. *Schizophrenia Bulletin*, *45*(5), 1092–1100. https://doi.org/10.1093/schbul/sby154

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Hohwy, J. (2016). The self-evidencing brain. *Nous, 50(2)*, 259–285. https://doi.org/10.1111/nous.12062

Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580-593. https://doi.org/10.1002/wcs.142

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 8-21.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kiebel, S. J., Daunizeau, J., & Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLOS Computational Biology*, *4*(11), e1000209. https://doi.org/10.1371/journal.pcbi.1000209

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712-719. https://doi.org/10.1016/j.tins.2004.10.007

Körding, K. (2007). Decision theory: What "should" the nervous system do? *Science*, *318*(5850), 606–610. https://doi.org/10.1126/science.1142998

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *JOSA A*, *20*(7), 1434–1448. https://doi.org/10.1364/JOSAA.20.001434

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique effect: a neural basis in the visual cortex. *Journal of neurophysiology*, *90*(1), 204–217. https://doi.org/10.1152/jn.00954.2002

Liu, J., Li, J., Feng, L., Li, L., Tian, J., & Lee, K. (2014). Seeing Jesus in toast: neural and behavioral correlates of face pareidolia. *Cortex*, 53, 60-77. https://doi.org/10.1016/j.cortex.2014.01.013

Limanowski, J., & Friston, K. (2018). 'Seeing the dark': grounding phenomenal transparency and opacity in precision estimation for active inference. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00643

Lupyan, G. (2015). Cognitive penetrability of perception in the age of prediction: predictive systems are penetrable systems. *Review of Philosophy and Psychology*, *6*(4), 547–569. https://doi.org/10.1007/s13164-015-0253-4

Lupyan, G., & Clark, A. (2015). Words and the world: predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*, *24*(4), 279–284. https://doi.org/10.1177/0963721415570732

Lyndon, S., & Corlett, P. R. (2020). Hallucinations in posttraumatic stress disorder: insights from predictive coding. *Journal of Abnormal Psychology*, *129*(6), 534. https://doi.org/10.1037/abn0000531

Macpherson, F. (2017). The relationship between cognitive penetration and predictive coding. *Consciousness and Cognition*, *47*, 6–16. https://doi.org/10.1016/j.concog.2016.04.001

Marcus, G., Marblestone, A., & Dean, T. (2014). The atoms of neural computation. *Science*, *346*(6209), 551–552. https://doi.org/10.1126/science.1261661

Martin J. M., Solms M., Sterzer P. (2021). Useful misrepresentation: perception as embodied proactive inference. *Trends in Neurosciences, 44(8), 619-628.* https://doi.org/10.1016/j.tins.2021.04.007

Metzinger, T. (2014). How does the brain encode epistemic reliability? Perceptual presence, phenomenal transparency, and counterfactual richness. *Cognitive Neuroscience*, *5*(2), 122–124. https://doi.org/10.1080/17588928.2014.905519

Montague, M., & Bayne, T. (2017). *Cognitive phenomenology*. Oxford University Press.

Orlandi, N. (2016). Bayesian perception is ecological perception. *Philosophical Topics*, *44*(2), 327-352. https://doi.org/10.5840/philtopics201644226

Orlandi, N. (2018). Predictive perceptual systems. *Synthese*, 195(6), 2367-2386. https://doi.org/10.1007/s11229-017-1373-4

Ramachandran, V. S. (1988). Perception of shape from shading. *Nature*, *331*(6152), 163–166. https://doi.org/10.1038/331163a0

Rescorla, M. (2015). Bayesian perceptual psychology. In: M. Matthen (Ed.), *The oxford handbook of the philosophy of perception* (pp. 694-716). Oxford University Press.

Rossi-Pool, R., Zainos, A., Alvarez, M., Parra, S., Zizumbo, J., & Romo, R. (2021). Invariant timescale hierarchy across the cortical somatosensory network. *Proceedings of the National Academy of Sciences*, *118*(3), e2021843118. https://doi.org/10.1073/pnas.2021843118

Schmack, K., Gomez-Carrillo de Castro, A., Rothkirch, M., Sekutowicz, M., Rossler, H., Haynes, J. D., Heinz, A., Petrovic, P., & Sterzer, P. (2013). Delusions and the role of beliefs in perceptual inference. *Journal of Neuroscience*, *33*(34), 13701–13712. https://doi.org/10.1523/JNEUROSCI.1778-13.2013

Spohn, W. (2008). *Causation, Coherence and Concepts: A Collection of Essays*. Springer Science & Business Media.

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*, *112*, 92-97. https://doi.org/10.1016/j.bandc.2015.11.003

Stawarczyk, D., Bezdek, M. A., & Zacks, J. M. (2021). Event representations and predictive processing: the role of the midline default network core. *Topics in Cognitive Science*, *13*(1), 164–186. https://doi.org/10.1111/tops.12450

Sterzer, P., Adams, R. A., Fletcher, P., Frith, C., Lawrie, S. M., Muckli, L., Petrovic, P., Uhlhaas, P., Voss, M., & Corlett, P. R. (2018). The predictive coding account of psychosis. *Biological Psychiatry*, *84*(9), 634–643. https://doi.org/10.1016/j.biopsych.2018.05.015

Teufel, C., & Fletcher, P. C. (2020). Forms of prediction in the nervous system.

*Nature Reviews Neuroscience*, *21*(4), 231–242. https://doi.org/10.1038/s41583-020-0275-5

Tschantz, A., Millidge, B., Seth, A. K., & Buckley, C. L. (2022). Hybrid Predictive Coding: Inferring, Fast and Slow. arXiv preprint arXiv:2204.02169.

Vance, J. (2015). Cognitive Penetration and the Tribunal of Experience. *Review of Philosophy and Psychology*, *6*(4), 641–663. https://doi.org/10.1007/s13164-014-0197-0

Vance, J., & Stokes, D. (2017). Noise, uncertainty, and interest: predictive coding and cognitive penetration. *Consciousness and Cognition*, *47*, 86–98. https://doi.org/10.1016/j.concog.2016.06.007

Wiese, W., & Metzinger, T. K. (2017). Vanilla PP for philosophers: A primer on predictive processing. In T. Metzinger and W. Wiese (Eds.)*, Philosophy and Predictive Processing* (pp. 1-19). MIND Group. https://doi.org/10.15502/9783958573024

Williams, D. (2019). Hierarchical minds and the perception/cognition distinction. *Inquiry*, 1–23. https://doi.org/10.1080/0020174X.2019.1610045

Williams, D. (2020). Predictive coding and thought. *Synthese*, *197*(4), 1749–1775. https://doi.org/10.1007/s11229-018-1768-x

# Chapter 5

# Belief Now, True Belief Later: The Epistemic Advantage of Self-Related Insights in Psychedelic-Assisted Therapy

This work was published as “Caporuscio, C. (2022). Belief now, true belief later: The epistemic advantage of self-related insights in psychedelic-assisted therapy. *Philosophy and the Mind Sciences*, *3*.” Slight revisions have been made to fit the dissertation format.

It was supported by funding from the Deutsche Forschungsgemeinschaft
(RTG2386)

## 5.0. Abstract

Chris Letheby’s defence of psychedelic therapy hinges on the premise that psychedelic-facilitated insights about the self are in a better epistemic position than those about the external world. In this commentary, I argue that such a claim needs further defending. More precisely, I argue that one element is underexplored in Letheby’s otherwise compelling picture: namely, that unlike new beliefs about the external world, beliefs about oneself have the capacity to turn into self-fulfilling prophecies. Recognising the psychedelic experience and the subsequent integration process as opportunities not only to apprehend certain facts about the self but also to actively shape and redetermine those facts is key to understanding the epistemic differences between insights patients have about themselves and about the external world.

