Review of Philosophy and Psychology
https://doi.org/10.1007/s13164-021-00585-y

Introspection and Belief: Failures of Introspective Belief Formation

Chiara Caporuscio [1,2,3]

Accepted: 6 September 2021/Published online: 25 September 2021

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.

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

✉ Chiara Caporuscio
caporusciochiara1@gmail.com

[1] Otto-von-Guericke Universität Magdeburg, Magdeburg, Germany

[2] Research Training Group 2386 "Extrospection", Humboldt-Universität zu Berlin, Berlin, Germany

[3] Present address: Berlin School of Mind and Brain, Faculty of Philosophy, Humboldt-Universität zu Berlin, Berlin, Germany

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[1]:

> *"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,

[1] 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.

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).

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[2]. 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.

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

[2] 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)

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-V as "false beliefs based on incorrect inference about *external reality* [...]" (American Psychiatric Association 2013); 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 The Dissonance of Introspection

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.

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*"[3]. In judgments of this kind, the epistemic intersects with the phenomenal: demonstratives are not epistemically 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

[3] "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.

experiences, in this case by conveying that it was similar to perception and different from memory, imagination, and metaphorical thinking.

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**).**

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

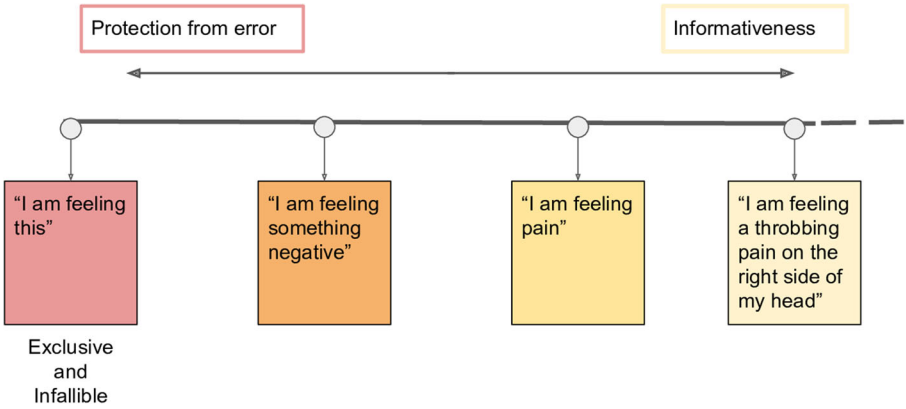

Fig. 1 The trade-off between informativeness and protection from error. This figure represents the trade-off between the two continuous properties that I believe to be at the center of the philosophical debate on introspection: informativeness and protection from error. "I am feeling this", "I am feeling 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"

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.

3 Informative Introspection and Belief Formation

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[4] 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 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 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).*

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

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

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.

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 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.*[5]

   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.*

[5] This is relatively common in psychiatric patients (Kind, ms)

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.

4 Objections and Future Directions

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.

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 Diagnostic and Statistical Manual of Mental Disorders, 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 2013). 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).

One reason to think that the external reality condition should be maintained is that delusions are thought to be pathological failures in belief formation[6], 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.

[6] 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.

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; 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.

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.

Acknowledgements I would like to thank Sascha Benjamin Fink, Joshua Martin, Emre Fatih, Barnaby Crook, the.
members of the RTG-2386 Extrospection and two anonymous reviewers for helpful feedback.
and discussion on the manuscript.

Availability of Data and Material Not applicable.

Code Availability Not applicable.

Funding Open Access funding enabled and organized by Projekt DEAL. German Research Foundation (DFG) – 337619223/RTG2386.

Declarations

Conflicts of Interest/Competing Interests The author has no relevant financial or non-financial interests to disclose.

References

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

American Psychiatric Association. 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., P. Delfabbro, C. Galletly, and T. Woodward. 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., A.A.A. Cárdenas, G.Z. Pazmiño, and P.A. Herrera. 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., L. Sass, and J. Parnas. 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 P.W. Halligan. 2020. Delusions and theories of belief. *Consciousness and Cognition* 81: 102935.

Connors, M.H., and P.W. Halligan. 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 P.W. Halligan. 1996. Cognitive neuropsychiatry [editorial]. *Cognitive Neuropsychiatry* 1: 1–3. https://doi.org/10.1080/135468096396659.

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

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

Davies, M., A. Davies, and M. Coltheart. 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., A.W. Young, A.H. Quayle, and K.W. De Pauw. 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., J. Kilner, and L. Harrison. 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., W. Muellbacher, and A. Nowak (1995) Imagery without perception—A case study of anosognosia for cortical blindness. *Neuropsychologia* 33 (11): 1373-1382https://doi.org/10.1016/0028-3932(95)00070-J.

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

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

Haselton, M.G., D. Nettle, and D.R. Murray. 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 S.A. Basha. 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 M. Coltheart. 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., M. Solms, and P. Sterzer. 2021. Useful misrepresentation: Perception as embodied proactive inference. *Trends in Neurosciences*, in press 44: 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 T. Wilson. 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 E.M. Hubbard. 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): 308-312https://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.