Self-Knowledge – A Defense of an Extrospective Account

By
Michael J. Roche

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Philosophy)

at the
UNIVERSITY OF WISCONSIN-MADISON
2013

Date of final oral examination: 12/17/13

The dissertation is approved by the following members of the Final Oral Committee:
Lawrence Shapiro, Professor, Philosophy
Sarah Paul, Assistant Professor, Philosophy
Alan Sidelle, Professor, Philosophy
Elliott Sober, Hans Reichenbach and William Vilas Professor, Philosophy
Charles Kalish, Professor, Educational Psychology

To the memory of my sister
Julie Ann Roche
whom I miss very much

Table of Contents

Acknowledgements

Earning a Ph.D. is quite the challenge. In my case, at least, it happened thanks to an abundance of help. I begin by acknowledging my family. My parents, Kent and Brenda, have always been extremely supportive, both emotionally and financially. I cannot overemphasize the role they have played in my successfully completing graduate school; with less supportive parents, I might very well not be writing this. I am forever grateful to them for their support. My brother, Bill, introduced me to philosophy and there is surely no person with whom I have talked more about philosophy than him. By far the majority of my work in philosophy—both in this dissertation and elsewhere—has been discussed at some point and to varying degrees with Bill. I thank him (and, at times, begrudge him!) for leading me down this path. I am extremely appreciative of his support and friendship.

Although my partner, Danielle, has been in my life for a much shorter time than my brother, she is rapidly gaining on him with respect to time discussing philosophy with me. I am certain that she has spent much more time than she would have liked talking with me about *my* work! Her willingness to spar with me philosophically has been a tremendous asset. She has also quite often talked me off of the proverbial ledge, offering me much needed encouragement about my philosophical work and abilities. Having her by my side the last six years has been invaluable, both professionally and, most importantly, personally. I thank her for this and very much look forward to our future together.

I also wish to thank my advisor, Larry Shapiro, for his guidance over the years. Larry has never turned down a request to read my work or to answer a question; at the

time I turned in this dissertation, he had already commented on earlier drafts of all but one of its seven main chapters. As my advisor, Larry has always made me feel as though I belong in this field, which I really appreciate. I should also note that his philosophical work has provided me with an excellent model for how to do and write philosophy (more on this below).

The remaining members of my dissertation committee, Alan Sidelle and Sarah Paul, also deserve thanks. Alan provided very helpful comments on an earlier version of chapters six and seven, and commented on the entire dissertation once it was completed. I really appreciate his willingness to provide such thorough and challenging feedback on my written work. Sarah's seminar on self-knowledge in the fall of 2011 gave me a tremendous confidence boost and, more importantly, established Sarah as someone I could go to for insightful feedback on my work. I have spent much time discussing self-knowledge with her over the last few years! Especially noteworthy is her assistance on chapters four, six, and seven. Of course neither she nor anyone else (aside from me) is responsible for any shortcomings or errors contained in the pages that follow.

Thanks as well to Elliott Sober who has had a tremendous influence on my philosophical work. Like Larry, Elliott has served as a model for how to do and write philosophy. I have spent a great deal of time reading and thinking about their work; undoubtedly, this has positively influenced my own work (although not as much as I would like!). I should also thank Elliott for the opportunity to be his project assistant in the spring of 2013. While this experience was rewarding in itself, it also gave me extra time to pursue my own work.

Brie Gertler also deserves thanks. Although I have met her only once (and only very briefly), she has been kind enough to read early drafts of much of the work in this dissertation (specifically, chapters four through seven). In addition, her work on self-knowledge has been tremendously valuable to my own understanding of the subject.

Both Chuck Kalish and Elliott Sober served as non-readers at my oral defense. They were my top choices and, not at all surprisingly, each significantly strengthened the discussion. I thank them for giving up some of their time to watch me squirm. I also thank Patty Winspur for helping me with the logistics surrounding both the submission of this dissertation and the oral defense. Moreover, I thank Patty, along with Lori Grant and Christy Horstmeyer, for being so friendly, encouraging, and helpful over the years.

Finally, I would like to thank the many friends and mentors that I have met over the course of my time in graduate school. I would especially like to single out Russ Shafer-Landau, Eric Margolis, Martha Gibson, Jesse Steinberg, Shannon Spaulding, Brynn Welch, Armin Schulz, Casey Helgeson, John Basl, Matt Kopec, Bo Kim Kopec, Michael Goldsby, Matt Barker, Jeff Behrends, Gina Schouten, Stewart Eskew, Eric Stencil, and Josh Filler. I will forever look back fondly on my years in Madison.

Chapter 1 - Introduction

1.1 The Topic of this Dissertation

We seem to enjoy a special kind of access to our own mental states. For one, I seem able to attain knowledge of my mental states that is epistemically superior to, and different in kind from, the knowledge that you are able to attain of my mental states. Regarding epistemic superiority, I seem able to attain knowledge of my own mental states that is, at the very least, better justified than the knowledge that you are able to attain of my mental states. Regarding difference in kind, this superior justification is not merely a matter of my knowledge having a greater degree of the same kind of justification as your knowledge; rather, I seem able to attain knowledge of my own mental states that has a different kind of grounding than the knowledge that you are able to attain of my mental states.

Moreover, I seem to have a method for coming to know about my own mental states that is not suited for coming to know about your mental states. While I am forced to infer the presence of such states in you by observing your behavior (verbal and otherwise), I seem able to know about these states in me via a uniquely first-personal method of some kind or another. To borrow Alex Byrne's (2005) terminology, we each seem to enjoy both 'privileged' and 'peculiar' access to our own mental states. These alleged features of self-knowledge are important in what follows.[1,2]

---

[1] Often, our access to our own mental states is characterized as simply being 'privileged'. But this is often taken to include both of the features just described, as well as others. Gertler (2003) points out that there are multiple senses of 'privileged access' that have occurred in the philosophical literature and refers the reader to Alston (1971) for a more detailed discussion of these various senses. For this reason, I appreciate Byrne's separating and clearly defining two features that one might have in mind when claiming that we enjoy special access to our own mental states.

Note, however, that no contemporary philosopher or psychologist thinks that we enjoy such access to *all* of our mental states. For example, one's character traits, assuming such traits exist,[3] are often cited as mental states to which one does not enjoy special access. Similarly, both deeply repressed Freudian mental states and sub-personal mental states (e.g., some of those involved in visual processing) are not thought to be accessible in privileged and peculiar ways. There are thus limits on privileged and peculiar access.

While there are notable exceptions, philosophers have traditionally assumed that the intuitive picture just sketched is correct. That is, they have traditionally assumed that we enjoy privileged and peculiar access to a significant portion of our mental lives.[4] The challenge is to explain *how* such access is possible.

Traditionally, attempts to meet this challenge have produced accounts on which privileged and peculiar self-knowledge is attained via an *inwardly directed* method. Both the 'inner sense theory' and the 'acquaintance theory' are part of this tradition. According to the former, we enjoy a perception-like access to our own mental states; on some versions of the theory, we are equipped with a perception-like mechanism whose function it is to detect our own mental states.[5] According to the latter, we are able to bear a non-

---

[2] I have slightly tweaked Byrne's usage of these terms, but I think I have done so for the better. For one, he does not characterize privileged access as involving a difference in kind. But I take it to be somewhat obvious that this is a mistake. Consider an individual, S, whose self-knowledge is attained exclusively via inference from observational evidence concerning S's behavior. Although S has more evidence of this kind than anyone else, we should not want to say that S's knowledge is thereby privileged.

[3] Gilbert Harman has argued that there is no evidence that people have character traits; he claims that we think otherwise due to our mistakenly downplaying the influence of situational factors on our behavior. See Harman (1999 and 2009).

[4] Gilbert Ryle (1949) is one such exception. See Peter Carruthers (2009, 2011) for an up-to-date defense of this skeptical position, one that draws heavily on recent findings in cognitive science.

[5] Philosophers who have recently defended the inner sense theory include David Armstrong (1968, 1981), Alvin Goldman (1993, 2006), William Lycan (1987, 1996), and Shaun Nichols and Stephen Stich (2003). John Locke (1690/1975) is often cited as an early defender of this theory.

causal, metaphysically direct relation to some of our own mental states; this metaphysically direct relation is referred to as 'acquaintance'.[6] On both accounts, one attains privileged and peculiar self-knowledge by (in some sense) directing one's attention "inward", towards one's mental states. I shall call this approach to self-knowledge an 'introspective approach', and I shall call accounts in this tradition 'introspective accounts'.

In contrast to this approach, some philosophers have proposed accounts on which privileged and peculiar self-knowledge is attained via an *outwardly directed* method. This approach takes as its inspiration the following remark from Gareth Evans (commenting on a remark from Wittgenstein):

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward – upon the world. If someone asks me 'Do you think there is going to be a third world war?,' I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' (Evans 1982, 225).

Self-knowledge, on this approach, is attained by "looking through" one's mental states, attending to their contents, i.e., that which they are about (e.g., a third world war).[7]

For accounts in this tradition, we attain privileged and peculiar self-knowledge via the cognitive mechanisms and processes that allow us to think about the (non-mental) world; no inwardly directed mechanisms or metaphysically special relations need be posited. This is an important feature of such accounts, and is one that many find quite

---

[6] Philosophers who have recently defended the acquaintance theory include Laurence BonJour (2003), David Chalmers (2003), Richard Fumerton (1995), and Brie Gertler (2001, 2012). Bertrand Russell (1912) was an early proponent of this theory.

[7] Given this metaphorical "looking through", such accounts are often referred to as 'transparency accounts' of self-knowledge. Note, however, that Carruthers (2011) misleadingly uses this label to refer to *any* account according to which we have special, non-interpretive access to our own mental states. Partly for this reason, I have opted for the 'extrospective' label.

attractive. I shall call this approach to self-knowledge an 'extrospective approach', and I shall call accounts in this tradition 'extrospective accounts'.[8]

1.2 The Thesis and Plan

In this dissertation I limit my focus to a particular class of mental states, namely, the propositional attitudes. This class includes, most prominently, beliefs, desires, and intentions. As suggested by the name, a subject has a particular propositional attitude at a time by virtue of bearing a certain kind of attitude towards a given proposition. The type of attitude one bears towards that proposition determines the type of propositional attitude one has at that time (e.g., belief, desire, intention, etc.). Note that one can bear various attitudes towards the same proposition. I say more about the propositional attitudes in Section 2.3.

I believe that we enjoy privileged and peculiar access to our propositional attitudes. My project in this dissertation is to defend an extrospective account of this access. Much of my focus on this front is on our knowledge of our beliefs. Interestingly, I think, defending an extrospective account of our privileged and peculiar access to our beliefs is not as straightforward as has been previously supposed. My story for how we attain such self-knowledge is thus quite complex. Fortunately, though, I think the complexity pays off. The account I construct for belief can be extended to the other propositional attitudes in a fairly straightforward manner.

Part I of the dissertation concerns the two aforementioned introspective accounts of self-knowledge. The inner sense theory is the focus of chapters two through four.

---

[8] Philosophers who have recently defended extrospective accounts of self-knowledge include Alex Byrne (2005, 2011), Fred Dretske (1994, 1995), Jordi Fernandez (2003, 2007, 2013), Robert Gordon (2007), Richard Moran (2001), and Michael Tye (1995, 2003).

Although I reject this theory, my aim in these chapters is not to argue against it. Instead, my aim is to explore the commitments and consequences of the theory. After offering a fairly brief description of the inner sense theory in Chapter 2, I argue in Chapter 3 that the theory is committed to a previously unrecognized claim about the neurological realization of the propositional attitudes. Uncovering this commitment promises to advance the study of the inner sense theory.

In Chapter 4 I examine a particular kind of argument that is used by both proponent and opponents of the inner sense theory. This argument is premised on the claim that if the inner sense theory is true, there should exist individuals with various kinds of first-personal deficits. Proponents of the theory focus on deficits that they claim are exhibited by certain individuals in the population (e.g., schizophrenics with passivity experiences) and conclude that the presence of such individuals confirms the theory. Opponents of the theory focus on deficits that they claim are not exhibited by individuals in the population and conclude that the absence of such individuals disconfirms the theory. I argue that the method subserving confabulation significantly complicates the predictive step of this two-step argument and undermines particular instances of it.

I close Part I with a discussion of the acquaintance theory. I think this theory is fairly easily shown to be inadequate as an account of how we know our propositional attitudes in a privileged and peculiar way; I make this case in Section 5.7. In the remaining sections of this chapter my aim is to explain and raise various criticisms of the acquaintance theory. My discussion examines Brie Gertler's (2011a) presentation of the view.

Part II of the dissertation concerns the extrospective approach to self-knowledge. In Chapter 6 I give a brief description of this approach to self-knowledge and the so-called 'paradox of transparency'. I also describe in detail Alex Byrne's (2005) extrospective account of our privileged and peculiar access to our beliefs. While I am sympathetic to Byrne's view, I argue that it rests on a contentious assumption concerning the relationship between judgment and belief. In response to this discussion, I develop and defend in Section 6.6 my own extrospective account that does away with this assumption. This account is the centerpiece of Part II of the dissertation and is further discussed and expanded on in its remaining chapters. In this chapter I also offer a brief remark concerning the epistemology of the extrospective approach to self-knowledge.

Chapter 7 focuses on a particular objection to the extrospective approach to self-knowledge due to Brie Gertler (2011b). This objection is particularly aimed at Byrne's account, but also applies to my own. In addition, it places a constraint on how I am able to defend a particular component of my account. While I defend my account from this objection, the main task of this chapter is to defend this component in a way that is consistent with the aforementioned constraint.