## 5.1. Introduction

According to the Comforting Delusion Objection to psychedelic therapy (henceforth CDO), psychedelic therapy should not be recommended even though its psychological effects are beneficial.  The reason for such skepticism is that psychedelics produce positive psychological effects only because they induce comforting beliefs in a joyous cosmology, a divine consciousness, or an ultimate reality. Because such beliefs are incompatible with naturalism and therefore most likely false, defenders of the CDO argue that the therapeutic potential of psychedelics comes with large epistemic costs that outweigh their psychological benefits (Letheby 2021, p. 2).

Chris Letheby's book The Philosophy of Psychedelicsis an elaborate and largely convincing rebuttal of the CDO. Letheby accepts that the ethical status of psychedelic therapy hinges on its epistemic status. However, he contends that mental wellbeing is promoted not by changes in metaphysical beliefs but by changes in beliefs about the self: for example, "I am in touch with my emotions" (Watts et al. 2017), "My identity is not tied to being a smoker" (Noorani et al. 2018). According to Letheby, these self-related beliefs are less likely to be delusional than metaphysical beliefs, making the mechanism of psychedelic therapy epistemically (and therefore ethically) innocent.

When assessing Letheby's argument, we are immediately faced with the following question: If psychedelic-induced beliefs about external reality are probably false, why should the same not be true of psychedelic-induced beliefs about ourselves? Unlike the metaphysical beliefs questioned by the CDO, beliefs about the self are compatible with naturalism; however, this does not necessarily make them more likely to be true. Letheby makes a good case for the non-propositional epistemic benefits of psychedelic self-related insights, such as the acquisition of knowledge-how and knowledge by acquaintance, and for the indirect epistemic benefits gained through increased psychological well-being (see Chapters 8.4–8.8). However, his argument about the direct acquisition of knowledge-that about the self through psychedelics, discussed in Chapter 8.3, is less compelling.

In what follows, I present a three-part response to Letheby's chapter.

1. I reconstruct Letheby's argument for the acquisition of knowledge-that through psychedelics, and suggest that it leaves self-related insights in a very similar epistemic position to that of beliefs about the external world.
2. I argue that, if psychedelics are equally likely to bring about false beliefs about the self and about the world, Letheby's reply to the CDO is not very convincing.
3. I offer an alternative argument, underexplored in Letheby's book, as to why self-related insights might after all be in a better epistemic position than insights about the external world.

## 5.2. Letheby's argument for the acquisition of knowledge-that

First, Letheby's argument. Chapter 8.3 defends that we have grounds to believe that at least some of the self-related insights facilitated by the psychedelic experience are genuine and can promote the acquisition of new propositional knowledge, or knowledge-that. The argument for the possibility of obtaining knowledge-that about oneself from psychedelic administration goes like this:

1. Decreasing the weighting of self-related priors can increase the probability of accurately apprehending certain facts about oneself (from general Predictive Processing theory)
2. Psychedelic administration temporarily decreases the weighting of self-related priors (from the REBUS model and the predictive self-binding theory)
3. Therefore, psychedelic administration can increase the probability of accurately apprehending certain facts about oneself (Letheby 2021, p. 169).

There is a problem with this argument, which Letheby himself points out: The fact that psychedelic administration can increase the possibility of accurately apprehending certain facts does not mean that all insights will be accurate.

As Andy Clark (2016, p. 288) puts it, priors are always both constraining and enabling. By filtering out evidence that contradicts them, they can impair access to certain facts, but they can also prevent implausible hypotheses from being considered and accepted. A temporary loss of confidence in the brain's prior knowledge does not imply that all resulting beliefs will be veridical, nor that they will all be false: "[...]when it comes to propositional knowledge about our own mind, psychedelics facilitate both genuine insights and placebo insights, and there is no general formula for telling the two apart" (Letheby 2021, p. 171). According to Letheby, the only way to assess the accuracy of these insights is through sober integration after the psychedelic session.

### 5.3. Do self-related beliefs deserve special status?

Letheby's emphasis is on beliefs about the self. This is central to his overall argument: he rejects the CDO by arguing that the beliefs that mediate psychological wellbeing are those about the self, and that beliefs about the self are less likely to be false than metaphysical beliefs.

However, the argument about gaining knowledge-that reconstructed in section 5.2 can be applied to self-independent beliefs as well. An example of priors constraining our knowledge about the external world is provided by a popular explanation of the Hollow Mask Illusion, according to which a strong reliance on the assumption "faces are convex" causes our brain to ignore evidence to the contrary and perceive a rotating concave mask as popping out. According to this explanation, a loss of confidence in the brain's top-down priors causes patients with schizophrenia (Dima etal. 2010) or people on psychedelics (Millière et al. 2018) to perceive the mask more accurately than controls.

Hence, Letheby's argument for the possibility of obtaining knowledge-that about the self through psychedelic use can be generalised as follows:

1. Decreasing the weighting of priors can increase the probability of accurately apprehending certain facts (from general predictive processing theory) and inaccurately apprehending others
2. Psychedelic administration temporarily decreases the weighting of priors (from the REBUS model and the predictive self-binding theory).
3. Therefore, psychedelic administration can increase the probability of accurately apprehending certain facts and inaccurately apprehending others.

If the argument generalises, the special status reserved by Letheby for self-related beliefs seems unjustified. Of all new insights acquired after a psychedelic session, some of those about the external world may be veridical, and some of those about ourselves may be false (or vice versa).

Imagine your friend Maria (Letheby 2021, p. 162), who claims to have gained propositional knowledge from a psychedelic experience. When you ask her what she has learned, she responds by listing three new beliefs she has gained. First, she has discovered some aspects of herself that were hidden from her before, including a deep desire for human connection. Secondly, she has discovered something she feels is a deep truth about another person's mind: She realises that the actions of a family member that she had always thought stemmed from selfishness and greed are actually motivated by anxiety and insecurity. Thirdly, she has gained a metaphysical insight, namely that all existence in time is equally real.

All of these insights are plausible, and all of them have the potential to cause lasting psychological benefits to Maria. But is one of these insights more likely to bring epistemic benefits, and in particular new propositional knowledge? In other words, are there epistemic differences between psychedelic-mediated insights about the self, about another person, and about the external world?

I argue that relevant epistemic differences between self-, other-, and world-related insights cannot be found if we treat the epistemic status of the psychedelic experience as purely dependent on acquiring new knowledge of pre-existing facts. Maria's insight about herself and her two insights about self-independent objects are all likely to have been caused by a weakening of her prior beliefs, which allowed her to see herself and others in a new light.  But this is no guarantee that any of the newly acquired beliefs is true.

Thus, Letheby's optimism about the possibility of gaining propositional knowledge about oneself rather than the external world seems unmotivated. Psychedelics put Maria in an epistemically promising but uncertain position, where she is likely to have acquired true beliefs about the world, other people and herself that she can only tentatively differentiate from the false ones by carefully scanning them for plausibility after her session.

If it is the case that self-related insights have no firmer epistemic grounding than self-independent ones, Letheby's rejection of the CDO is considerably weakened. At least some of the self-related insights driving psychological improvement are probably still comforting delusions, and the epistemic status of psychedelic therapy is only partially rescued by indirect epistemic benefits and post-session evaluation. There are obvious epistemic faults in a therapeutic method that will sometimes work by convincing a person who lies continually that they are honest and dependable, for example. And because Letheby accepts the CDO's premise that epistemically bad means ethically bad, the possibility of comforting delusions about the self is a reason to refrain from recommending psychedelic therapy.

### 5.4. The power of self-shaping

I propose an alternative reason why Maria might be in a better epistemic position regarding her insights and newly acquired beliefs about herself than regarding those about the external world. More precisely, I will argue that one element is underexplored in Letheby's otherwise compelling picture: namely, the recognition of the psychedelic experience and the subsequent integration process as opportunities not only to apprehend certain facts about oneself but also to actively shape and redetermine those facts through exploratory thinking and behaviour.

Moran (2001) and McGeer (2008,1996; McGeer  and Pettit 2002) talk about the the power of first-person authority to shape the self. In their view, the authority of self-knowledge derives not from a passive, error-free ability to detect our mental states, but from our capacity to regulate our thoughts and actions in accordance with the claims we make about ourselves. For example, the thought "I hate laundromats" might contribute to creating and sustaining the declared hate for laundromats, thus turning into a self-fulfilling prophecy (Schwitzgebel 2011). Deciding as a child that your favourite colour is blue might causally influence yourchoice of outfits, objects, and self-expression, feeding into a growing appreciation for the colour blue.

According to McGeer (2008), self-shaping is not only a capacity, but a moral responsibility: In order to be intelligible as rational agents, we owe it to ourselves and others to behave and think in the ways we declare we do. Our core beliefs about ourselves can (and should) turn into self-fulfilling prophecies, providing us with familiar patterns of expression and behaviour, allowing us to act as predictable agents, and fulfilling the expectations that we have created in ourselves and others.

However, this also means we might end up stuck in our core beliefs about ourselves. In pathological cases, this is extremely problematic. A depressive patient who believes they are unable to find pleasure in going outside will behave in accordance with this belief – or, in terms borrowed from the active inference framework, they will sample their environment for evidence that will confirm the belief and avoid evidence that will disconfirm it (Ramstead et al. 2020). Not only will this behaviour reinforce the belief in an endless loop, but it may also make it true: by committing to act in line with the belief that they are incapable of getting out of bed and having a nice day, they will make it impossible for themselves to enjoy being outside.

In his book, Letheby argues that the psychedelic experience allows for a relaxing and rewiring of self-related priors, thus allowing patients with negative self models to access evidence about themselves that was previously hidden because it conflicted with those models. This is likely to be true, but not the whole story: If damaging self-related beliefs are (at least to some extent) self-fulfilling, losing confidence in them allows psychedelic users not only to access previously hidden evidence but also to create new evidence by acting and thinking in new, unconstrained ways.

Imagine, in line with Moran and McGeer's examples, that by thinking of herself and presenting to others as self-sufficient, independent, and emotionally distant, Maria has been committing for most of her adult life to act and think in a way that would fulfil that expectation. Because of this, at the time of her psychedelic experience, evidence of a deep desire for human connection is not only hidden by her self model, but scarce. She has been leading a life of voluntary isolation, keeping distance from her family and friends, and she has rarely been imagining or wishing for a different lifestyle.

However, by relaxing her belief about herself through psychedelic use, she is able to temporarily break free from her commitment to act and think in line with it. She can exploratorily entertain new thoughts, imagine new modal truths about how her life could be (Letheby 2015; 2021), and test out different behaviours, like opening up and expressing closeness to her trip sitters or companions. These new thoughts and behaviours then contribute to the evidence for her new, emerging insight that at her core is a deep desire for human connection. If Maria comes out of her trip with strong confidence in this belief then this confidence will, in turn, exercise its power to shape the self. During the sober integration period, Maria will consolidate her belief by maintaining and incorporating into her life the thoughts and behaviours she tentatively explored during the psychedelic experience.

Contrast this with the knowledge-acquisition story described by Letheby in Chapter 8 of Philosophy o Psychedelics. Maria is not only detecting previously unnoticed patterns of her prior behaviour that indicate her deep desire for human connection; she is (during the psychedelic experience) exploring new patterns of thought and action in line with her newly formed belief, and (during the subsequent integration period) sticking with them.

Let us return to the epistemic status of self-related insights mediated bypsychedelics. Two questions might be asked:

1. Has Maria acquired new propositional knowledge?
2. Is she in a better epistemic position regarding her insights about herself than regarding those about the external world?

Her epistemic position regarding her new belief that she is someone with a deep desire for human connection is an interesting one. Prior to the psychedelic experience, her position would probably have been inaccurate: She thought that she was happy by herself and never had thoughts or carried out behaviours indicating that she was longing for connection. However, months after the experience, Maria finds herself enjoying the company of others and letting barriers down with her loved ones, as she has learned to do during her trip and has consolidated the habit of doing during the integration period. Belief first, true belief later: She might have ended her trip with an illusory insight, but months later this has turned into new propositional knowledge.

Does the same apply to her self-unrelated belief that the actions of her family member were motivated by anxiety and insecurity? Imagine this second insight was also, prior to the experience, false: This person's actions were in fact guided by selfishness and greed. Does this inaccurate belief have any self-fulfilling capacity? Maybe, indirectly and to a much lesser extent: By reconnecting with the other person thanks to her favourable disposition, Maria might be able to partially influence their ways of acting and thinking in line with her belief. However, this is likely to be much harder and only successful in specific circumstances (for one, the other person must be receptive and willing to connect).

Finally, beliefs whose object is completely mind-independent, like Maria's metaphysical insight about the nature of time, are even more clearly not self-fulfilling. Her belief about the nature of time will not change the nature of time. Maria's capacity to shape and influence self-independent objects and people is not comparable to her capacity to shape and influence her own mental states and behaviour.

### 5.5. Concluding Remarks and Future Directions

The self-fulfilling character of these beliefs has some noteworthy implications. First, my account suggests a more prominent role for the integration period and psychological support following the psychedelic experience, which aligns with existing evidence that these elements are important predictors of successful outcomes (Johnson et al. 2008; Teixeira et al. 2022). In Letheby's account, follow-up sessions are important for epistemic purposes because patients can soberly scan insights for plausibility and distinguish accurate ones from placebos with the help of a trained therapist. While this is likely true, I argue that truth-testing is not the main mechanism that renders the months following the experience crucial in determining epistemic benefits.

The epistemic success of psychedelic therapy happens in two steps: The session is for discovery, and its aftermath is for consolidation and commitment. During the psychedelic experience itself, self-related beliefs are relaxed, allowing patients to momentarily escape the self-fulfilling effects of damaging self models, explore new and healthier ways of thinking and acting, and consequently acquire new insights about themselves. However, the process of self-shaping involves changing one's behaviour and thinking over much longer timescales than six or twelve hours. This is why most of it happens while sober. With the help of trained therapists, patients can consider which beliefs about themselves that have been tested out during the psychedelic experience are worth committing to, and plan for ways to change their behaviour to integrate and fulfil these new (or newly appraised) beliefs. If they do not adapt their lifestyle, thoughts, and actions to conform to their psychedelic-induced insights, their newly formed beliefs are unlikely to turn out to be true and will probably be abandoned.

The second consequence of the account I have presented is that while the self-fulfilling nature of self-characterisations helps mitigate the epistemic risks of psychedelic therapy, it also carries serious psychological risks. Not only is there nothing intrinsically true about psychedelic-mediated insights, there is also nothing intrinsically positive: While some people report feelings of bliss and newfound self-acceptance, others experience anxious spiralling, negative emotions, and self-deprecating thoughts. If positive insights can actively shape users' future mental states and behaviour, so can psychologically damaging ones. In this account, a bad trip triggering the new belief that one is deeply incapable of being happy carries more than just epistemic risks.

In order to prevent damaging self models from being accepted and incorporated as a result of a negative psychedelic experience, it is extremely important to take the danger seriously and carefully assess whether and to what extent it can be reduced. Further work is needed before psychedelic-assisted therapy can be safely recommended. I suggest that, in order to mitigate the risk, we should focus future research efforts on the following two aspects. First, appropriate preparation before the psychedelic experience, moderate dosage and controlled set and setting can significantly reduce the chances of a bad trip. Secondly, psychological and behavioural support during the integration period can help an individual not only to integrate psychologically and epistemically beneficial insights into their life, but also to recognise and discard damaging ones.

Psychedelics can facilitate a state of temporary flexibility where damaging self models can be discarded and new ones can take their place, be consolidated and powerfully shape one's future self. Old beliefs are abandoned, and new beliefs are accepted and turned into reality. However, this increased flexibility is neutral in itself: it is what happens before, during and after the session that really determines whether the newly adopted models will be harmful or therapeutic.