My account developed and defended in chapters six and seven concerns knowledge of belief. In Chapter 8 I extend this account to propositional attitudes other than belief, namely, desire and intention. Finally, in Chapter 9, I take stock. Specifically, I consider the question of whether my extrospective account of privileged and peculiar access to our propositional attitudes should be preferred over the inner sense theory examined in Part I. I argue that, although the case is inconclusive, my extrospective account should be preferred over that theory.

I. The Introspective Approach

Chapter 2: What the Inner Sense Theory Is

2.1 Introduction

In this chapter I explain the inner sense theory of introspection. The chapter is relatively short due, in part, to the fact that the theory itself is quite simple. This is not, however, to say that the issues surrounding the theory are quite simple. They are not. Many of these issues are touched upon in chapters three and four. These three chapters collectively should provide a fairly comprehensive understanding of the inner sense theory. The purpose of this chapter is to present the essentials of the theory (2.2) and to provide a concrete example of a contemporary version of the theory (2.3).

2.2 The Perceptual Analogy

According to the inner sense theory, humans have a special method for detecting their own mental states. This method is contrasted with the method that we use to attribute mental states to others. While the latter is, at the very least, indirect and highly interpretive, relying on the gathering of data concerning another's behavior (verbal or otherwise), the former is not thought to be interpretive in this way. According to the inner sense theory, we are able to know our own mental states via a method that is direct and quasi-perceptual. David Armstrong, a prominent defender of the inner sense theory, writes that introspection by inner sense "is a perception-like awareness of current states and activities in our own mind" (1981, 61). In addition to Armstrong, proponents of the inner sense theory include William Lycan (1987, 1996), Shaun Nichols and Stephen Stich

(2003), and Alvin Goldman (1993, 2006); John Locke (1690/1975) is often cited as an early defender of this theory.[9]

The inner sense theory thus likens self-knowledge to perceptual knowledge. Just as perceptual knowledge, although fallible, is more secure than knowledge arrived at via inference, the inner sense theory claims that self-knowledge by inner sense is more secure than our inferential knowledge of the minds of others. Moreover, that this knowledge is perceptual shows that self-knowledge via inner sense has a different kind of grounding than knowledge of others' mental states. Given these features of the theory, it seems capable of explaining the alleged privileged access that we have to our own minds. Because this alleged perception-like method cannot be used to attain knowledge of others' mental states, the theory also seems capable of explaining the alleged peculiar access that we have to our minds.

Note, however, that, at least according to Armstrong, this peculiarity is only contingent. Interestingly, although I think appropriately, he writes the following:

> In introspection we have direct, noninferential, awareness of our own mental states. We have no such direct, noninferential, awareness of the mental states of other people. It is, however, perfectly conceivable that we should have direct access of the mental states of others. In Materialist terms, we have scanners that can scan some of our own inner states, but no scanners that can scan the inner states of others. However, I take the claim that telepathic knowledge exists to be the claim that we do in fact have some direct awareness of the mental states of others (1968/1993, 124).

[9] Importantly, I am *not* in this paper concerned with the inner sense theory understood as a theory of consciousness. Rather, I am concerned with it only as a theory of how we come to know (or have beliefs about) our own mental states. Nichols and Stich explicitly deny that their inner sense account is intended to explain consciousness. Goldman similarly avoids issues of consciousness. Although Lycan uses his account of inner sense to explain some aspects of consciousness, he also intends for it to explain the much more mundane phenomenon just described.

Lycan seems to agree with Armstrong on this point, writing that "… [b]ut neither you nor I could have this functionally direct access to someone else's mental states (except by some futuristically special rewiring)" (1996, 49).

Likening inner sense to perception, and thus likening self-knowledge to perceptual knowledge is fine, but what more can be said about the inner sense theory? Sydney Shoemaker (1994a), an ardent critic of the theory, helpfully identifies eight features of ordinary sense perception. An inner sense account need not satisfy each of these features in order to warrant the analogy with sense perception. But Shoemaker claims, and inner sense theorists appear to agree, that satisfaction of the following two features of ordinary sense perception are essential to the analogy:

> i. Causal Feature: "perceptual beliefs are causally produced by the objects or states of affairs perceived, via a causal mechanism that normally produces beliefs that are true" (253).
>
> ii. Independence Feature: "the objects and states of affairs which the perception is of, and which it provides knowledge about, exist independently of the perceiving of them" (254).

Formulating these two features in terms of self-knowledge will be useful in later sections:

> i'. Causal Feature$_{s\text{-}k}$: introspective beliefs are causally produced by the mental states perceived, via a causal mechanism that normally produces beliefs that are true.
>
> ii'. Independence Feature$_{s\text{-}k}$: the mental states of which the perception is of, and which it provides knowledge about, exist independently of the perceiving of them.

Just as my perceptual belief that there is a tree before me is caused by that very tree, via a causal mechanism that normally produces true beliefs, so too, according to the inner sense theory, is my privileged and peculiar belief that I intend to go to the store caused by my intention, via a causal mechanism that normally produces true beliefs. And just as my perceptual belief that there is a tree before is independent of that very tree (the tree's existence does not depend on my having any belief about it), so too, according to the inner sense theory, is my privileged and peculiar belief that I intend to go to the store independent of my intention to go to the store.

Shoemaker, in addition to requiring that an account of inner sense possess these two features, also takes their presence to be sufficient for an account's being one of inner sense. An account that adopts just these two features of ordinary sense perception, he calls a 'broad perceptual account'; in contrast, he calls an account that adopts these two features plus others an 'object-perceptual account'.[10] While I agree that these features are necessary for an account of inner sense, I deny that they are jointly sufficient. I will say more about this later on in Section 3.4.

Among the features of ordinary sense perception that Shoemaker regards as inessential to the inner sense theory, there are two that he claims the inner sense theory rejects. These are:

> iii. Sense perception involves the operation of an organ of perception whose disposition is to some extent under the voluntary control of the subject" (204-05).

[10] Shoemaker's arguments against the object perceptual model of inner sense differ from his arguments against the broad perceptual model of inner sense. He argues against the former in his 1994a and against the latter in his 1988 and 1994b.

> iv. Sense perception involves the occurrence of sense-experiences, or sense impressions, that are distinct from the object of perception, and also distinct from the perceptual belief (if any) that is formed (205).

Armstrong (1968/1993, 1981) explicitly denies the existence of an organ of inner sense. He points out, however, that inner sense has this in common with proprioception, i.e., our perception-like awareness of the state of our body, thereby undermining the alleged disanalogy.

To see the idea behind the rejection of (iv), consider one's awareness of a pain sensation in one's knee. This introspective awareness appears to involve exactly one sensation, namely, the pain sensation; there does not appear to be an additional sensation of that pain sensation. I find this plausible, but cannot say anything more on its behalf. At any rate, Shoemaker is certainly correct that inner sense theorists do not adopt this feature of ordinary sense perception.[11]

Next, I would like to describe a particular version of the inner sense theory, namely, Nichols and Stich's (2003) Monitoring Mechanism account. Doing so will provide a concrete example of a contemporary version of the inner sense theory. Note, however, that it is just one such account. I will describe Goldman's (2006) inner sense account later on in sections 3.3 and 4.2. While his account is quite similar to Nichols and Stich's, there are some differences worth noting.

---

[11] The remaining features of ordinary sense perception identified by Shoemaker are: (v) "sense perception provides one with awareness of facts … by means of awareness of objects"; (vi) "[s]ense perception affords 'identification information' about the object of perception"; (vii) "[t]he perception of objects standardly involves perception of their intrinsic (non-relational) properties"; and (viii) "[o]bjects of perception are potential objects of attention" (1994a, 205-206).

2.3 Nichols and Stich's Monitoring Mechanism Account

Nichols and Stich's account, although not limited to the detection of the propositional attitudes, is usefully presented with respect to such states. Thus, before turning to their account, I would like to say a bit about the nature of the propositional attitudes, and, relatedly, what the detection of these attitudes involves; I call the detection of propositional attitudes 'attitudinal detection'. This background will make clearer both the presentation of their account and my arguments in Section 3.2.

2.3.1 Propositional Attitudes and Attitudinal Detection[12]

There are at least two components of a propositional attitude: a propositional c*ontent* and an *attitude* taken towards that content. Beginning with the former, Nichols and Stich assume that the mind possesses a representational medium. A representational medium allows for the realization of contentful mental states. On this assumption, roughly, a given propositional content is tokened in a thought process if and only if a state with that very same content is tokened via the mind's representational medium. The nature of this medium is irrelevant for current purposes.[13] All that is being assumed is that the mind has a representational medium, and is thereby capable of realizing contentful/representational states.

Turning next to the attitudinal component of a propositional attitude, an agent can take various attitudes towards a given propositional content. For example, one might

---

[12] See Fodor (1987) and Rey (1997) (especially Chapter 1 and Chapter 8) for a more detailed discussion of the propositional attitudes. Each has influenced what follows.

[13] There are a variety of ways by which the mind might realize contentful mental states. For example, following Fodor (1975), the representational medium could be language-like, and thus, if true, cognition would take place in something like a language of thought. In this case, a given content would be tokened in one's thought processes via the tokening of a specific sentence in *mentalese*. Alternatively, the representational medium could be pictorial, and thus, if true, a given content would be tokened via a non-linguistic, pictorial tokening. And of course the nature of the representational medium might be something else entirely.

believe some propositional content, or one might instead *desire*, *fear,* or *hope* for that content. Importantly, talk of an agent "taking an attitude towards" or, alternatively, "bearing a certain relation to" or "having in his Belief Box" a propositional content is merely shorthand for the idea that the corresponding representational state plays a certain causal/functional role within that agent's cognitive system.

Thus, if a state with the content *Obama is president* enters into the causal/functional relations characteristic of belief, then that state is a belief, and we can say that the agent "takes the belief attitude towards", "bears the belief relation to", or "has in his Belief Box" the propositional content represented by that state, namely *Obama is president*. If, on the other hand, that state enters into the causal/functional relations characteristic of desire, then it is a desire. Finally, notice that this is *not* to assume functionalism about the mind, but only functionalism about that which determines attitude type. I take this weaker claim to be quite uncontroversial among philosophers of psychology; I assume Nichols and Stich do as well.

This brief characterization of the propositional attitudes is relevant to understanding what is involved in attitudinal detection. The detection of any mental state should lead to the formation of a belief that one is in that mental state. In the case of attitudinal detection, as demanded by the nature of the propositional attitudes, the belief must specify the mental state's propositional content *and* the type of attitude taken towards that content. Attitudinal detection would thus seem to involve at least two tasks: (1) the detection of content and (2) the detection of attitude type. I refer to these tasks, respectively, as the 'content task' and the 'typing task'. This second task will feature prominently in my argument in Section 3.2.

2.3.2 The Account

According to Nichols and Stich, a relatively simple mechanism suffices to explain the seemingly special, direct access that we have to our own mental states. They offer the following succinct description:

> To have beliefs about one's own beliefs, all that is required is that there be a Monitoring Mechanism (MM) that, when activated, takes the representation *p* in the Belief Box as input and produces the representation *I believe that p* as output. This mechanism would be trivial to implement. To produce representations of one's own beliefs, the Monitoring Mechanism merely has to copy representations from the Belief Box, embed the copies in a representation schema of the form: *I believe that* ____, and then place the new representation back in the Belief Box. The proposed mechanism (or perhaps a distinct but entirely parallel mechanism) would work in much the same way to produce representations of one's own desires, intentions, and imaginings (160-1).

The content task, on their account, is handled rather simply. Their proposed mechanism simply copies the content of a propositional attitude and then embeds the copied content within an appropriate representation schema. Put in terms of input and output, the way in which the content of the input state appears in the output representation is completely straightforward: the content is simply 're-deployed'.[14]

The typing task, on the other hand, is not so simple. According to the quoted passage, executing the typing task requires selecting the appropriate representation schema within which to embed the copied content. If the targeted mental state is a belief, then the copied content should be embedded in the '*I believe that* ___' schema. If it is a desire, then the copied content should be embedded in the '*I desire that* ___' schema.

[14] Redeployment accounts of self-knowledge are quite popular. As already noted, Goldman (2006) develops and defends his own inner sense account. Although his account differs from Nichols and Stich's in various ways, both accounts claim that the content task in handled by re-deployment. Arguably, some who reject the inner sense theory also endorse re-deployment. See, e.g., Dretske (1995), Peacocke (1999), and Gordon (1995, 2007). Goldman (2006, 238-42) discusses this issue at some length. Note, however, that redeployment cannot be the whole story. Because the higher-order belief must always be a *belief*, the attitude type of a targeted state cannot be redeployed.

And so on for the other attitude types. Frustratingly, Nichols and Stich are silent about how the typing task is carried out on their account.[15] I will investigate this issue in depth, however, in the following chapter. Despite this gap in their explanation, see Figure 1 (on p 186) for a sketch of how Nichols and Stich's Monitoring Mechanism is supposed to handle the detection of a desire.

Although the above quoted passage concerns only the detection of the propositional attitudes, Nichols and Stich go on to offer a parallel account for the detection of perceptual experiences. In fact, their account can arguably be extended (with some adjustments) to accommodate any mental state type that has a representational content. The content task would be handled via redeployment, while the typing task would be handled in whatever way it is handled in attitudinal detection. Nichols and Stich leave open whether there is single monitoring mechanism equipped with representation schemas for all mental state types, or multiple monitoring mechanisms, each specialized for a certain type of mental state.

Notice, however, that extending the account beyond the propositional attitudes complicates each of these tasks. First, some representations seem to be such that they can be re-deployed in a (higher-order) belief only after being *translated* into a form suitable for belief. The representational content of a given visual experience, for example, may not be suited for belief; if not, then that content would need to be translated into a medium appropriate for belief. Second, because there would be a much larger class of mental state types to choose from (including pains, itches, visual experiences, auditory

[15] It might seem that this can be easily addressed. However, I argue below that there are significant difficulties with explaining how the Monitoring Mechanism detects which "box" a given propositional attitude resides in.

experiences, etc.), selecting the appropriate representation schema within which to embed a copied content would become more complicated.