### 5.6. References

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190217013.001.0001

Dima, D., Dietrich, D. E., Dillo, W.,  and Emrich, H. M. (2010). Impaired top-down processes in schizophrenia: A DCM study of ERPs. *NeuroImage,52(3), 824–832.*https://doi.org/10.1016/j.neuroimage.2009.12.086

Johnson, M., Richards, W., & Griffiths, R. (2008). Human hallucinogen research: Guidelines for safety. *Journal of Psychopharmacology, 22(6), 603–620.* https://doi.org/10.1177/0269881108093587

Letheby, C. (2015). The philosophy of psychedelic transformation. Journal of Consciousness Studies, 22(9-10), 170–193.

Letheby, C. (2021). *Philosophy of Psychedelics*. Oxford University Press.

McGeer, V. (2008). The moral development of first-person authority. *European Journal of Philosophy, 16(1), 81–108.* https://doi.org/10.1111/j.1468-0378.2007.00266.x

McGeer, V. (1996). Is "self-knowledge" an empirical problem? Renegotiating the space of philosophical explanation. *The Journal of Philosophy, 93(10),* 483–515. https://doi.org/10.2307/2940837

McGeer, V., & Pettit, P. (2002). The self-regulating mind. *Language & Communication, 22(3), 281–299.* https://doi.org/10.1016/S0271-5309(02)00008-3

Millière, R., Carhart-Harris, R. L., Roseman, L., Trautwein, F.-M., &

Berkovich-Ohana, A. (2018). Psychedelics, meditation, and self-consciousness. *Frontiers in Psychology, 9.* https://doi.org/10.3389/fpsyg.2018.01475

Moran, R. (2001). *Authority and Estrangement: An Essay on Self-knowledge.* NJ: Princeton University Press.

Noorani, T., Garcia-Romeu, A., Swift, T. C., Griffiths, R. R., & Johnson, M. W. (2018). Psychedelic therapy for smoking cessation: Qualitative analysis of participant accounts. *Journal of Psychopharmacology, 32(7), 756–769.* https://doi.org/10.1177/0269881118780612

Ramstead, M. J. D., Wiese, W., Miller, M., & Friston, K. J. (2020). Deep neurophenomenology: An active inference account of some features of conscious experience and of their disturbance in major depressive disorder. [Preprint]. http://philsci-archive.pitt.edu/18377/.

Schwitzgebel, E. (2011). Introspection, what? In Smithies, D., & Stoljar, D. (Eds.) *Introspection and Consciousness (pp. 29–48).* Oxford University Press.

Teixeira, P. J., Johnson, M. W., Timmermann, C., Watts, R., Erritzoe, D., Douglass, H., Kettner, H., & Carhart-Harris, R. L.(2022). Psychedelics and health behaviour change. *Journal of Psychopharmacology, 36(1),* 12–19. https://doi.org/10.1177/02698811211008554

Watts, R., Day, C., Krzanowski, J., Nutt, D., & Carhart-Harris, R. (2017). Patients' accounts of increased "connectedness" and "acceptance" after psilocybin for treatment-resistant depression. *Journal of Humanistic Psychology, 57(5), 520–564*. https://doi.org/10.1177/0022167817709585

# Chapter 6
# Concluding Remarks and Future Directions

## 6.1. Overarching Conclusions

This dissertation is a collection of self-standing essays. Despite this, the work presented here has a lot of thematic overlap and some overarching conclusions about the relationship between experience and belief can be drawn.

The first one regards the direction from experience to belief. Experience can trigger and inform a a new belief: waking up to the smell of coffee can lead me to believe that it's morning, and feeling pressure in my chest can motivate the belief that I'm getting nervous about something. However, experience is rarely the sole driving force of belief formation, with the exception of uninformative judgements such as "I am feeling this" (Gertler 2012). Pre-existing beliefs, background knowledge and motivational factors all play a role in determining and shaping new beliefs, and can stray them away from being accurate depictions of experience. Accepting a new belief as true is the result of a process that involves a delicate balance between phenomenal evidence, namely experience, and cognitive evidence, such as background beliefs (Chapter 2).

This is true for delusional beliefs too. However, in pathological cases it is common for the balance between experiential and cognitive evidence in informing new beliefs to be disrupted. Aberrant experiences sometimes disproportionally drive belief formation; or in the opposite scenario, strong background beliefs or motivational factors might motivate new delusions in spite of someone's phenomenal experience. The first case applies to delusions that are very close in content to hallucinatory experiences, such as the belief of having an extra limb caused by the experience of phantom pain. The background belief that humans normally only have two limbs is present, but not sufficient to override the perceive significance of the experienced pain. The second case could apply to introspective delusions, of which Anton-Babinski syndrome is a possible candidate: patients might not be having visual experiences, but the motivational force of blindness denial could be strong enough to override experience when they form the belief that they are seeing (Chapter 3).

The second conclusion regards the direction from belief to experience. Our beliefs can affect the way we experience the world: not only through cognitive penetration, namely by directly altering our lower-level perceptual states, but also by guiding action and therefore affecting how we actively sample the external world. This is especially true for beliefs about the self: by acting in accordance to our beliefs about who we think we are, we turn our self models into self-fulfilling prophecies. This process takes repeated behavioural and thinking patterns over time: for this reason, self-related beliefs have an epistemic advantage when it comes to beliefs about the self that are built over time (Chapter 5) but the same advantage does not always apply when it comes to immediate detection of our mental states (Chapter 2).

Cognition does not always penetrate perception, and perception does not always penetrate cognition. Sometimes, perceptual and cognitive systems appear as if they are to an extent informationally segregated. Perception and belief might be predictions at different levels of the same inferential hierarchy, as Predictive Processing suggests, but their function is different. Inflexible constraints present at lower levels of perceptual systems can allow perception to be fast and computationally efficient, capable of quickly identifying danger or opportunities for action. On the other hand, the absence of such constraints at higher levels of the hierarchy makes cognitive predictions less computationally efficient and slower, but better able to flexibly take into account context-dependent information (Chapter 4).

### 6.2. Future Directions

An in-depth discussion of future directions can be found at the end of every chapter. Here, I want to focus on two intriguing questions that I am planning to explore further in my future research.

Chapter 1 and 2 focus on introspection, and argue that we should take the possibility of introspective errors seriously. This provides an opportunity to challenge the idea of introspective incorrigibility, which asserts that introspection always supersedes any external access when there is a disagreement between the two. Especially when put into the context of psychiatry, this leads to an important ethical question: Is it ever ethical to question someone's perspective on their own experience based on second or third-person evidence?

The ethical dimension of rejecting introspective incorrigibility is going to be increasingly important in the future, as third-person methodologies continue to evolve and enhance their accuracy and reliability. It is crucial to note that assessing introspective reports should not entail disregarding them. While assessing disputes it is important not to take introspection at face value because of an assumption of infallibility; however, someone's first-person perspective is still an extremely valuable source of insight into their own experience. First, second, and third-person perspectives should be all considered as complementary sources of evidence about the same object, and any conclusion about someone's experience should be arrived at in collaboration with them.

Another question regards the safety of psychedelic experiences as a means of therapeutic improvement. As defended in Chapter 5, psychedelic-assisted therapy is a new experimental paradigm that yields promising results in the treatment of mental disorders. However, there are very important concerns as to whether psychedelic-assisted therapy can ever become regulated and accepted as a standard practice in psychiatry. Psychedelics are capable of inducing powerful experiences that can cause radical belief changes. This raises doubts about the safety and ethics of such substances: can they be challenging to one's sense of personal identity? Can subjects give inform consent to largely unpredictable experiences? Can the therapist provide safety, comfort and integration practices without disproportionately influencing the patient's ideas due to their vulnerable and malleable state? What kind of procedures should be in place to maximize therapeutic benefit and psychological safety? This is a new ethical territory to navigate, and assuming that the risks can be controlled, there will be a need for additional or different ethical guidelines that are not currently considered in traditional therapy.