Exactly how much these considerations complicate the story, I do not know. But that they *do* complicate the story is, I think, clear. From this point on, I will have this expanded version in mind when discussing Nichols and Stich's Monitoring Mechanism account.

2.4 Conclusion

In this chapter I have explained the basics of the inner sense theory and have provided a concrete example of a contemporary version of this theory. With this foundation in place, I explore the theory in more detail in the following two chapters. More specifically, I expose a previously unrecognized commitment of the theory (Chapter 3) and I argue that there is a significant difficulty with a particular way of arguing for (or against) the theory (Chapter 4).

Chapter 3: The Inner Sense Theory's Psycho-Neural Commitment

3.1 Introduction

In this chapter I argue that the nature of the propositional attitudes is such that the inner sense theory can explain their detection only if a specific claim concerning their neurological realization is true. This psycho-neural claim is, as far as I can tell, completely absent from the literature on the inner sense theory. My arguments thus draw out a previously unrecognized empirical commitment of the inner sense theory. Doing so promises to make more tractable the empirical investigation of that theory.

I first make my case with respect to Nichols and Stich's version of the inner sense theory, namely, their Monitoring Mechanism account. This takes up the bulk of the chapter (3.2-3.3). Having established my conclusion with respect to this account, I then argue that it extends to the inner sense theory more generally (3.4). This then leads to a discussion of the significance of my conclusion (3.5).

3.2 The Typing Task and the Need for [CORRELATION]

Recall that redeployment cannot provide the needed explanation. While the content of a mental state targeted by inner sense must be included in the resulting higher-order belief (e.g., a true belief that S desires that *p* must involves the content of S's first-order state, namely, *p*), the same is not true of attitude type. No matter what the attitude type of a mental state targeted by inner sense, the resulting higher-order belief must always be a *belief*. The attitude type of a targeted mental state thus cannot be re-deployed. To explain the execution of the typing task something more is needed.

The requisite supplementation might appear to not be far off. Specifically, the Monitoring Mechanism must simply be given the capacity to detect the box in which a given propositional attitude resides. If the Monitoring Mechanism detects that a propositional attitude with the content *p* is from (or in) the Belief Box, then it "knows" to embed the copied content into an 'I believe that _____' representation schema. If it instead detects that it is from the Desire Box, then it "knows" to embed the copied content into an 'I desire that _____' representation schema. Problem solved!

But of course this is a mistake. Recall from Section 2.3.1 that this talk of 'attitude type boxes' is metaphorical. Obviously nothing resembling a Desire Box can literally be found inside the head. Desires are likely not stored in a common area, let alone a bordered one. Rather, talk of a "Desire Box" refers to nothing more than a set of representations each of which plays the causal/functional role characteristic of desire; and likewise, mutatis mutandis, for the "boxes" of the other attitude types. Consider the following from Nichols and Stich (2003), commenting on their boxology:

> [P]ositing a 'box' which represents a functionally characterized processing mechanism or a *functionally characterized set of mental states* does not commit a theorist to the claim that the mechanism or the states are spatially localized in the brain (11, emphasis added).

Appreciating this rather mundane fact, though, has significant consequences. The Monitoring Mechanism is put forth as an explanation of attitudinal detection, and, as such, it must detect both the content *and* attitude type of a given propositional attitude. But if there are literally no attitude type boxes, or localized areas containing all and only representations of a certain attitude type, then it is unclear how Nichols and Stich's

mechanism is supposed to execute the typing task. An explanation is surely needed. Unfortunately, and as already noted, Nichols and Stich are silent on this issue.

Because causal/functional roles, not boxes, are that which determine attitude types, the Monitoring Mechanism would seem to require the capacity to detect the causal/functional roles of the mental states it targets. In the absence of such a capacity, the execution of the typing task would seem to be a matter of pure guesswork. Notice, however, that a mental state's causal/functional role is a relational property of that state. A mental state's causal/functional role is a function of the causal relations that it bears to other mental states, the causal relations that it bears to various inputs and outputs of the cognitive system, and the various cognitive processes in which it participates.

Importantly, if an object has such a relational property, this property cannot be directly detected via inspection of just that object; if one is inspecting just that object, the relational property can be detected only indirectly via the direct detection of certain of the object's non-relational properties. Or at least this seems obvious to me. An example will help to clarify this distinction between direct and indirect detection.

John is married and wears exactly one ring. I meet John, notice (through vision) that he is wearing a ring, and register him as a ring-wearer. John's non-relational property of being a ring-wearer[16] is responsible for my registering him as such; if John were not wearing the ring, but all else was the same, I would not have registered him as a ring-wearer. In this sense, my detection of John's being a ring-wearer is *direct*.

[16] If there is a worry about whether 'being a ring-wearer' is a non-relational property, John's wearing a ring can be replaced with John's having the message 'I am married' tattooed on his forehead. (And we can imagine that such tattoos are how John's society signifies marriage). I could then detect that he is married via the direct detection of this tattoo.

Suppose further that John's ring is on his ring finger. Given that I have certain beliefs concerning rings and marriage, I also register that John has the relational property of being married. Unlike the previous case, however, this relational property of John is not responsible for my registering him as being married; if John were not married, but all else was the same (including his wearing a ring on his ring finger), I would still have registered him as being married. In this sense, my detection of John's being married is indirect; it proceeds via my direct detection of his ring. Similarly, while I can directly detect that the piece of paper in my wallet is green, has the numeral '1' printed on it, etc., I can only indirectly detect that it is a piece of currency worth one hundred pennies.[17]

As is the case with being married, or being a piece of currency, so is the case with playing a certain causal/functional role. A psychological mechanism that takes as input mental state, *m,* can detect *m*'s causal/functional role only by directly detecting a distinct, non-relational property of *m* that indicates that *m* plays that role. The fate of Nichols and Stich's Monitoring Mechanism thus rests on its being able to indirectly detect the causal/functional roles of its inputs. That is, the account requires something like the following:

> [Correlation] For each causal/functional role definitive of a given attitude type, there exists a directly detectable property had by all and only mental states that play that causal/functional role.

While this requirement is essentially correct, there are two ways in which it is unnecessarily strong.

---

[17] Boghossian (1989) makes a similar distinction when discussing self-knowledge. His focus, however, is on mental *content* (construed as being determined relationally), not attitude type. I assume here that the inner sense theory has no problem explaining how mental content is detected. I will comment on Boghossian's objection again in section 3.3.2.

First, there need not be just a single directly detectable property correlated with each causal/functional role; a 1:1 correlation is not needed in order for Nichols and Stich's Monitoring Mechanism to indirectly detect these roles. Imagine, for example, that all and only mental states that play the causal/functional role definitive of belief have either of two directly detectable properties, *P* and *Q*. In this case, the detection of either property would indicate that the mental state is a belief. And of course we could imagine that all and only mental states that play the causal/functional role definitive of belief have either of *three* directly detectable properties, *P*, *Q*, and R. Any number of such properties will work.

Second, Nichols and Stich's Monitoring Mechanism does not aspire for infallibility.[18] For this reason, the indirect detection that is required for the execution of the typing task need not be perfect. Their mechanism, when properly functioning, should only be required to be highly reliable.[19] This level of accuracy, though, can be achieved even if the requisite correlations are imperfect. Assuming the correlations are steady enough, the Monitoring Mechanism could exploit them in order to reliably self-attribute propositional attitudes.

In light of these points, the above claim should be modified as follows:

> [CORRELATION] For each causal/functional role definitive of a given attitude type, there exists a directly detectable property (*or set of*

[18] Nor does the inner sense theory in general aspire for infallibility. By likening introspection to perception, and self-knowledge to perceptual knowledge, the theory welcomes fallibility.

[19] Nichols and Stich's Monitoring Mechanism is a psychological mechanism and so, like any psychological mechanism, can be damaged. Of course, if damaged, it should not be expected to reliably detect one's propositional attitudes. Damage to inner sense will be discussed in the following chapter.

> *properties*) had by *mostly* all and *mostly* only mental states that play that causal/functional role.

I submit that Nichols and Stich's Monitoring Mechanism can successfully execute the typing task only if [CORRELATION] is true. The Monitoring Mechanism will be able to successfully execute the typing task only if the propositional attitudes have directly detectable properties that fit the profile just articulated.

3.2.1 Goldman's Argument and the Multiple Monitoring Mechanisms Response

Before moving on, I would like to comment on an argument put forth by Alvin Goldman (2006) concerning Nichols and Stich's account. The reason for this is twofold. First, I would like to distinguish his argument from my own, for they might appear to be quite similar to one another. Second, this discussion will allow me the opportunity to respond to an important objection to my argument.

Consider the following from Goldman, commenting on Nichols and Stich's account:

> A monitoring mechanism might simply keep track of which box the [input representation] was taken from. But on the standard philosophical use of box symbolism, boxes are merely convenient shorthand for functional roles ... The question therefore arises how the monitoring mechanism determines that a given [representation] occurring in the mind has this or that functional role. We are told nothing about this (2006, 239).

This no doubt sounds quite similar to my own discussion of Nichols and Stich's account offered above. However, despite the seeming similarity, continuing on with this passage makes clear that Goldman and I have something different in mind:

> The problem can be *avoided* by positing a separate monitoring mechanism for each attitude type. If each mechanism is specialized for its proprietary attitude type, it knows what prefix to assign before putting an attributed syntax into the belief box (ibid, emphasis added).

Here, Goldman appears to acknowledge that multiple monitoring mechanisms would solve the problem that he has identified. This is significant because Nichols and Stich explicitly allow for the possibility of multiple monitoring mechanisms (2003, 161). Goldman's criticism of their account thus rests on ruling out this possibility. To this end, he appeals to parsimony as a reason to doubt that there are multiple monitoring mechanisms (2006, 239).

Importantly, my argument that Nichols and Stich's account requires [CORRELATION] is not threatened by the presence of multiple monitoring mechanisms. The challenge, recall, is to explain how a mechanism is to detect a mental state's causal/functional role, given that this property is relational. If there is a single mechanism, then this challenge applies only to it. But if there are multiple mechanisms, then the challenge simply applies to each mechanism. Multiplying the number of mechanisms only multiples the number of entities to which the challenge applies.

Further, claiming that each mechanism is "specialized for its proprietary attitude type" does not help. Saying this is simply to say that each such mechanism is somehow able to detect the causal/functional role characteristic of its proprietary attitude type. But an explanation of how this specialization is achieved is precisely what is needed!

Perhaps what Goldman has in mind when considering the possibility of multiple monitoring mechanisms is that each such mechanism would receive as input only mental states of a certain attitude type. If so, then there would be no question of whether a given input is a belief, or desire, or intention, etc. Any such mechanism would have only a single representation schema, namely, 'I X that ____', where X is whatever attitude type

the mechanism is specialized for ('believe', 'desire', 'intend', etc.). The problem with this suggestion is that it either collapses into the already rejected idea that mental states of a certain attitude type reside together in the same area of the brain, or presupposes [CORRELATION], and so does not avoid the challenge posed in the previous section.

To see why, consider the possibility that there are multiple monitoring mechanisms, each of which receives as input only mental states of a single attitude type. The Belief Monitoring Mechanism receives as input only beliefs, the Desire Monitoring Mechanism receives as input only desires, and so on. How could such an arrangement come about? Perhaps each such mechanism is located near only propositional attitudes of the relevant type. For example, the Belief Monitoring Mechanism is located near only beliefs, and, for this reason, never encounters non-beliefs.[20] But of course this suggestion is implausible. It assumes that propositional attitudes of a given type are located in a specific area of the brain, separated from propositional attitudes of other types.

A more promising explanation of the arrangement under consideration is that each such mechanism is "built" in such a way that it can process only propositional attitudes of its proprietary type. So although each mechanism has access in some sense to each of the various types of propositional attitudes, it can take as input only those with the right attitude type.[21] But this suggestion seems to require [CORRELATION]. If the Desire Monitoring Mechanism, for example, is capable of taking as input only desires, then it seems that this is because (most) desires have a non-relational property (or set of properties) that is lacked by (most) non-desires, and to which the mechanism is somehow

[20] If it *were* to encounter a non-belief, it would attempt to monitor it, but would (incorrectly) classify it as a belief.

[21] Imagine each monitoring mechanism has a lock and only mental states of the mechanism's proprietary attitude type can unlock it, and so be monitored by it.

sensitive. Without this kind of non-relational property difference, I cannot see how the mechanism would be "built" for desires.

Contra Goldman, positing multiple monitoring mechanisms is not a solution to the problem described in the previous section. Such a move will not allow Nichols and Stich's account to avoid the need for [CORRELATION]. That Goldman seems to think otherwise suggests either that his argument is different from my own, or that he does not appreciate the full force of his argument.

3.3 Filling Out [CORRELATION]

I now want to return to [CORRELATION]. Specifically, I want to consider which kind of property is most likely to fit the profile specified in [CORRELATION]. Because this profile calls for properties that are directly detectable, the focus is on the non-relational properties of token propositional attitudes. The question is whether there are properties of this kind that could indicate causal/functional role, and thus attitude type.

In pursuing this question I will borrow a bit from a discussion from Goldman (2006). In defending his own inner sense account, Goldman claims that mental states are classified as pains, tickles, visual perceptions, propositional attitudes, etc. via something resembling a transduction process. He writes that "[a] transduction process features inputs – events or properties to which the process is causally sensitive – and outputs – representations generated in response to these inputs" (2006, 246).[22] But what are the inputs to this process? That is, by which of a mental state's properties does inner sense

[22] He admits that talk of a transduction process in this context is metaphorical: "it is not literally being proposed that introspection involves a change in the form of *energy* by which information is transmitted" (257, fn 17).

determine that a given mental state is a pain, as opposed to a tickle, or itch, or perception, or desire? This is the question that Goldman must answer.