# APPENDIX

This appendix contains a book review and a commentary that I have published during the course of my PhD. Although they are not strictly part of my PhD work, I decided to include them here because I consider them to be relevant to my work. As I note in the introduction to this dissertation, both Levy's *Bad Beliefs* (2021), and Bermudez' *Frame it Again* (2020) have the merit of paying attention to contemporary cases of flawed beliefs, such as climate change denial, conspiracy theories or political debates, with the intention of understanding some of the real, imperfect mechanisms that inform our belief formation processes.

## Appendix A. Is Framing Irrational?

This book review has been published as “Caporuscio C. (2022) Is framing irrational?, *Philosophical Psychology*, DOI: 10.1080/09515089.2022.2158072”

It was supported by funding from the Deutsche Forschungsgemeinschaft (RTG2386).

### 1. Is framing irrational?

“Frame It Again. New Tools for Rational Decision-Making” by José Luis Bermúdez is a powerful defence of a traditionally unappreciated aspect of human cognition: framing effects, namely, the tendency to value the same thing differently when it is accompanied by a different narrative. “Pro-life” sounds very different than “against-choice”. “Leaving your home behind” evokes different emotions and might lead to a different decision than “starting a new adventure”. This remains true even if you are consciously aware of the fact that the first two expressions both refer to people that are against legal abortion, and the latter are just different ways to look at the choice to move to a new country.

It is easy to see how framing got a bad reputation. After all, “That which we call a rose/by any other name will smell as sweet”: attaching one narrative rather than another to the same object doesn’t alter the object itself. Why then should it change the value we assign to it, or alter our decision making? The mainstream answer to this question is clear: it should not, and yet it does. This is because humans are fundamentally irrational: they often do not reason how they ought to reason.

In Chapter 1, Bermúdez calls this idea “the litany of irrationality”. The litany of irrationality permeates the literature on belief formation and decision making, and is made up for two parts: one is a normative theory of ideal rational reasoning, and one is a descriptive narrative on how humans actually reason. According to the litany, the latter deviates from the former regularly and spectacularly. Framing effects are often considered a clear example of this deviation.

If judging the same object differently when it's framed in different ways is irrational, then defenders of the litany of irrationality are right in judging that humans don't do very well at rationality. As Bermúdez convincingly shows in chapters 2 and 3, people fall victims to framing effects all the time - in experimental contexts, but also and especially when real gains and losses are concerned. Some infamous effects that regularly trick investors into bad decisions can be created or eliminated by framing. One example is the disposition effect: investors are way more likely to move out of a loss and into a new position if it is not framed as "taking a loss", but as "transferring assets".

So, are framing effects irrational? According to Bayesian decision theory, yes. Chapter 4 argues that, in Bayesian terms, rationality means internal consistency. This has two important consequences. One is extensionality: the identity of a set is fixed by its members, irrespective of how the members themselves might be described. The second is that a standard requirement of internal consistency is that preferences be transitive: if you prefer A to be and B to C, then you should prefer A to C. Framing effects cause non-transitive preferences, thus contradicting this rule. Take the case of the Greek leader Agamemnon, faced with the alternative of sacrificing his daughter to please Artemis or not being able to leave for the Trojan war, thus failing his ships and people. Agamemnon's dilemma comes from framing the first choice in two very different ways: *murdering his daughter* or *following Artemis' will*. *Following Artemis' will* sounds much better to him than *failing his ships and people*, but even f*ailing his ships and people* is not as bad as *murdering his daughter*. So far, so good - A is better than B that is better than C. But A and C - following Artemis' will and murdering his daughter - are the same outcome, and Agamemnon knows it. He thus finds himself with non-transitive preferences (or quasi-cyclical preferences, as Bermúdez calls them) due to framing effects.

No hope left for rational frames, then? Not quite. In Chapter 5, Bermúdez confronts us with some compelling critical cases where rational decision-makers are confronted with quasi-cyclical preferences. These are all complex cases, where the frames highlight different (and equally compelling) aspects of an outcome. Some of them have to evoke complicated and contradictory emotions, or force the decision-maker to a choice between opposing values. In all of these cases, even when a decision is reached, the quasi-cyclical preferences do not seem to completely resolve, giving rise to regret. Agamemnon finally makes the choice to sacrifice Iphigenia, but this does not eliminate the *murdering his daughter* frame from his mind. And rightly so, because both frames highlight an important aspect of what is going on: on one side, Agamemnon's private role as a father, on the other, his public role as a leader. Agamemnon is both a father and a leader, and considering only one of these perspectives would mean oversimplifying the situation. In contrast, the figure of Abraham, who does not hesitate one second to kill his son to follow God's orders, does not seem more rational or a better decision maker. It is hard to resist the author's point here: if anything, the biblical character seems to lack depth compared to the tormented decision-making process of Agamemnon.

Bermúdez convincingly argues that this is because one important principle of rationality is left out in the discussion about framing effects: the due diligence requirement. "In setting up a decision problem, rational decision-makers need to be appropriately sensitive to as many potential consequences of the different courses of action available to them as possible" (Chapter 6). It would be a failure of rationality for Agamemnon to ignore his role as a father of Iphigenia or as a leader of the Greeks, because each of these framings highlights different values, emotions, and consequences that are all part of the decision he's about to make. As long as there are no explicit contradictions regarding objective facts, considering as many frames as possible is rational, even if it might lead to quasi-cyclical preferences.

In Chapters 7, 8, 9 and 10, Bermúdez supports his claim with an array of practical examples of framing effects and quasi-cyclical preferences in the wild. The challenges of self-control, game theory, political deadlocks - all of these can be explained, understood and sometimes even used to our advantage through framing effects. The clashes between the left and the right on matters like gun control, taxation, or abortion are ultimately clashes between frames. And one key thing about frames is that they tend not to disagree on factual, objective matters, like whether Iphigenia is the daughter of Agamemnon or whether a foetus at the third month has a beating heart. They are clashes of values.

Where does this leave us? Considering different frames can be rational, but how to make a rational decision when there is no way to factually decide between two perspectives? This takes us to the most interesting (and underdeveloped) part of the book. In the last chapter, Bermúdez develops four steps or techniques for frame-sensitive reasoning. These involve stepping outside one's own frame, imaginatively simulating different frames or perspectives, holding multiple frames in mind at the same time, and weighting the reasons they generate on a single scale. This will not always be successful, Bermúdez admits, and often it won't produce one real winner or a clear course of action - but the rationality of frame-sensitive reasoning is determined by the process, rather than by the result.

I found the last section to be less convincing than the rest of the book. In particular, Bermúdez does not really explain to a satisfactory extent how the last step, namely the weighting of reasons generated by different frames on a single scale, is supposed to happen. He has an example of successful frame-sensitive reasoning regarding the Community Charge in Great Britain, a fixed flat-rate tax charged to every adult resident. Earlier in the book, Bermúdez argues that different ways of framing the Community Charge reflected a clash between the values of fairness as equity versus fairness as equality. Through the final step of frame sensitive reasoning, the reasoner might come to the view that fairness as equity subsumes fairness as equality. I suspect, however, that this view would only be endorsed by someone who already started with the equity frame in mind. Values are, by Bermúdez' definition, not factual or objective states of affairs. They are ways of seeing the world and they are often not tied to a single problem or dilemma, but to the way we see ourselves and the world at large. And philosophers have often defended that values are incomparable, namely, they cannot be measured on a single scale in order to decide which one is superior to the other (see for example Chang 2002). Thus, I am sceptical as to whether we can abstractly reason ourselves out of values, and I do not think that Bermúdez offers a sufficient explanation for this step in his last chapter.