While Goldman's question is certainly similar to my own, I should note that Goldman never articulates anything nearly as specific as [CORRELATION]. Whether he has something like it in mind (applied to all mental sates rather than just the propositional attitudes) is unclear. In section 3.3.3.1, however, I will offer a piece of evidence that suggests that he does not.

Goldman considers four kinds of properties that might serve as the inputs to his proposed transduction-like process: functional properties, phenomenal properties, representational properties, and neural properties. Because I am concerned with [CORRELATION], and because [CORRELATION] requires *non-relational* properties that indicate causal/functional roles, I will put aside functional properties. The remaining three kinds of properties should thus be construed non-relationally. Finally, I will assume in what follows that the propositional attitudes have these three kinds of non-relational properties in question.

3.3.1 Phenomenal Properties

If phenomenal properties are plugged into [CORRELATION], the following claims is obtained:

> [CORRELATION_P] For each causal/functional role definitive of a given attitude type, there exists a directly detectable phenomenal property (or set of properties) had by mostly all and mostly only mental states that play that causal/functional role.

The question is whether this claim is at all plausible. From Goldman's discussion, we can extract three alleged reasons for thinking that it is not:[23]

> [a]ccording to orthodoxy, attitudes do not have phenomenal properties … [e]ven if all occurrent attitude tokens had some phenomenology, this would not guarantee the possibility of typing them *by* their phenomenology. For this, it would be necessary that each type have a *distinctive* phenomenal property (or set of properties) that could serve as the causal trigger of the corresponding introspective classification … [i]f it is granted that nonconscious mental states can be introspected (or monitored), that is a problem for the phenomenal-properties answer, because nonconscious states presumably lack phenomenal properties (249).

I find the first reason offered suspect. There is by no means a consensus on whether the propositional attitudes have phenomenal properties. Moreover, even if they do not, they might nevertheless typically be *associated* with, or *accompanied* by, such properties. For example, perhaps desires are sometimes accompanied by visceral feelings towards their objects. Even if the phenomenology does not strictly speaking belong to the attitude, it is nevertheless present and could potentially be exploited to detect the mental state's attitude type. According to a suitably altered [CORRELATION$_P$], phenomenal properties are not required to be 'had by' (mostly all) and (mostly only) mental states of the relevant type; 'had by' would be replaced with 'associated with' or 'accompanied by'.

The second and third reasons offered by Goldman are more serious. Supposing propositional attitudes have (or are associated with) phenomenal properties, I side with Goldman in finding unlikely the possibility that each attitude type has a "*distinctive* phenomenal property (or set of properties) that could serve as the causal trigger of the

[23] Interestingly, this is a shift for Goldman. His earlier work on introspection (1993, 2000) endorses the position that mental states are classified via their phenomenal properties. It is unclear whether he also thinks that such properties are used to classify input mental states along the content dimension. That such an account would be doomed to failure is, I think, fairly obvious. For this reason, the phenomenological account is best viewed as assigning the content task to a distinct method (redeployment, likely).

corresponding introspective classification" (249). Although Goldman does not expand on this point, seeing how one might is not difficult.

Consider the various kinds of propositional attitudes that exist. While I have spoken so far mostly of beliefs and desires, there are also intentions, fears, hopes, seemings, imaginings, etc. If one reflects on one's own phenomenology, I suspect that one will not attest to there being a distinctive phenomenology for *each* of these propositional attitude types. Since this is precisely what [CORRELATION$_P$] requires, this is reason to reject [CORRELATION$_P$].

Of course this is consistent with there being *some* attitude types that have (or are associated with) a distinctive phenomenal property (or set of phenomenal properties). If so, then inner sense could perhaps type such attitudes by these properties. However, this would mean that the detection of some propositional attitudes would be fundamentally different than the detection of others. For some, inner sense would execute the typing task via a detection of the attitude type's distinctive phenomenal property (or set of properties). For others, either inner sense would execute the typing task by some other means, or this task would be executed by a mechanism or method distinct from inner sense. The inner sense theory would presumably reject this second possibility. But if there exists another, non-phenomenal, means for executing the typing task, why would inner sense not use this means in all cases?

Moreover, I think Goldman is correct to bring up the relevance of non-conscious propositional attitudes. *If* such states can be detected by inner sense, and *if* non-conscious states lack phenomenal properties (or, more weakly, are not associated with any phenomenal properties), then the indirect detection of attitude type cannot proceed via the

direct detection of phenomenal properties, at least not if (as seems eminently reasonable) non-conscious states lack phenomenal properties. An inner sense account on which non-conscious states can be detected by inner sense must therefore provide a different, non-phenomenal, way of filling out [CORRELATION].

There is a connection between the second and third points that I would like to point out. I suppose one could argue that non-conscious mental states have phenomenal properties, and thus that some phenomenal properties are non-conscious. If so, then this undermines the significance of my claim that there does not *seem* to be a distinctive phenomenology had by (or associated with) each attitude type. Despite how things (consciously) seem, there might exist such distinctive phenomenology at the non-conscious level. While possible, I find this suggestion to be both implausible and unmotivated; invoking non-conscious phenomenal properties in order to save [CORRELATION$_P$] seems *ad hoc*.

Finally, I would like to consider the possibility that instead of there existing a unique phenomenal property (or set of properties) had by (or associated with) each attitude type (e.g., belief, desire, intention, fear, hope, etc.), there is a unique phenomenal property (or set of properties) had by (or associated with) each of the two main classes of attitude types, namely, the *cognitive* and *conative* attitude types. The former are characterized by a world-to-mind direction of fit, while the latter are characterized by a mind-to-world direction of fit.[24]

---

[24] A belief, e.g., has a mind-to-world direction of fit, given that a belief aims to fit a *representation* (of the world) to some state of the *world*. A desire, e.g., has a world-to-mind direction of fit, given that a desire aims to make some state of the *world* fit a *representation* (of the world). Anscombe (1957) seems to be the first to clearly articulate this distinction, although Austin (1953) discusses direction-of-fit in the context of speech-acts.

The virtue of this suggestion is that it reduces the number of distinct types to two: cognitive and conative. However, even if each attitude type neatly falls within one of these two classes, and even if each of these classes possesses a distinctive phenomenology, humans nevertheless self-ascribe the more specific attitude types. Thus, even if true, these two assumptions only show that the Monitoring Mechanism could use phenomenology to detect whether a given propositional attitude is cognitive or conative. The more fine-grained task would be left unexplained. I thus conclude that phenomenal properties are not the solution to the problem identified in Section 3.2.

3.3.2 Representational Properties

Perhaps representational properties will fare better than phenomenal properties. Plugging representational properties into [CORRELATION] yields the following claim:

> [CORRELATION$_R$] For each causal/functional role definitive of a given attitude type, there exists a directly detectable representational property (or set of properties) had by mostly all and mostly only mental states that play that causal/functional role.

On one interpretation, this claim is obviously false. After all, one can bear different attitudes towards the same proposition. If a propositional attitude's representational content is exhausted by the relevant proposition, then, e.g., the belief that *p* and the desire that *p* have the same representational properties. They are different propositional attitudes in virtue of a non-representational difference, namely, the difference in the

causal/functional roles that they play. On this interpretation, then, [$\text{CORRELATION}_{\text{R}}$] is false.[25]

There is, however, a different way of interpreting the above claim. This interpretation comes from Goldman's discussion of whether representational properties might serve as the inputs to his account's transduction-like process. He seems to have in mind representational features *in addition* to those that determine an attitude's propositional content. Such features might distinguish the belief that *p* from the desire that *p*. I suppose we might imagine that the content is tagged in a certain way, e.g., $p_{\text{-BEL}}$ and $p_{\text{-DES}}$. While Goldman rejects this possibility, he offers little more than a denial. He writes that "[w]hat distinguishes a state of believing from a state of desiring with the same content isn't a matter of some further intentional content (251)". Although I agree with this claim, I think a bit more can be said on its behalf.

The representational properties in question are supposedly distinct from those that determine a propositional attitude's content. They are thus not needed in order to explain why the desire that *p* has the content that it does. Moreover, as I have noted before, a mental state's attitude type is determined by the causal/functional role that it plays within the cognitive system of which it is a part. If a given state did not play the causal/functional role of belief (e.g., if it did not interact with desires to form intentions,

---

[25] Notice that this is so even on the assumption that representational content is "inside the head". This distinguishes my objection to inner sense accounts of attitudinal detection from the much discussed charge that content externalism of the sort advocated by Putnam (1975) and Burge (1979) is incompatible with privileged access; in addition to Boghossian (1989), see McKinsey (1991) and Brown (1995). This is not my objection. In fact, I am sympathetic to the compatibilist solution to this problem advocated by Burge (1988) and John Heil (1988); see also McLaughlin and Tye (1998). The problem posed by Boghossian and others centers on the conflict between privileged access and the alleged extrinsic/relational nature of mental content. Although the problem I am pressing here also concerns extrinsic/relational properties, these properties are not representational properties, but are instead the causal/functional properties that determine attitude type. Thus, as noted in an earlier footnote, my argument is importantly different than Boghossian's.

or if it did not interact with other beliefs in drawing inferences), then it would not be a belief. And this would be so even if that mental state had, in addition to its propositional content, some representational content distinctive of beliefs. Such content on its own thus cannot explain why a given state is a belief, as opposed to a desire or intention.

What, then, could be the purpose of this content? Perhaps it helps to determine the causal/functional role that the mental state plays. That is, perhaps a given belief plays the causal/functional role definitive of belief because it has a certain kind of content that is unique to belief. And mutatis mutandis for the other attitude types. While this suggestion is certainly possible, I nevertheless find it a bit unmotivated. Moreover, there appears to be a set of properties that are better suited for this purpose. The properties I have in mind are neural properties, to which I now turn.[26]

3.3.3 Neural Properties

Because the propositional attitudes are realized in the brain, they will have various neural properties. Moreover, a given state's neural properties are obviously relevant to the sorts of causal relations it participates in. Given this connection to causal/functional role, and thus also to attitude type, neural properties are seemingly the best candidates for fitting the profile specified in [CORRELATION]. That they are stronger candidates than phenomenal and representational properties is, I think, clear. The following instantiation of [CORRELATION] is thus the most promising:

[26] Plausibly, the mind has multiple representational mediums. Perhaps, then, each kind of propositional attitude uses its own distinctive medium, in which case attitude type could be (indirectly) detected via representational properties. While I do not doubt that the mind has multiple representational mediums, I am quite skeptical that each attitude type has its own distinctive medium. For this reason, I will not discuss this possibility any further.

> [CORRELATION$_N$] For each causal/functional role definitive of a given attitude type, there exists a directly detectable neural property (or set of properties) had by mostly all and mostly only mental states that play that causal/functional role.

I thus conclude that Nichols and Stich's Monitoring Mechanism account of introspection should be paired with [CORRELATION$_N$]. Such a pairing gives the account the best chance of being true. For practical purposes, then, Nichols and Stich's account is committed to this claim about the neurological realization of the propositional attitudes; their account can explain the execution of the typing task only if psychological attitude types are correlated with neural properties in the way specified by [CORRELATION$_N$]. This is the psycho-neural claim mentioned in the title of this chapter.

3.3.3.1 Goldman On Neural Properties

Before turning to the inner sense theory more generally, I should note that Goldman also settles on neural properties. He claims that these properties serve as the inputs to his account's transduction-like process. Importantly, though, he never articulates anything nearly as specific as [CORRELATION$_N$]. Whether he has something like it in mind (as applied to all mental sates, not just the propositional attitudes)[27] is unclear.

While speaking of sensations, he cites the work of A.D. Craig (2002), which suggests that there are dedicated neural circuits for different types of sensations.

---

[27] Recall that Goldman, like Nichols and Stich, is concerned with providing an account of inner sense that works for all those mental states we are able to know in a first-personal way; the propositional attitudes are a mere (proper) subset of this class of mental states.

Goldman takes this to support his claim that inner sense classifies mental states via their neural properties. But are such circuits likely to exist for the propositional attitudes? Interestingly, Goldman admits that they are not, offering the following rather cryptic remark intended to show that this is not a problem for his account:

> But such circuits are unlikely to be available for the propositional attitudes. Does this fact scotch the neural-properties approach? No. The existence of distinctive neural properties that are usable by an introspection device does not require dedicated circuits. A dedicated circuit involves the "front end" of a mental-event type, and introspection isn't concerned with the front end. So the neural properties approach remains very promising (252-3).

I must confess that I do not know what to make of this passage. Why exactly does the absence of dedicated neural circuits for the propositional attitudes not cast serious doubt upon what he calls the 'neural-properties approach'? The final two sentences of the quotation are obviously meant to address this question, but how they do so is unclear to me. My main point is that whether Goldman appreciates the need for something like [CORRELATION$_N$] is not at all obvious. Thus, despite the similarities that exist between my main argument in this chapter and his discussion of inner sense, I take my conclusion to be both important and original.

3.4 Moving Beyond Nichols and Stich's Account

I have argued that Nichols and Stich's Monitoring Mechanism account is committed to [CORRELATION$_N$]. Given the influence of their account, this is significant. Ideally, however, my argument would have wider application. Fortunately, I think it does. I believe that any account warranting the name 'inner sense' is committed to [CORRELATION$_N$]. The argument for this more general conclusion can be briefly stated.

Any inner sense account must handle attitudinal detection. This means that any such account must explain the detection of both the content and attitude type of a propositional attitude. Because attitude type is determined by causal/functional role, any such account must explain how a propositional attitude's causal/functional role is detected. A propositional attitude's causal/functional role is a relational property of that state, and so can be detected only indirectly. But a propositional attitude's causal/functional role can be indirectly detected only if [CORRELATION] is true. Finally, as I argued in Section 3, neural properties are the properties most likely to fit the profile specified in [CORRELATION], thus showing that the indirect detection of a propositional attitude's causal/functional role likely requires [CORRELATION_N]. An inner sense account's ability to explain attitudinal detection thus likely requires the truth of this claim.

I must admit that implicit in this argument is a particular understanding of the inner sense theory, which I should now make explicit. Recall from Section 2.2 that the inner sense theory is not committed to the existence of a dedicated organ of inner sense; this is one respect in which the analogy with ordinary sense perception is not perfect.[28] While I of course recognize this feature of the theory, I nevertheless assume that the theory is committed to there being *something* (or a collection of *somethings*) that identifies the introspected mental states by virtue of causally interacting with their properties. Call this something (or collection of somethings) 'S'.[29] I take S to be distinct from both the resulting introspective beliefs and the introspected mental states. The

[28] Although recall that Armstrong (1968/1993) notes that there is no dedicated organ of proprioception, which he regards as a kind of perception.

[29] Just as Nichols and Stich allow for multiple monitoring mechanisms, so too could there be multiple *S*s, each charged with monitoring a certain kind of mental state.

accounts of Nichols and Stich, Goldman, and Lycan seem to be of this type. My claim is that S can categorize a propositional attitude's attitude type (i.e., execute the typing task of attitudinal detection) only if [CORRELATION$_N$] is true.

There is, however, a potential problem with this way of understanding the theory. Recall that in Section 2.2 I described a version of the inner sense theory that Shoemaker (1994a) calls the 'broad perceptual model'. This version of the theory takes on only two of the eight features of ordinary sense perception identified by Shoemaker. These two features are those I earlier labeled the 'causal' and 'independence' features. The resulting view is one on which: (i) introspective beliefs are causally produced by the mental states perceived, via a causal mechanism that normally produces beliefs that are true and (ii) the mental states of which the perception is of, and which it provides knowledge about, exist independently of the perceiving of them.

Although the causal condition references a 'causal mechanism', it does not say that this mechanism identifies introspected mental states by virtue of causally interacting with their properties. Consider the following quote from Armstrong (1968/1993) whom Shoemaker claims endorses the broad perceptual model of inner sense:

> Eccentric cases apart, perception, considered as a mental event, is the acquiring of information or misinformation about our environment. It is not an 'acquaintance' with objects or a 'searchlight' that makes contact with them, *but is simply the getting of beliefs*. Exactly the same must be said of introspection. It is the getting of information or misinformation about the current state of our mind" (1968/1993, 326, emphasis added).

He later writes that introspection, like perception, is "a mere flow of information or beliefs" (ibid).

These remarks suggest a picture on which a mental state, perhaps under certain conditions, simply causes the subject to come to believe that it is present. For example, a desire that *p*, perhaps under certain conditions, simply causes one to come to believe that one desires that *p*. The belief is logically independent of the desire, the desire causally produces the belief, and yet there is no S charged with the task of determining the desire's content and attitude type. If this is a version of the inner sense theory, then the main argument from this chapter does not apply to the inner sense theory as a whole.

I have two points to make in response to this issue. First, although Armstrong's remarks suggest the picture just sketched, other remarks seem to suggest a quite different picture. For example, he repeatedly claims that introspection is a "self scanning process in the brain" (1968/1993, 324). But 'scanning' suggests the presence of a scanner (or scanners). Moreover, a scanning process would seem to involve a scanner (or scanners) causally interacting with the properties of that which is scanned. Because the result of the "self scanning process" is an introspective belief, the belief would appear to be the result of the scanner (or scanners) causally interacting with the properties of the introspected mental state. Thus, on this characterization of the account, there *is* an S.

Of course whether or not Armstrong endorses the broad perceptual model is not important. What matters is whether the broad perceptual model is a version of the inner sense theory. If it is, then I cannot claim that the inner sense theory as a whole is committed to [CORRELATION$_N$]. At most, I can say that a particular kind of inner sense theory is committed to [CORRELATION$_N$]. This brings me to my second point.

The broad perceptual model is too dissimilar to ordinary sense perception to warrant the title 'inner sense'. As I look at the objects on my desk right now, I see a

laptop, a pile of books, a coffee mug, etc. I have the perceptual belief that there is a coffee mug [there]. The coffee mug is independent of my perceptual belief. It is also part of the causal sequence that produced that belief. My perception of the mug thus satisfies the two conditions of the broad perceptual model.

But there is much more to the story than these two facts. My perceptual belief is the product of a process that is causally sensitive to certain properties of the coffee mug. These properties allow me (or my visual system) to recognize the coffee mug as a coffee mug. The perceptual belief is in some sense the result of an *inspection* of the coffee mug and its properties. This is exactly what is missing from the broad perceptual model. It denies, for example, that my introspective belief that I have a dull pain in my left shoulder is the result of a process of recognition. Given this, I am reluctant to classify the account as one of inner *sense*. Because my aim in this chapter is the inner sense theory, my argument's failing to apply to this kind of account is unproblematic.

3.5 The Significance of [CORRELATION$_N$]

I have argued that the inner sense theory is committed to a previously unrecognized empirical claim about the neurological realization of the propositional attitudes. But is [CORRELATION$_N$] true? Unfortunately, current science is not helpful here. As far as I know, there is no data on the neurological realization of the propositional attitudes.

Nevertheless, I would like to briefly sketch two lines of argument, one for and one against [CORRELATION$_N$]. First, I noted in Section 3.3.3 that a mental state's neural properties are surely relevant to the causal/functional role that it plays. A belief that *p* plays the causal/functional role definitive of belief, as opposed to desire, in part due to

the fact that certain of its neural properties enable the kinds of causal interactions that are definitive of belief, but not desire. Or at least this seem reasonable to me. Given this connection between a mental state's causal/functional role (and thus attitude type) and its neural properties, finding the kinds of correlations specified by [CORRELATION_N] would not be too surprising.

On the other hand, causal/functional roles appear to be the sorts of properties that are prime candidates for being multiply realizable. *Corkscrew* is a causal/functional kind.[30] Corkscrews have the function of removing corks from wine bottles. However, there are many ways to achieve this end. A waiter's corkscrew and a double-lever corkscrew achieve this end in different ways: while the former consists only of a screw and a single lever, the latter consists of a screw and two wing-like appendages that rise as the screw enters the cork. Arguably, the waiter's corkscrew and the double-lever corkscrew are members of the kind corkscrew, but are different realizations of that kind in virtue of their different ways of fulfilling the relevant function. And what is true of corkscrews is also true of carburetors, mousetraps, computer programs, and so on.

The relevance of multiple realizability to the current discussion is this. If there are multiple physical ways in which a mental state with the content *p* might be realized so as to play the causal/functional role definitive of desire, then this role (and thus the attitude type *desire*) is multiply realizable. If so, then even if a mental state's neural properties largely determine its causal/functional role, the mental state would be able to play that

[30] The following example comes from Shapiro (2004). One of the main theses of his book is that philosophers are often too quick to conclude that a given kind is multiply realized. Shapiro argues that a mere physical difference between two realizations, r_1 and r_2, of a kind, K, does not suffice for r_1 and r_2 being different realizations of K. For that, r_1 and r_2 must differ in their 'r-properties', i.e., those properties that causally contribute to the functional task that defines K. See also Shapiro (2000).

role with *different* neural properties. Thus, multiple realizability might be thought to threaten [CORRELATION$_N$].[31]

Notice, though, that the multiple realizability of attitude type does not *on its own* threaten [CORRELATION$_N$]. In order to threaten this claim, attitude type must be multiply realiz*ed*. Moreover, it must be multiply realized in a particularly strong way. This is because, as discussed in Section 3.2, [CORRELATION] is consistent with an attitude type's being correlated with a *disjunctive* set of properties. [CORRELATION$_N$] is thus consistent with an attitude type's being correlated with a disjunctive set of neural properties.

Having briefly surveyed two lines of argument concerning the truth of [CORRELATION$_N$], it seems that neither is particularly persuasive. For this reason, whether the inner sense theory's commitment to [CORRELATION$_N$] helps or hurts the theory is unclear.

3.6 Conclusion

In this chapter I have argued that the inner sense theory is committed to a claim about the neurological realization of the propositional attitudes. This claim is [CORRELATION$_N$]. While the status of this claim is unclear at this time, the exposure of this commitment has the potential to further the empirical investigation of the inner sense theory.

---

[31] Neural plasticity is perhaps relevant here. The brain appears to be highly flexible/plastic; see Levin and Grafman (2000) and Freund et al. (1997) for details on neural plasticity. If a region of the brain subserving some cognitive process is damaged, other parts of the brain not previously responsible for the functioning of that process will often "take over", ultimately serving as the new neural realizer of that process. In short, there are data suggesting that the connections between many cognitive processes and the neural properties subserving those processes are highly contingent. Of course this is not to say that this is the case for all cognitive processes. Arguably, mirror neurons are needed in order to subserve the processes responsible for face-based emotion recognition (FBER).

Chapter 4: How *Not* to Argue For or Against the Inner Sense Theory

4.1 Introduction

The inner sense theory is commonly thought to predict the existence of individuals with first-personal deficits/impairments of one kind or another. One can thus argue for (or against) the theory by noting the presence (or absence) of such deficits/impairments. My focus in this chapter is primarily on the first, predictive step of this two-step argument. I argue that various first-personal deficits/impairments that might appear to be predicted by the theory are *not* so predicted. Moreover, I argue that those deficits/impairments that are legitimately predicted by the theory will be quite subtle, and so difficult to detect. Together these two arguments show that the two-step argument in question is much more difficult to successfully execute than was perhaps previously thought.

I begin with a very brief description of two recent presentations of the inner sense theory (4.2). Although these accounts were discussed in the previous chapter, briefly doing so here is, I think, useful. Next, I look at two versions of the two-step argument from the recent philosophical literature (4.3). The first is an argument *for* the inner sense theory, put forth by Shaun Nichols and Stephen Stich (2003). The second is an argument *against* the inner sense theory, put forth by Mark Engelbert and Peter Carruthers (2010). In the bulk of what remains, I develop the two arguments mentioned in the previous paragraph (4.4-4.7).

4.2 The Inner Sense Theory – Two Recent Accounts

Nichols and Stich (2003) offer a fairly simple explanation of the seemingly special, direct access that we have to our own mental states. They claim that to explain such access to, for example, our beliefs, all that is needed is a monitoring mechanism that: (i) copies the representations in one's "Belief Box," (ii) embeds the copied representations in an 'I believe that ____' representation schema, and (iii) places the new meta-representations into one's Belief Box. This approach can be extended to other mental state types by simply adding representation schemas appropriate to those types: an 'I desire that ____' schema (for desires), an 'I intend that ____' schema (for intentions), an 'I am having a visual experience that ____' schema (for visual experiences), etc.

Nichols and Stich leave open whether there is single monitoring mechanism equipped with representation schemas for all mental state types, or multiple monitoring mechanisms, each specialized for a certain type of mental state. Inner sense, for Nichols and Stich, is constituted by this simple kind of monitoring.

Goldman (2006) offers an account of inner sense that is quite similar to Nichols and Stich's. There are, however, a few differences worth noting. First, Goldman opts for a single mechanism of inner sense, denying multiple monitoring mechanisms on grounds of parsimony. Second, unlike Nichols and Stich, he explains how the monitoring mechanism selects the appropriate representation schema within which to embed a copied representation. Notice that Nichols and Stich's boxology is metaphorical; a "box", for them, refers to a functionally characterized set of mental states. For this reason, a monitoring mechanism cannot make the selection simply by noting the "box" in which a mental state resides. On Goldman's account, the monitoring mechanism *recognizes* a

mental state's type via its neural properties. This component of Goldman's account strengthens the perceptual analogy at the core of the inner sense theory.

4.3. The Two-Step Argument

If there is a mechanism of inner sense, as the inner sense theory claims, then damage to this mechanism should be possible. Moreover, there appears to be agreement that such damage should bring with it first-personal deficits of one kind or another. The presence or absence of such deficits has thus been used to argue for or against the inner sense theory. This is the two-step argument that is the focus of this paper. I will now briefly sketch two recent versions of the argument, one for the inner sense theory and one against it.

4.3.1 Nichols and Stich's Argument For the Inner Sense Theory

We have just seen that, for Nichols and Stich, inner sense is constituted by one or more monitoring mechanisms. Importantly, they regard these mechanisms as being *distinct* from those mechanisms (whatever their nature) that subserve our ability to attribute mental states to others, and to reason about those states in predicting and explaining behavior; this ability is a major component of what is known as 'mindreading'. Given this distinctness, Nichols and Stich claim that either of these could be damaged while the other is not. They write:

> [O]n our theory it is possible for one or more of the [monitoring mechanisms] to malfunction, causing a deficit in one or more aspects of first-person mental-state detection, while the [heterogeneous collection of mental mechanisms which subserve mindreading] continue to function normally (2003, p. 188).[32]

[32] They also note the possibility of the opposite pattern of dissociation, where mindreading is impaired while inner sense is not. In this paper, however, I am concerned only with arguments for and against the inner sense theory that make use of the possibility of damage to inner sense. For this reason, I will set aside this direction of Nichols and Stich's argument.

This is the first, predictive step of their argument for the inner sense theory. They go on to claim that this prediction is borne out, thus confirming the inner sense theory.[33] I will discuss this second step in Section 4.7.1.

4.3.2 Engelbert and Carruthers' Argument Against the Inner Sense Theory

Engelbert and Carruthers (2010) reject the inner sense theory, taking aim at the two versions of the theory described above. Their central claim is that, on both accounts, inner sense is *dis*unified. Inner sense is constituted by either multiple distinct monitoring mechanisms (one for each mental state type), or a single mechanism with multiple distinct channels (one for each mental state type).[34] Such disunification, they claim, makes possible selective damage, i.e., damage to some mechanisms/channels, but not others. They allege that such damage should lead to "people who can self-attribute beliefs but not desires, or who can self-attribute visual experiences but not auditory ones, and so forth" (246).