Another point that I found underdeveloped is the relation to what Bermúdez calls "the litany of irrationality". Under the umbrella term of "framing effect" are included some of the most standardly used textbook examples of human reasoning errors. This, however, is not to say that the litany of irrationality is false.

Bermúdez does an excellent job at persuading us that not all framing effects are irrational, but he does not exclude that some might be. I suspect that a lot of the preferred examples of the defenders of human irrationality belong to this last category. In experiments like the Asian Disease Paradigm, where framing effects disappear when subjects are made aware of the trick, the effect seems to be more the result of an automatic association than of a clash between irreducible values. Overall, a lot of the examples in Chapters 2 and 3 seem quite different in this sense than the examples in Chapters 7, 8, 9 and 10. In the former, the effect seems to be rooted in objective, factual errors due to psychological mechanisms that Kahnemann (2001) would attribute to System One. In the latter, frames are different and equally valid perspectives of looking at a complex situation. If something very different is going on in these two cases, to the point that one is rational and the other is textbook irrational, it might be worth investigating whether the term "framing effect" really refers to one category of phenomena. Are the proponents of human irrationality really talking about cases like Agamemnon's when they refer to framing effects?

Despite these doubts, "Frame it again" is an excellent book. The argumentation is clear and to the point, confronting the reader with a compelling defence of framing effects. The examples of rational frames brought forward by the author are convincing and pervasive, ranging from games to political decisions, from mythological and literary characters to our day-to-day life. I suspect that his analysis will encourage researchers to look at framing effects from a more positive perspective and explore underdeveloped applications. For example, stepping out of our old frames and imaginatively creating new ones could be an interesting perspective to look at strategies for mental health and therapy.

Bermúdez' main point is compelling: the human tendency to hold different perspectives at the same time is a feature, not a bug. Considering different frames when confronted with difficult (private and public) decisions is an epistemic requirement. There is more to ideal rationality than Bayes.

## 2. References

Bermúdez, J. L. (2020). *Frame it again: New tools for rational decision-making*. Cambridge University Press.

Chang, R. (2002). *Making Comparisons Count*, New York: Routledge.

Kahneman, D. (2011). *Thinking, Fast and Slow.* New York: Farrar, Straus and Giroux.

## Appendix B. What Is Left of Irrationality?

This work was published as "Murphy-Hollies, K. & Caporuscio, C. (2023) What is left of irrationality?, *Philosophical Psychology*, DOI: 10.1080/09515089.2023.2186220"

It was supported by funding from the Deutsche Forschungsgemeinschaft (RTG2386).

### 0. Abstract

In his recent book *Bad Beliefs and Why They Happen to Good People*, Neil Levy argues that conspiracy theories result from the same rational processes that underlie epistemic success. While we think some of Levy's points are valuable, like his criticism of the myth of individual cognition and his emphasis on the importance of one's social epistemic environment, we believe that his account overlooks some important aspects. We argue that social deference is an active process, and as such can be helped or hindered by epistemic virtues and vices. With this in mind, holders of bad beliefs acquire more responsibility than is considered by Levy.

### 1. Introduction

Conspiracy theories and misinformation are widespread phenomena. Claims on which the scientific community has long reached a consensus, like anthropogenic climate change, evolution, or the efficacy of vaccinations, are disputed by unreliable sources, whose alternative stories are believed by large parts of the population despite the lack of evidential support. The story according to which vaccinations are an elaborate masterplan by pharmaceutical companies to implant chips into members of the population has been repeatedly rejected by experts and yet, it remains for many a more attractive theory of the purpose of vaccinations than the mainstream view. Why would a rational agent believe such a far-fetched, convoluted and discredited story rather than one that is linear, simple and largely supported by evidence?

A popular answer to this question is that people often do not act as rational agents. Since the Enlightenment, rationality has been characterised as a largely individual process: our capacity to collect first-order evidence, weigh it appropriately and come to the most plausible conclusion, with the help of epistemic virtues such as being open-minded, tenaciously logical, conscientious, humble, and so on. According to this conception, accepting conspiracy theories and other bizarre beliefs is a failure of rationality that is to be blamed on individual biases, epistemic vices and reasoning errors which affect people's capacity to deal with first-order evidence. Bad beliefs are the result of irrational processes.

In his book "Bad Beliefs: Why They Happen to Good People", Neil Levy (2021) presents an alternative account that sees these bizarre beliefs as resulting from the same rational processes that underlie epistemic success. Levy argues that much of human cognition does not rely on first-order evidence as much as it does on social deference and higher-order evidence. Most people do not have the resources or background knowledge necessary to assess the first-order evidence for or against anthropogenic climate change, for example. Both the scientifically-minded individual and the conspiracy theorist need to trust second-order sources of information which have already analysed and interpreted first-order evidence. The only difference between the two individuals is where they place this trust.

According to Levy, this means that conspiracy theories are not to be blamed on individual epistemic agents, but on the epistemic environment they are immersed in. It is a rational choice to defer to trusted sources of information on topics where the first-order evidence is too complicated to deal with yourself. However, the polluted epistemic environment we live in makes it so that the higher-order sources that are more present and vocal in the lives of many are not the ones that should be trusted. Thus, if we want to combat the rise of conspiracy theories, we need to clean up the epistemic environment first; by making it so that people who are not experts cannot exhibit expert status, and by increasing credibility signals to the scientifically supported opinion. This way, we can make it easier for people to know and recognize which higher-order sources to trust.

A lot of Levy's points are timely and well-taken. We appreciate Levy's project of showing that the average person is not a stupid irrational being, but someone who makes choices which make sense to them. We also appreciate the emphasis on just how significant and important one's social epistemic environment is, and that knowledge production is very much a fundamentally shared enterprise. However, we believe that his account can overlook some important parts of the story: the social aspect of epistemic virtues and vices, and the role of active choice in belief formation. When considering these aspects, we think that falling for conspiracy theories and bad beliefs acquires more epistemic responsibility than Levy allows.

In section 2, we take a closer look at some of the examples discussed by Levy and consider how they affect what rationality, and opposingly, irrationality, mean. These examples look rational with hindsight but don't involve comprehensive understanding. Sometimes, we want more than this; we want to innovate the processes and conclusions we acquire socially by altering them and improving them and this takes closer engagement. This is especially true when it comes to high-stakes beliefs about climate change, or the safety of vaccines. We also consider that epistemic virtues can play a more valuable role here than Levy allows. In section 3, we argue that belief formation is an active process of picking sides based on one's self-conceptions, rather than a passive process where beliefs that are prevalent in our environment "happen" to us. This puts responsibility back into the picture and suggests that cleaning up the epistemic environment won't be enough to solve the problem of bad beliefs.

## 2. Rationality as luck

Levy provides a convincing and comprehensive account of how some of our strangest beliefs and practices can in fact be understood as rational. However, we worry that in managing to rationalise such practices, we start to lose out on a useful picture of *irrationality*. We'll look at two of Levy's examples to demonstrate this. Firstly, Levy discusses the Naskapi hunters who heat a caribou shoulder over coals until it cracks, and then decide where to hunt on the basis of how the pieces fall. Secondly, Indigenous American peoples who cooked corn with wood ash or ground sea shells or lime, and so subsequently the corn did not give them Pellagra, which was a disease affecting corn-eaters elsewhere. Both practices are just 'the done thing', with both populations having little grasp of the mechanisms by which they work. In the former case, the random nature of how the bone fragments fall ensures that the hunters don't fall prey to the tendency to get superstitious and see illusory patterns in which hunting spots are best. In the latter case, the added ingredients to the corn were alkalis which released the niacin in the corn, which in turn meant that the corn did not give people the Pellagra disease (caused by niacin deficiency).