This is the first step of their argument against the inner sense theory. They go on to claim that this prediction is not borne out, thus disconfirming these two versions of the inner sense theory.[35]

---

[33] They are primarily concerned with arguing against the 'theory theory of self-awareness', according to which a single method subserves both first-personal and third-personal mental state detection. For an articulation and defense of this view, see Gopnik (1993) and Gopnik and Wellman (1994).

[34] The details of their argument for why a single monitoring mechanism must have distinct channels is not important for my purposes; the point can be granted.

[35] Their aim in attacking the inner sense theory is to support what Carruthers (2009) calls the 'mindreading is prior model', according to which (roughly) we have no special method for self-attributing mental states; we self-attribute mental states via the very same method that we use to attribute mental states to others. For current purposes, this view can be seen as equivalent to the 'theory theory of self-awareness' mentioned in footnote four. See Carruthers (2009, 2010, 2011) for a nuanced discussion of this view.

4.3.3 The Two Predictions

Making explicit the predictions from the two arguments just described will be useful in what follows:

> Nichols and Stich's Prediction: there should be individuals who suffer deficits in first-person mental state detection, but not third-person mental state detection.
>
> Engelbert and Carruthers' Prediction: there should be individuals who are capable of self-attributing certain types of mental states, but incapable of self-attributing other types.

Notice that while the prediction from Engelbert and Carruthers' argument concerns a specific first-personal deficit, namely, an *inability* to self-attribute certain kinds of mental states, Nichols and Stich's prediction does not. Their prediction is compatible with various first-personal deficits, including but not limited to the one described by Engelbert and Carruthers. Importantly, given that various kinds of first-personal deficits are possible, the first step of the two-step argument can be instantiated in various ways. In this chapter I will consider multiple such instantiations.[36]

4.4 Engelbert and Carruthers' Argument and the Dual Method Theory

I want to begin by looking at the prediction from Engelbert and Carruthers' argument. They claim that damage to inner sense should block the individual's capacity to self-attribute certain kinds of mental states. The issue I want to examine is whether this first-personal incapacity is a genuine prediction of the inner sense theory. If it is not, then

[36] To take two examples, I will consider both (i) the possibility that one is *unreliable* in self-attributing certain types of mental states and (ii) the possibility that one is *less* reliable than others in self-attributing certain types of mental states. I am thus using 'deficit' to cover problems of this sort.

the alleged absence of individuals with this incapacity cannot disconfirm the inner sense theory via the two-step argument.[37]

Importantly, damage to inner sense should lead to this first-personal incapacity only if the following auxiliary assumption is true:

> [Auxiliary Assumption]: humans have no method for attributing mental states that (i) can be applied to oneself and (ii) is independent of any alleged mechanism of inner sense.

The sense of independence stated in condition (ii) is important. A method for attributing mental states satisfies this condition so long as it can operate without inner sense. A method's satisfying condition (ii) is thus consistent with inner sense's being ordinarily involved in its operation; it need only be such that it *can* operate without inner sense. This issue will be further discussed in the following section.[38]

The role of [Auxiliary Assumption] in Engelbert and Carruthers' argument is a bit unclear. Because they oppose the inner sense theory, they would certainly deny the assumption. However, it is very doubtful that the inner sense theory is committed to [Auxiliary Assumption]. Many inner sense theorists, including both Goldman and Nichols and Stich, endorse what is called the 'dual method theory'.[39] They believe that, in addition to inner sense, humans sometimes self-attribute mental states via an indirect, interpretive method of the sort used to attribute mental states to others; the main reason why they endorse the dual method theory will become clear in the next section. If the

---

[37] Also, if it is not a prediction of the theory, then even if there were such individuals, they would not *confirm* the inner sense theory via the two-step argument.

[38] Perhaps, despite what they wrote, Engelbert and Carruthers did not intend to claim that those with damaged inner sense should be *incapable* of self-attributing certain kinds of mental states. Perhaps they had a weaker prediction in mind, one which does not require [Auxiliary Assumption]. I will consider numerous weaker predictions in Section 4.6.

[39] As far as I know, this name comes from Goldman (2006).

dual method theorist takes this interpretive method to be independent of inner sense (in the manner specified above in condition (ii)), then she must reject [Auxiliary Assumption].

In any case, I will argue in the following section that [Auxiliary Assumption] is false. Moreover, as will become apparent, its falsity in no way jeopardizes the inner sense theory; the inner sense theory is perfectly compatible with the falsity of [Auxiliary Assumption]. For this reason, the inner sense theorists should reject [Auxiliary Assumption], thereby undermining Engelbert and Carruthers' argument against that theory. The importance of my argument against [Auxiliary Assumption] goes beyond the reasons just given. Various parts of the argument will play crucial roles in the remainder of this chapter, especially Section 4.6.

4.5 An Argument Against [Auxiliary Assumption]

The falsity of [Auxiliary Assumption] may appear obvious. After all, humans are mindreaders, i.e., they can attribute mental states to others in a fairly reliable way. Arguably, mindreading is central to social cognition, and both philosophers and psychologists have done a great deal of work exploring the mechanism(s) thought to subserve it.[40] If mindreading satisfies conditions (i) and (ii) from above, then [Auxiliary Assumption] is false. I suspect that many philosophers and psychologists already believe that it does satisfy these conditions, and so already reject [Auxiliary Assumption]. I do

[40] There are those who doubt mindreading's importance to social cognition. See, e.g., Tad Zawidzki (2008).

not wish to rest content here, however. I will now offer an argument against [Auxiliary Assumption] that does not rely on any assumptions about the nature of mindreading.[41]

The argument is based on the phenomenon of confabulation. A person confabulates when she self-attributes a mental state that is not present, but which, if present, would help to make sense of her behavior.[42] Confabulations are the products of an *interpretive* method that generates beliefs about one's mental life based on evidence concerning one's behavior and situation. Much psychological work has demonstrated the occurrence of confabulation for a variety of mental state types, including intentions (Wegner and Wheatley, 1999), judgments (Nisbett and Wilson, 1977; Eagly and Chaiken, 1993; Gazzaniga, 1995), and decisions (Brasil-Neto, et al., 1992). These studies, and others, are discussed in Carruthers (2009).

In the next two sub-sections I will describe in detail two of these studies, namely, Wegner and Wheatley (1999) and Gazzaniga (1995). These studies nicely illustrate some of the features of confabulation that are important to both the main argument of this section and other arguments to come.

4.5.1 Wegner and Wheatley (1999)

In Wegner and Wheatley's study participants and confederates shared control of a mouse connected to a computer monitor displaying various pictured objects. Participant-confederate pairs took part in multiple trials, each consisting of a period of joint-movement and a period of joint-stopping. Both participants and confederates wore headphones and were told that they would hear one spoken word per trial; music was

[41] By denying [Auxiliary Assumption], I am of course not claiming that all humans have a method for attributing mental states that satisfies conditions (i) and (ii). Surely, this is false. Rather, I am claiming that *normal* humans have a method of this kind.

[42] This definition is not perfect, but it will suffice for my purposes.

used to cue the period of joint-stopping. Each spoken word named an object, although only sometimes was the object pictured on the monitor. Participants were told that the words were intended to serve as distracters, while, in fact, they were meant to *prime* thoughts in the participants about the objects named. Participants were under the *false* impression that their confederate partners were hearing the same kind of recording, but with different words spoken at different times. In fact, confederates were being instructed by the experimenters to make particular movements at particular times.

There were two kinds of trials. In forced stop trials, the confederate would manipulate the mouse so that the cursor would come to a stop on the pictured object named over his partner's headphones. In unforced stop trials, the confederate would not manipulate the mouse, allowing the participant to freely bring the cursor to a stop. Importantly, in the unforced stop trials, there was no statistically significant difference between where the cursor stopped when the participant's word named an object pictured on the monitor and where it stopped when it did not name an object pictured on the monitor. This was interpreted as showing that the priming did not cause the participants to form intentions to stop their cursors on the pictured objects named.

Nevertheless, in the forced stop trials, participants were on average inclined to (falsely) self-attribute the intention to stop their cursor on the pictured object named over their headphones. Much more importantly, though, the degree of intentionality self-attributed in these cases varied in ways strongly suggesting that these self-attributions were produced by an indirect, interpretive method. Specifically, when the priming word was heard *before* the (forced) stop, the strength of intentionality self-attributed was inversely proportional to the amount of time between when a participant heard the

priming word and when the cursor stopped on the corresponding pictured object. Participants were unwilling to (falsely) self-attribute an intention when the priming word was heard one second *after* the (forced) stop.

Important to notice here is the reasonableness of the confabulated intentions. In a forced stop trial, the cursor stopped on the pictured object that the participant was primed to be thinking about. In such a case, the interpretation that one intended to stop the cursor on that object is perfectly reasonable. This reasonableness is even greater, if the thought immediately preceded the stop. And, as I just described, this was reflected in the judgments of the participants. In more ordinary contexts, such interpretations would most likely be true. This point will be relevant in a later section.

These points strongly suggest that the self-attributions in the forced stop trials were produced by an indirect, interpretive method sensitive to at least two factors: (i) whether the cursor stopped *before or after* the word was heard and (ii) the *amount of time* between when the cursor stopped and when the word was heard. The participants employed some kind of method (a sub-personal one, most likely) that took these factors to be relevant to whether and to what extent they had intended to behave as they did. The data thus strongly support the existence of an indirect, interpretive method for self-attributing mental states.

4.5.2 Gazzaniga (1995)

Next consider the research on so-called ‘split-brain’ individuals. In cases of severe epilepsy, some patients opt for cerebral commissurotomy, an operation that removes one’s corpus callosum and anterior commissure. As a result, the brain hemispheres cannot communicate with each other; information received by the right

hemisphere is closed off from the left hemisphere, and vice versa. Because the left hemisphere is verbal and the right hemisphere is nonverbal, speech for these patients is the sole product of the left hemisphere.

For these reasons, when information presented/available only to the right hemisphere prompts behavior, the left hemisphere's spoken explanation of this behavior must be arrived at via interpretation. This phenomenon forms the basis of a number of fascinating studies on split-brain patients conducted by Michael Gazzaniga; he summarizes these findings in his (1995).[43]

In one study, Gazzaniga presented commands to split-brain patients' right hemispheres. Not surprisingly, these commands were obeyed. What is interesting is the patients' explanations of these commanded behaviors. In one case, a patient's right hemisphere was presented with the command 'laugh', thereby causing the patient to laugh. When asked why he was laughing, the (left hemisphere of the) patient answered "You guys come up and test us every month. What a way to make a living!" (Gazzaniga, 1995, 1393). According to Gazzaniga, "[h]owever this type of test is manipulated, it always yields the same kind of result" (ibid).

In another study, Gazzaniga presented split-brain patients with two different images, each being shown exclusively to one hemisphere. Patients were then presented with an array of photos (in full view of both hemispheres) and asked to select two photos corresponding to the original images shown. The catch was that patients had to select one photo with their right hands (controlled by their left hemispheres) and the other with their left hands (controlled by their right hemispheres).

---

[43] See also his 2000.

In the case of the patient, P. S., an image of a chicken claw was flashed to his left hemisphere, while an image of a snow scene was flashed to his right hemisphere. When asked to select photos corresponding to these flashed images, P.S.'s left hand selected a photo of a shovel (suitable for the snow scene image), while his right hand selected a photo of a chicken (suitable for the chicken claw image). When asked to explain why he selected these photos, (the left hemisphere of) P. S. answered "Oh that's simple. The chicken claw goes with the chicken, and you need a shovel to clean out the chicken shed".

In these cases, the patients' explanations of their behavior are obviously incorrect. In the first case, the patient laughed because he was commanded to do so, *not* because he found the testing funny. In the second case, the patient selected the photo of the shovel because he judged it to be the photo that best corresponded to the snow scene image, *not* because of anything having to do with chickens. Clearly, these explanations were confabulated. These explanations, although reasonable in light of the evidence available to (the left hemispheres of) these patients, were clearly the products of an interpretive method.

4.5.3 The Lesson

The data from both Wegner and Wheatley's study and Gazzaniga's studies make a strong case for the existence of an indirect, interpretive method for self-attributing mental states. Such an interpretive method is clearly orthogonal to the direct, perception-like method that inner sense is alleged to be. For this reason, the data make a strong *prima facie* case for this method's independence from inner sense. Such a method, though, is exactly what [Auxiliary Assumption] denies. These studies thus suggest the

falsity of [Auxiliary Assumption]. Importantly, this conclusion stands even in the face of those skeptical of the claim that humans *often* confabulate.[44] My denial of [Auxiliary Assumption] requires only that humans have the capacity to confabulate, not that they often exercise this capacity.

4.5.4 *Not So Fast!* – The Simulation Theory

I must admit that consistent with what I have argued so far is the possibility that confabulation is subserved by a method that involves inner sense in a way that violates condition (ii). One way of motivating this possibility is to recognize that inner sense plays an important role in some accounts of mindreading. The simulation theory is such an account.[45] But, as I will now explain, there is good reason to think that inner sense plays only a *dispensable* role in mindreading, on this theory. If so, then condition (ii) is not violated.[46]

The basic idea behind the simulation theory of mindreading is that when attempting to understand what some individual, S, is thinking, how S will act, or why S acted as she did, one attempts to "step into S's shoes." Consider the following

---

[44] Typically, dual method theorists are such skeptics. They doubt the pervasiveness of confabulation, while accepting that it occurs sometimes. Moreover, the existence of confabulation is typically the reason why one moves to a dual method theory; Goldman (2006), e.g., cites the confabulation data when discussing his endorsement of the dual method theory.

[45] Not all simulationists are committed to inner sense. Robert Gordon (1995), for example, offers a simulationist account that rejects inner sense.

[46] There is a more direct, albeit unlikely, way in which confabulation might be subserved by a method that involves inner sense in a way that violates condition (ii). Perhaps confabulation is subserved by inner sense alone. This possibility is implausible for two reasons. First, recall that confabulations are, by my above definition, incorrect. In confabulation studies, there are very good reasons to suppose that the self-attributed states are not present. If inner sense is that which subserves confabulation, then we must explain why inner sense gets things so wrong in these situations. Presumably, the participants in these studies have properly functioning inner senses (assuming, of course, that the inner sense theory is true). Second, recall how reasonable confabulations are in light of the context and behavior of the confabulator. Confabulations are plausible-sounding, although false, mentalistic explanations of one's behavior. But there is simply no explanation for why inner sense should get things wrong in such a constructive, reasonable way. Inner sense is supposed to be a *detection* method, not an inferential, interpretive one. This all suggests that we should reject the hypothesis that inner sense on its own is that which subserves confabulation.

oversimplified example.[47] Suppose you see S behave in some specific way, X. In order to understand why S (intentionally) behaved in this way, you try to imagine what S's mind was like prior to S's behavior. Suppose you generate various hypotheses, including the hypothesis that S had a certain belief, Y, and a certain desire, Z. (For now, put aside the issue of how these hypotheses are generated.) The next step is to test these various hypotheses, by running them through your own decision-making process. By doing this, you are essentially determining what decision to act you would have made if *you* had had a given hypothesized belief-desire pair.

Inner sense contributes to mindreading by detecting the outcome of this procedure. Suppose that, after testing the Y-Z belief-desire hypothesis, inner sense detects that the outcome is a decision to X. Given that this decision matches S's apparent decision (as evidenced by S's doing X), this is evidence in favor of that particular belief-desire hypothesis. The final step in this process is to attribute this belief-desire pair to S. I will call mental state attributions arrived at in this way 'simulation-based attributions'.

Importantly, this example shows that simulation-based attributions are not arrived at purely by simulation. Prior to running a simulation, a mindreader must generate various belief-desire hypotheses. Moreover, because there is an indefinite number of mental hypotheses that can make reasonable S's doing X, the generated hypotheses must be narrowed. Mindreaders, on this account, must have some way of intelligently narrowing (or framing) the pool of possible hypotheses. Again, this narrowing process

[47] The following sketch of the simulation theory is based on Goldman's presentation from Chapter 2 of his 2006.

must occur prior to the running of any simulations. For these reasons, what we might call a 'pure simulation theory' must be rejected in favor of a hybrid theory.[48]

The relevance of these points should be clear. There is significant pressure for simulation theorists to posit a component of mindreading that is capable *on its own* of generating and narrowing reasonable hypotheses about the mental lives of others. Presumably, then, on this theory, if one's inner sense were to become damaged, one would still be capable of attributing mental states to others in a fairly reliable way. All that follows is that these attributions would not be simulation-based. For this reason, even if confabulation is subserved by a simulationist method, and even if this method involves inner sense, the method likely satisfies condition (ii) from above. Accordingly, the simulation theory does not, after all, pose a threat to my argument against [Auxiliary Assumption].

I know of no account of mindreading other than the simulation theory that could possibly be suspected of having inner sense as an indispensible component. I thus conclude that the indirect, interpretive method that I have argued subserves confabulation satisfies *both* conditions (i) and (ii), and thereby undermines [Auxiliary Assumption].[49]

---

[48] This is not a criticism of the simulation theory. The simulation theory remains a substantive competitor to non-simulationist theories so long as simulation-based attributions play a significant role in mindreading.

[49] There is one further complication that I would like to quickly address. Suppose that the method argued for in this section is such that its proper development requires a properly functioning inner sense. If this were the case, then those with damaged inner sense from birth should not have this method. This is certainly true. However, the inner sense theory should not be interpreted as being committed to this developmental supposition. The theory is certainly compatible with its falsity, and I suspect many inner sense theorists would deny it. For this reason, when examining the two-step argument that is the focus of this paper, the truth of this supposition should not be assumed. To assume its truth would be to shift the focus of the argument: it would no longer concern the inner sense theory, but rather the *conjunction* of the inner sense theory and the controversial supposition in question.

4.6 Revisiting the First Step of the Two-Step Argument

The result of the previous section is that the method subserving confabulation (whatever its exact nature) undermines [Auxiliary Assumption]. At this point, however, I see no reason to resist concluding that this method just *is* that method which subserves mindreading. To deny this, is to claim that there exist two distinct indirect, interpretive methods for attributing mental states, one subserving mindreading and the other subserving confabulation. While this is possible, I find the alternative much more plausible. I will thus proceed on the assumption that the method subserving mindreading satisfies conditions (i) and (ii), and so undermines [Auxiliary Assumption].

Recall the importance of undermining [Auxiliary Assumption]. Without it, Engelbert and Carruthers cannot justify the predictive step of their two-step argument. Specifically, because [Auxiliary Assumption] is false, and because the inner sense theory is compatible with its falsity, the inner sense theory does not predict that there should exist individuals who are incapable of self-attributing certain types of mental states (but not others). For this reason, contra Engelbert and Carruthers, the apparent absence of such individuals does not disconfirm the inner sense theory via the two-step argument. Engelbert and Carruthers miss their mark.

However, as I noted in Section 4.3.3, there are other kinds of first-personal deficits/impairments, and thus other ways of instantiating the first, predictive step of the two-step argument. So far I have blocked only one such way. Perhaps, then, there are other ways that succeed. In the remainder of this section I will adduce various considerations that undermine several seemingly plausible ways of instantiating this predictive step. In each case, keep in mind that Nichols and Stich's prediction will be

borne out by any first-personal deficit that exists in the absence of any third-personal deficits.[50]

4.6.1 Mindreading is Reliable

I argued a few paragraphs back that the method argued for in Section 4.5 is likely just that method which subserves our ability to mindread others. I assume that most believe that mindreading others is a fairly reliable process; the deliverances of the mindreading faculty are true often enough. Mindreading oneself should thus be at least as reliable. Indeed, there is reason to think that it should be even *more* reliable. As many philosophers and psychologists have previously noted, there is a much greater bank of relevant behavioral evidence when one's mindreading target is oneself.[51]

Of course mindreading oneself delivers the *wrong* answers in confabulation studies, and this might appear to cast doubt upon the reliability of mindreading. But recall, from Section 4.5, the reasonableness of confabulated self-attributions. In the Wegner and Wheatley study, the confabulated intentions were perfectly reasonable, given the behavior and context of the participants.

Presumably, the self-attributions made in confabulation studies are false because of the nature of the experimental setups. These studies are designed so that there is a mismatch between the true explanation of one's behavior and the mentalistic explanation of one's behavior that is seemingly most reasonable. In normal circumstances this mismatch is presumably absent, and thus such self-attributions would most likely be true.

---

[50] Again, I will consider the alleged first-personal deficit that Nichols and Stich argue bears out their prediction in Section 4.7.1.

[51] Gilbert Ryle (1949) has famously made this point, although not in terms of 'mindreading'. More recently, Carruthers (2009, 2010) has used an updated version of this thought in defense of his account of self-knowledge.

And this is what is important for my purposes, for these will be the typical situations of those with damaged inner sense.

Accordingly, if the inner sense theory is true, those individuals with damaged inner sense will nevertheless have a reliable means of self-attributing mental states.[52] This point is significant, for it shows that the inner sense theory does not predict individuals who unreliably self-attribute mental states. This blocks another way of instantiating the first step of the two-step argument.

4.6.2 Less Reliable Self-Attributions Would Likely Go Unnoticed

Consistent with the reliability of mindreading oneself is the possibility that this method is *less* reliable than any alleged mechanism of inner sense. And, indeed, this is likely the case. Although the inner sense theory downplays the specialness that self-knowledge is often taken to exhibit, self-knowledge by inner sense is standardly thought to be quite reliable. After all, the theory likens self-knowledge by inner sense to perceptual knowledge, which is quite reliable. Those with damaged inner sense should thus be expected to self-attribute mental states less reliably than those with unimpaired inner sense. Nevertheless, there are good reasons to think that this would go unnoticed.

First, there is a presumed authority that we grant others with respect to their self-attributions. When what is at issue is the state of another's mind, we typically defer to the judgments of that individual. Consider, for example, the oddness of questioning another's claim to be depressed or to be thinking about one's day. Thus, many false self-

[52] Although if mindreading typically involves inner sense, then the reliability of mindreading in normal cases cannot be transferred to abnormal cases where inner sense is impaired. However, for the reasons given in section 4.2, I suspect that in the absence of inner sense, mindreading would still be reliable, even if less reliable than when inner sense is present. As I argued there, versions of the simulation theory that involve inner sense must nevertheless include non-simulationist components that are capable on their own of generating reasonable mental state attributions. A *decrease* in reliability is a possibility considered in the next sub-section.

attributions would likely go unnoticed.[53] Second, self-attributions can often be self-fulfilling. That is, even if one falsely self-attributes, for example, an intention to Φ, one might eventually come to form the intention to Φ in response to this (initially false) self-attribution. In such cases, the (initially false) self-attribution will appear to be true to both the self-attributor and others.

Together, these points suggest that on the assumption that damage to one's inner sense should be expected to decrease the reliability of one's self-attributions, this decrease should likely go unnoticed. Accordingly, the inner sense theory does not predict that those with damaged inner sense should stand out in virtue of their less reliable self-attributions. This way of instantiating the first step of the two-step argument thus also fails.[54]

4.6.3 Mindreading Oneself Need Not Feel Distinctively Third-Personal

I have argued that the method subserving mindreading could fill in for a damaged inner sense. Of course this method differs greatly from inner sense. As I have noted numerous times, inner sense is thought to be a direct, perception-like method for detecting one's own mental states. In contrast, the method subserving mindreading is indirect and interpretive. Intuitively, then, the operation of these methods should feel quite differently to their subjects.

---

[53] Deference is not an absolute rule; we can imagine situations where such questioning is appropriate. The idea is rather that individuals are taken to be *default* authorities on their minds. Some philosophers, interpreting Wittgenstein (1953), have suggested that such authority is partially constitutive of mental state self-attributions; see Wright (1989) for discussion along these lines.

[54] Incidentally, the points about deference and self-fulfillment are relevant to the discussion from Section 4.6.1. Suppose that, contrary to what was argued there, the method subserving mindreading is *not* reliable, and thus that those with damaged inner sense should often make false self-attributions. The points about deference and self-fulfillment show that this possible unreliability would not obviously be noticed.

The research on split-brain individuals described in Section 4.5 is relevant here. Recall that such individuals (or, more precisely, their left hemispheres) offer confabulated explanations of behavior caused by their right hemispheres. To see the relevance of these individuals to the current discussion, consider the following from Gazzaniga:

> It is interesting to note that while the patients posses at least some understanding of their surgery, they never say things like, 'Well I chose this because I have a split brain and the information went to the right, nonverbal hemisphere'. Even patients who have higher IQs than P.S., based on IQ testing, view their responses as behaviors emanating from their own volitional selves, and as a result they incorporate these behaviors into a theory to explain why they behave as they do (1995, 1394).

Apparently, these individuals do not detect their own confabulations. This is so despite their being aware that, due to their commissurotomies, they are likely to confabulate in certain situations.

Indeed, as Carruthers (2009) reports (via communication with Gazzaniga), these individuals were often reminded of their situations during the experiments. He writes that "[o]n a number of occasions testing was paused and the experimenter said something like, 'Joe, as you know, you have had this operation that sometimes will make it difficult for you to say what we show you over here left of fixation. You may find that your left hand [controlled by his right hemisphere] points to things for that reason, ok?' Joe assents, but then on the very next series he is back to the interpreter effect once again" (126). This strongly suggests that the method producing these confabulated explanations, although indirect and interpretive, does not feel that way to the patients. If it did, then, presumably, patients on the lookout for confabulations would notice them.

I see no reason to resist extrapolating these results to normal, *non*split-brain individuals. That is, these results suggest that in normal individuals, self-attributions that are clearly confabulated can nevertheless feel direct and non-interpretive. Denying this is to suppose, implausibly, that a commissurotomy changes the feeling of the indirect, interpretive method housed in one's left hemisphere.

Individuals forced to self-attribute mental states via their mindreading method should thus not be assumed to notice this fact. Like Gazzaniga's split-brain patients, many of their (confabulated) self-attributions may nevertheless feel direct and non-interpretive. Perhaps this should not be too surprising. After all, normal participants in confabulation studies presumably do not (always) feel as though they are interpreting themselves; presumably, the (false) mental explanations that they offer are often delivered with a confidence suggesting that they feel as though they are directly reporting on the contents of their own minds.[55]

I conclude, then, that the inner sense theory does not predict that those individuals with damaged inner sense should feel as though they have only an indirect, third-personal access to their own minds. Another way of instantiating the first step of the two-step argument is thus undermined.

4.6.4 Motivated Mindreaders Will Issue Direct Self-Attributions

Inner sense, in virtue of being perception-like, is capable of detecting mental states *directly*, in the absence of relevant behavioral evidence. Presumably, though, this is not the case for the indirect, interpretive method subserving mindreading. For this reason,

[55] Indeed, I believe that mindreading *others* can often feel direct and non-interpretive. Sometimes, as it were, you can simply "see" that another is sad, or skeptical, or desiring food, etc. I will not, however, rely on this claim in what follows.

individuals with damaged inner sense might be thought to refrain from self-attributing mental states in the absence of such evidence. This is incorrect. There are various reasons to suppose that those with damaged inner sense should be expected to issue direct self-attributions.