Levy describes both of these practices as perfectly rational and reflective of the crucial role of culture in knowledge production. He emphasises the severe costs if individuals break away from doing 'the done thing' here and question the practices; they are less successful and risk illness or even death. Because the individual who breaks away from these practices and questions them risks so much and neglects such valuable social knowledge, it is a better epistemic position to be in to just go along with the practices even if the mechanism isn't understood. Levy describes these practices as therefore manifesting 'intelligence' (page 49). It is a special skill of humans to imitate every step of a routine shown to them even if some steps are clearly functionally redundant. Chimpanzees will not do this, skipping out the unnecessary steps. This is, in Levy's view, to their loss because it gets in the way of accumulating very valuable cultural knowledge over time and generations, which would be impossible for individuals working alone. However, he also says that "we owe our success to the fact that we are in some ways less—or at any rate less directly—rational animals than chimps." (page 45) This hints at the possibility of what is rational and what is 'successful' or 'intelligent' coming apart, but it is by these same processes that Levy goes on, throughout the book, to defend bad beliefs as rational.

Our worry is that these practices manifest intelligence from an external point of view, of mother nature, perhaps. They work, in the long run. But this doesn't tell us very much about people and rationality, with the latter turning into a matter of luck. Specifically, luck with regards to whether you are an Indiginous American with the custom of cooking corn with ash or shells or lime, or if you are based elsewhere and do not do this. Irrationality becomes no fault at all, but just 'wrong place wrong time'. At first this fits Levy's picture to some extent - people are not irrational, just their environments are unideal and either they have good customs or they don't - but Levy also allows that within the same culture, some cultural practices will be 'shallow' and require straightforward imitation, whereas others will be 'deep' and require *innovation* in order to achieve the valuable cultural knowledge accumulated over time.

Returning to the shells example, we want the Indiginous Americans not to question their corn-cooking practices, but for the Europeans to do so. How are we ever to know which position we are in? Levy describes the intelligence of the caribou-shoulder burning as the overriding of the human disposition to lose signal in noise by seeing illusory patterns, but the practice around corn-cooking would have *initially* been a pattern which could have been just as illusory as the superstitions which damage hunting prospects - because there is no understanding of the underlying mechanism to guide this decision-making process. These cooking practices clearly turn out to be worthwhile, but from the point of view of the human beings involved, it's a poorly understood ritual that they are 'falling prey to' in the same way that the hunters would be 'falling prey to' biases of superstition.

Medical professionals were in this position when investigating why rates of Pellagra were so high outside of the Indigenous American population, and asking questions about the mechanism (or, innovating) is what brought answers. This is where we want good old fashioned individual rationality to come in; in ascertaining when to question and innovate, and when to go with the flow and imitate faithfully. Investigating the underlying mechanism and ascertaining how 'illusory' the pattern really is will be crucial to this. Levy acknowledges that we do sometimes respond to hints in deciphering this; if the person we are imitating seems to be acting very intentionally we are less likely to innovate. If the person we are imitating seems to be distracted or getting around another problem (their hands are full, for example), we are more likely to innovate and not straightforwardly imitate.

But this is difficult to apply to helping us know when we are dealing with shallow or deep cultural knowledge. It is difficult to apply to cases of individuals faced with the question of whether climate change is real, or whether they should support figures like Donald Trump. Levy describes how in the latter case of Never Trumpers especially, it is the social outsourcing of beliefs and falling in line with what those around us believe because they are perceived to be ‘people like us’ (and may also be prestigious) which explains how the seemingly irrational change of opinion is ecologically rational. In other words, it’s a mechanism which usually gets us good and well-supported beliefs, but in these kooky epistemic environments they get us bad beliefs. We accept that these individuals often have good reasons for rejecting the mainstream view given their epistemic environment and peer group, but nevertheless the option of innovation surely doesn’t go away. Levy says that innovation is appropriate for *shallow* cultural knowledge, as opposed to deep cultural knowledge, but this merely pushes the question one step along to - how do we know whether an issue pertains to shallow or deep cultural knowledge? Given that Levy appears to allow for the issue of whether to support Trump, and other beliefs which *feel* personally deep, to in fact be shallow (page 65) because they are abandoned relatively quickly in response to social pressure, this is a sticky problem. But importantly it surely makes it possible that beliefs regarding the truth of climate change can also count as shallow and the option of innovation remains. We are suggesting that some epistemic responsibility may come into the picture at ascertaining this; whether we should innovate or imitate, even if perfectly rational processes can lead us astray once we pluck for one of these.

The next question is likely to be, what does ‘innovation’ look like in the face of considering whether to accept the truth of climate change? Levy warns against individualistic solutions; ‘doing your own research’ or having epistemic virtues. We look at each of these in turn.

We accept that innovation does not have to be an individualistic affair, and can instead be just as socially embedded as Levy’s picture of rational processes is. This is in contrast with his description of ‘doing your own research’ as being very individualistic. He describes agents as facing a choice of either shrugging their shoulders, or ‘doing their own research’ and digging into argumentation, when they come across surprising or bizarre conclusions, such as that climate change is not real. He suggests that they ought to shrug their shoulders and move on (page 94), given the risks which the individual incurs when engaging in questioning (such as in the cases of Indigenous Americans who would question cooking corn in their traditional way, and Naskapi hunters who would question why they use caribou shoulder fragments to pick where to hunt). However, we think this is too simplistic. Firstly, both these options are individualistic; shrugging shoulders or engaging and trying to tackle spurious arguments ourselves. But Levy criticises only the second for being individualistic. We also do not think this option has to be individualistic. It only looks this way when we needlessly limit the time span we are looking at, to the immediate aftermath of coming across a strange conclusion and/or argument. In reality, we think there is a path between the two options of shoulder shrugging and individual research. This is something like, holding the strange idea in the back of our minds, and seeing what happens in the near future. Do you notice other people mention it? Do other people ask you about it? Do you come across specific people you think could give really valuable insight? Does it pop up on twitter or in meetings? Perhaps you follow up on that when you might not have before. We hope it's clear to see the role that other people play here in tackling a surprising new idea. But it is not an attempt to, independently, master complex expert literature or 'science the shit' out of something, nor is it a passive shoulder shrug given that we already know what we and our peers think about some issue.

Zooming out, our picture is one of agents, over time, sometimes finding themselves alone with a new idea and perusing argumentation behind it, reflecting on how it strikes them, and being in moments where they have no choice but to be individualistic. But at others, they draw on the thoughts and ideas of others consciously or subconsciously, perhaps to then think about privately again later. In this cycle, there are individualistic moments which Levy captures but eschews, whereas we think they can still have a part to play in a broader process which draws on social influences at other times.

We think something similar is the case with Levy's account of epistemic virtues. He thinks that epistemic virtues are not as risky as doing your own research, but still worries that they are too individualistic to help - "they appear to aim to bring us each to inculcate the virtues in ourselves and then, guided by our intellectual excellences, to tackle hard problems largely on our own" (page 91). He would prefer something which better enables apt deference to others. However, we see a much closer link between epistemic virtues and exactly this - apt deference to others. Things like open-mindedness, humility, arrogance, sociability, can all have a significant role to play in ensuring apt deference to others. If we are arrogant, we are unlikely to take much of what anyone else says seriously. If we are humble, we are more likely to take seriously what we hear from others, and not just people who look like us or are familiar to us. If we are sociable, we're likely to be in more, and more intimate contact with a wider variety of other people from all walks of life, and therefore be more likely to come across lots of valuable information from them. In many ways, the virtue epistemologist has to battle the same problems as the virtue ethicist in accepting that unideal environments - particularly unideal social environments - do place individuals at significant disadvantages in their epistemic, or ethical, development. They both hope that agents will respectfully defer to others, have relevant trustworthy epistemic institutions - or moral role models - available to them, and have relevant educational experiences to learn from. They are both concerned not just with ways of analysing and interpreting first-order evidence (of what to believe or how exactly to act), but of being well disposed such that you're in a good position to defer aptly to others when needed and bring what is learnt there to bear on relevant situations.

So, we do not see 'doing your own research' or virtue epistemology as individually as Levy seems to. Although this lets into the picture all the ways that unideal social environments can create bad beliefs, much of Levy's description of which we are on board with, we still also think there are some opportunities here for slightly more independent choice and rationality to be exercised.

### 3. Beliefs, action and responsibility

Levy's take on rationality hinges on some underlying assumptions about how we form and sustain beliefs. Beliefs "happen" to people: the title itself implies that belief formation is not an active process of choosing the most rational option after analyzing the available first-order evidence, but it is a somewhat passive process that can be reliably predicted given certain environmental factors. If the prevalent sources of information that hold epistemic authority in my environment says that x, the belief that x will likely “happen” to me. Rationality does not need to involve active thinking or choosing, but imitating practices and deferring to one’s social environment. If the epistemic environment is polluted and filled with misinformation, the otherwise rational act of social deference will fail and conspiracy theories will proliferate.

This has important consequences on Levy’s proposed solution to the proliferation of conspiracy theories and bad beliefs. If the environment is the primary force determining beliefs, we can artificially generate better beliefs by cleaning up the epistemic environment: by making clear which are the mainstream, scientifically supported views and nudging people towards credible positions, we will counterbalance the rise of conspiracy theorists and produce more successful epistemic agents.

Levy does get at something here. The environment and the available evidence are certainly part of the reasons for epistemic success or failure. Someone growing up in a family of scientists is probably less likely to succumb to conspiracy theories than someone whose social bubble is made entirely of flat-earthers. Having credible information available and easily recognizable is an important prerequisite to form good beliefs, and our current epistemic environment is not ideal in this sense: unless one has learned a fairly complicated set of skills to help them recognize which sources to trust, the amount of contradicting and (at least at first sight) credible-looking information it is not easy to navigate.

However, talking about beliefs this way ignores another very important aspect of belief formation. We do not only form beliefs by passively absorbing information - we actively select and choose what or who we want to trust in based on our personality, epistemic strategies, identity and core beliefs. Making it obvious what views are mainstream and scientifically credited will only help people who have already made the choice of trusting mainstream, scientific views.

A devoted Christian will likely keep believing in creationism even if they end up in an epistemic environment where the most prevalent sources of information say otherwise, like a largely non-religious city or a science class. A left-leaning scientist will keep believing in climate change if they move to a very conservative town where everyone thinks green politics are a hoax. Both epistemic agents are perfectly capable of ignoring the mainstream position in their current environment in favour of a minority position. More importantly,  they do so in spite of all pointers of epistemic authority (it being taught at school, for example) because of a background choice about their values and who they are. They know who the epistemic authority is in their social environment; they just decide to reject it because they perceive it as conflicting with their values. In extreme cases, they might come to reject any view that comes to their attention that is labelled as epistemically authoritative, even if they know nothing about it, precisely because it is presented as the consensus in an epistemic environment they feel strongly averse to. This is the narrative of the system; I do not fit in the system; this is not my narrative.

This is in contrast to Levy's description of even deeply held (or so it seems) beliefs about the self being shallow and often being outsourced - as he suggests is the case with the 'Never Trumpers' who denounce Trump but then ended up supporting him, in line with the rest of the social group they saw themselves as being a part of. Instead of having robust inner models of ourselves and our beliefs, we off-load these onto the outer world and tend to more so just respond to triggers when the time comes. In the most stark example of this, choice blindness studies show individuals explaining their recently expressed belief or choice x when prompted to by researchers, even though they actually expressed belief or choice y. So, they used this prompt by researchers as a trigger which told them what they actually thought/believed, demonstrating how shallow and impoverished our inner models are.

However, an alternative interpretation of this is that we actually see quite how tightly people cling to inner models of themselves. Individuals react to these prompts in this way because of some of their other self-related beliefs. Particularly, for example, that they are consistent, that they know what they said a minute ago, that they are the ultimate authority on their own views. This is a strange scenario in which deeply held beliefs about one's own consistency does cause inconsistency, given that evidence to the contrary is being overlooked. But this is still in the light of enjoying a particular self-image.

So, for similar reasons, a Flat Earther will likely not abandon their convictions just because the epistemic environment gets depolluted and the mainstream view is clearly indicated. This is because this depolluting doesn't contribute anything to the self-relating beliefs which the agent practically enjoys having, as they speak to their being consistent but also interesting and contrasting. Even more than religion or science, many conspiracy theories thrive in an “us versus them” dynamic. Such polarisation cannot only be explained by looking at a polluted epistemic environment and the difficulty to tell the scientific information apart from the unfounded sources: it is, in the first place, an active choice to pick one side rather than the other. We will briefly talk about which factors might influence this choice, and the consequences for moral and epistemic responsibility.

We believe that this is an area where epistemic virtues can actually be helpful. In particular, for example, we believe that a common epistemic vice among conspiracy theorists is a need for uniqueness - a desire to place oneself above others in terms of one’s epistemic capacities and knowledge, to hold the essentially contrasting minority position and defend controversial views in order to appear different from the majority. This could be thought of as a sort of, propensity to be an ‘epistemic special snowflake’ and emphasises how something that looks quite a lot like a vicious trait, has knock-on effects for aptly deferring to good sources and paying them the attention they deserve.

Accepting some wacky beliefs that might contradict one another at a superficial level is a worthy sacrifice to maintain stability in one’s core beliefs at an inner level. People who feel rejected and alienated by the system are much more likely to develop conspiratory beliefs (Pierre 2020). Losing faith in the system develops into actively researching narratives outside of it, which further fosters the feeling of opposition: we are smarter than them, we know something that they are trying to hide. This opposition is not only epistemic, but is often also experienced as moral, social and political. One recent example are COVID deniers (Bisiada 2021): refusing to comply with governmental policies led to them being pointed to as responsible for the continuation of the pandemic by people who were practising social distancing, wearing masks and getting vaccinated. In turn, being painted as ignorant led to further grudge and distrust towards the official narrative from people who were sceptical, uninformed or just unable to deal with the social and economical consequences of social distancing.

The deeper issue with many instances of adherence to conspiracy theories is not to be found in misinformation itself, but in the reasons why misinformation is picked and believed. In all these cases, mainstream views are rejected not because it's hard to know which view is the epistemically authoritative view in an epistemically polluted environment, but precisely because of the signals of authority they display. If this is true, cleaning up the epistemic environment will not have the impact that Levy hopes for - highlighting with banners and pointers what the consensus of epistemically authoritative sources is will not make these people gain trust in those authorities, rather it will make it easier to recognise which positions to reject, because adopting them does not emphasise and enhance the individual's self-conceptions. Without deeper work on the socio-political environment to stop fuelling "us versus them" narratives, increase trust in science and prevent groups from feeling marginalized, efforts to contrast the rise of bad beliefs will be meaningless.

## Conclusion

Where does this leave rationality and responsibility? We have argued that social deference can be rational not in virtue of a passive mechanism of absorbing or repeating the prevalent practice or opinion in one's immediate environment, but in virtue of a powerful active component: one's capacity to adequately pick who to trust. This choice is strongly influenced by one's identity and can be led astray by epistemic vices like a need for uniqueness, and alternatively helped by epistemic virtues like humility, and therefore it is not free from epistemic responsibility. At the same time, Levy is right in arguing that to change people's beliefs, we need to start from changing their environment. Rather than a simple epistemic depolluting, though, we suggest that social changes are necessary in the first place to prevent people from developing anti-scientific identities and consequently picking anti-scientific beliefs.

## 4. References

Bisiada, M. (2021). Discursive structures and power relations in Covid-19 knowledge production. *Humanities and Social Sciences Communications*, *8*(1), 1-10.

Levy, N. (2021). *Bad Beliefs: Why They Happen to Good People*. Oxford University Press.

Pierre, J. M. (2020). Mistrust and misinformation: A two-component, socio-epistemic model of belief in conspiracy theories. *Journal of Social and Political Psychology*, *8*(2), 617-641.