In the event that an individual with damaged inner sense were motivated to answer an inquiry about his or her mind (posed by herself or another), I suspect that the individual's mindreading method would become activated. This method would step in to perform the task that inner sense could not. Presumably, for the inner sense theorist, something like this is what happens when humans confabulate.[56] The issue, then, is whether the absence of any relevant behavioral evidence would prevent the mindreading method from issuing a self-attribution. I think there are reasons to expect that it would not.

The method subserving mindreading presumably has access to *the contents* of much of what one believes, including one's memories.[57] When mindreading others, we obviously use beliefs/memories concerning both their past mental states and behavior; such beliefs about one's spouse, for example, surely influence how one views his or her mental life. When a mindreading target exhibits no behavior that could assist the mindreading process, such beliefs will obviously be quite useful. Indeed, some of these beliefs might concern past mental state attributions made in the presence of relevant

[56] Consider, for example, the participants from Wegner and Wheatley's study. They did not intend their stops in the forced-stop trials, and thus no such intentions were present for inner sense to detect. I suppose an inner sense theorist will claim that this is why the participants turned to their interpretive method.
[57] The emphasis on 'contents' is important. The mindreading method cannot be assumed to have access to what one believes or remembers *as such*.

behavioral evidence. When the mindreading target is oneself, the number of relevant beliefs/memories about one's past will be great.

Imagine, for example, an individual with a completely non-functioning inner sense. Suppose he wonders whether he intends to stay at his current job, while exhibiting no behavior relevant to whether he has the intention in question. Further, suppose that he has memories of past behavior that are relevant to this question. These behaviors might include his constantly keeping tabs on job openings at other firms, his complaining to his friends and family about his co-workers, etc. These memories could provide the basis for his (reasonably) self-attributing an intention to *not* stay at his job.

There is, however, a seemingly more difficult case to consider. This case concerns *occurrent thoughts*. Imagine an individual sitting quietly in her chair, exhibiting no overt behavior whatsoever. Further, imagine that she wonders (for whatever reason) what she is occurrently thinking about. This case is more difficult, given that past behavior seems irrelevant here. In the previous example, the mental inquiry concerned an intention, which is a *standing state*. Standing states should be expected to influence behavior over the course of time that they are present. For this reason, past behavior is relevant when self-attributing a standing state. In contrast, occurrent thoughts, in virtue of their transitory nature, are seemingly disconnected from past behavior in this way. If the individual exhibits no overt behavior relevant to the inquiry, then her mindreading method would seemingly be at a loss.

But this is incorrect. Even if occurrent thoughts are not connected to past behavior in the way that standing states are thought to be, knowledge of past behavior (or one's past more generally) could provide the basis for plausible-sounding, *even if false and*

unjustified, hypotheses about one's occurrent thoughts. For example, suppose that our individual, prior to retiring to her chair for the evening, has been working hard in her office on a project for work. In light of this fact about her past behavior, the hypothesis that she is sitting quietly thinking about this project is plausible-sounding, even if false and unjustified.

Similarly, suppose she has an important presentation at work the next day. This fact about her future could provide the basis for the plausible-sounding, even if false and unjustified, hypothesis that she is sitting quietly thinking (worriedly) about her upcoming presentation. If someone else were forced to mindread this individual in this context, I suspect that these would be the sorts of hypotheses that he or she would generate (assuming the mindreader has access to the facts cited above). To expect such hypotheses from an individual mindreading herself is thus not unreasonable.[58]

Humans are expected, and consequently motivated, to know what is happening in their own minds. Accordingly, it is very likely that an individual, when pressed, will use all of her available resources to come up with a plausible-sounding report of what is going on in her mind. If the inner sense theory is true, those without a functioning inner sense should thus be expected to use their mindreading method on themselves in such situations. The current point is that one's mindreading method, in virtue of having access to much of what one believes and remembers, will be capable of forming plausible-sounding hypotheses about one's mental states (standing or occurrent) even in the

[58] I hasten to point out that I am *not* claiming that normal human beings self-attribute occurrent thoughts in this manner. Rather, I am claiming that, if the inner sense theory is true, then individuals with damaged inner sense could self-attribute such thoughts in this way.

absence of relevant behavioral evidence. The inner sense theory thus does not predict individuals who refrain from self-attributing mental states in such contexts.[59]

4.7 The Second Step of the Two-Step Argument

The previous section contains four arguments, each of which shows that an apparent prediction of the inner sense theory is *not* a prediction of that theory after all. This is in addition to my having already blocked the prediction from Engelbert and Carruthers' argument in sections 4.4 and 4.5. In total, then, I have blocked five ways of instantiating the first step of the two-step argument.

However, these arguments accomplish more than the mere blocking of various ways of instantiating the predictive step of the two-step argument. Collectively, they show that individuals with damaged inner sense should appear to be quite normal to both themselves and others, if they have intact mindreading. First, such individuals should be able to self-attribute any kind of mental state that they can attribute to others. Second, they should be able to do so reliably. Third, although they should do so less reliably than their unimpaired peers, this difference should go unnoticed. Fourth, there is evidence that their self-attributions should feel direct and distinctively first-personal. Finally, they should issue plausible-sounding self-attributions in the absence of relevant behavioral evidence.

These points significantly complicate the second step of the two-step argument, i.e., the step claiming that the prediction from the first step is or is not borne out. Because

[59] Finally, notice that although these plausible-sounding self-attributions would most often be false, this would likely go unnoticed for the reasons given in Section 4.6.2. Moreover, those reasons apply even more strongly here. Consider the point about deference. On what possible grounds could one reasonably deny another's claim about what she is occurrently thinking, when, by hypothesis, she is exhibiting no overt behavior? Deference is all the more appropriate in this kind of case.

individuals with damaged inner sense will appear to be quite normal, those first-personal deficits/impairments that *are* legitimate predictions of the inner sense theory will be quite subtle and consequently difficult to identify. Justifying the second step of the two-step argument will thus be quite difficult.

Take, for example, the legitimate prediction (noted in Section 4.6.2) that those with damaged inner sense should self-attribute mental states less reliably than their unimpaired peers. Imagine the difficulties with trying to establish the existence of such individuals. I suppose there is no *in principle* difficulty here, but, for the reasons given in Section 4.6.2, the practical difficulties seem significant. If this prediction were borne out, it is unlikely that anyone would know; if this prediction were not borne out, it is unlikely that anyone would know.

These points have two important consequences. First, opponents of the inner sense theory who wish to deny the existence of individuals with damaged inner sense cannot simply point out that such individuals are not known about. Because of the subtlety of those first-personal deficits that are genuinely predicted by the inner sense theory, a much more thorough investigation is needed. Second, proponents of the inner sense theory who wish to affirm the existence of individuals with damaged inner sense cannot simply expect that such individuals will be known about, if they exist.

The main point, then, is that successfully employing the two-step argument requires much more care and thoroughness than perhaps was previously thought. Significant difficulties exist at each step.

4.7.1 The Second Step of Nichols and Stich's Argument

This is a natural place to consider the second step of Nichols and Stich's argument, for it may appear to be at odds with the points just made. I will ultimately argue that it is not, but, first, I must briefly explain the apparent tension.

Recall the first step of their argument. They claim that the inner sense theory predicts individuals with first-personal, but not third-personal, deficits. They then go on to claim that schizophrenics with passivity experiences confirm this prediction. There are three types of passivity experience: (i) delusions of control, the belief that one's bodily movements are controlled by external forces;[60] (ii) thought insertion, the belief that the thoughts in one's mind are not one's own;[61] and (iii) thought withdrawal, the belief that one's thoughts are extracted from one's mind.[62] In short, Nichols and Stich suggest that these problems stem from an inability to detect one's own intentions.

Here, they are relying on the work of the psychologist Chris Frith (1992, 1994). The explanation applies most directly to delusions of control. The idea is that if one is unaware of the intentions that give rise to some of one's bodily movements, then those movements will feel unintentional, as though they are under external control. Because schizophrenics with passivity experiences (allegedly) have intact mindreading, Nichols and Stich claim that damage to inner sense is responsible for their problems detecting their own intentions.[63]

---

[60] "When I reach my hand for the comb it is my hand and arm which move, and my fingers pick up the pen, but I don't control them … I sit there watching them move, and they are quite independent, what they do is nothing to do with me …" (from Mellor 1970, 18).
[61] "Thoughts come into my mind from outer space" (from Frith et al. 2000a, 358).
[62] "I am thinking about my mother, and suddenly my thoughts are sucked out of my mind by a phrenological vacuum extractor, and there is nothing in my mind, it is empty" (from Mellor 1970, 16-17).
[63] On this point, they disagree with Frith (1992, 1994). He claims that schizophrenics are unaware of some of their own intentions because of damage to mechanisms *involved in* mindreading.

This is quite odd, given what I have argued so far. I have argued that those with damaged inner sense should appear to be quite normal, if they have intact mindreading. But the schizophrenics that Nichols and Stich present are anything but seemingly normal. If these individuals have intact mindreading, then why should damage to their inner sense wreak such havoc? Fortunately, there are two problems with Nichols and Stich's argument, each of which helps to show why schizophrenics with passivity experiences do not pose a problem for my arguments. I will now offer a brief sketch of each.[64]

First, whether these individuals have intact mindreading is debatable. Ben Wiffen and Anthony David (2009) offer a brief survey of the research on mindreading abilities in schizophrenics. They write that "[a] deficit in mindreading is clearly demonstrable in schizophrenia", citing a meta-analysis on the various research into this question (Sprong et al., 2007). This meta-analysis surveyed studies on schizophrenics of all types, not just those with passivity symptoms or paranoia, finding a large and statistically significant impairment with respect to mindreading. Sprong et al. conclude from their meta-analysis that mindreading impairment is a trait marker of schizophrenia. They also note the very small sample sizes from the studies cited by those, like Nichols and Stich, claiming that schizophrenics with passivity experiences have intact mindreading.[65]

This research casts doubt on whether schizophrenics with passivity experience have intact mindreading. If they do not, then their existence is not puzzling against the background of my arguments. All of my arguments have been based on there existing a "backup method" of the kind argued for in Section 4.5. Individuals without such a

[64] For a more detailed discussion of these problems, see Carruthers (2009, section 9, 2011, 293-97).
[65] For example, they note that the sample size from Corcoran et al. (1995) was seven, and that the sample size from Pickup and Frith (2001) was just one.

method thus fall outside the scope of my arguments. More importantly, though, if they do not have intact mindreading, then they do not confirm the prediction from step one of Nichols and Stich's argument.

Second, a close examination of Frith's research on passivity experiences reveals that his account, which Nichols and Stich endorse, does not obviously trace these symptoms to an inability to monitor one's own intentions. Indeed, Frith et al. (2000b) explicitly deny this explanation, noting that this marks a change from Frith's earlier work cited by Nichols and Stich. More recent work on the topic, including but not limited to that of Frith, strongly suggests that passivity experiences are caused by problems in the motor control system and, specifically, problems with the forward modeling of motor commands. To see why, I must give a brief description of Frith's account of delusions of control.

Central to this account is Frith's model of the motor control system.[66] On this model, goal-directed movements involve three representations. First, there is a representation of the intended goal state; e.g., a representation of one's right hand holding an apple. Second, there is a representation of the estimated current state of one's body; e.g., a representation of one's right hand being inside one's pocket. On the basis of these two representations and, specifically, the differences between them, the system's "controller" generates a sequence of motor commands that, if successful, will cause one to move one's body in a way that satisfies the intended goal. This sequence is called the 'inverse model'. Third, on the basis of (an efferent copy of) the inverse model, the system's "predictor" generates what is called a 'forward model'. A forward model

[66] The following is based on Frith et al. (2000a, 2000b) and Frith (2012).

represents the predicted consequences of executing the inverse model.

According to Frith (2012), the forward model represents these consequences in terms of both kinematics and sensations; e.g., a representation of one's right hand holding the apple in such and such way, at such and such time, feeling such and such way. Finally, consider the following from Frith et al. (2000a) concerning the way in which the inverse and forward models interact with one another in the motor control system:

> Since inverse modelling may be less accurate than forward modelling it is possible for a discrepancy to be detected between the predicted and the desired consequences of the action before the action is actually generated. This is because the forward model may show that the movements based on the inverse model would not achieve precisely the goal required. In this case it is possible for the inverse model to be refined on the basis of the errors revealed by the forward model. A more appropriate sequence of commands can then be computed without any movements actually taking place (Frith et al. 2000a, 359).

Frith (2012) claims that passivity experiences in schizophrenics are the result of damage to this motor control system and, specifically, to the predictor charged with generating forward models. At times, he seems to suggest that the forward models are not generated. At other times, he seems to suggest that the forward models are generated, but that the subject is not aware of them. Either way, though, schizophrenics are not aware of an accurate forward model on this hypothesis. Without an accurate forward model, an agent will not anticipate the movements that he or she makes. As a consequence, these movements will be unexpected, and will thus *not* feel intentional. Movements that apparently achieve some goal will thus appear to be the result of the intentions of another, thereby causing delusions of control.[67]

[67] Frith (2012) admits that this account does not obviously apply to thought insertion and withdrawal, noting that this will depend on whether we can think of thought as an action. On the other hand, Carruthers (2009 and 2012) thinks that much occurrent thought occurs in inner speech. Because the production of

On this explanation, then, passivity experiences in schizophrenics are not due to an individual's being unaware of his or her intentions. Rather, the passivity experiences are due to damage in the motor control system and, specifically, the parts concerned with forward modeling. Inner sense is left completely out of the explanation. Now I of course admit that this might not be the correct explanation of passivity experiences. But the fact that there is an account of these experiences currently on offer that does not tie them to deficits in inner sense is significant. It weakens Nichols and Stich's claim that schizophrenics with passivity experiences support the inner sense theory. There is an explanation of such experiences that is consistent with the arguments from this chapter. I conclude, then, that the data presented in the second step of Nichols and Stich's argument are not at odds with any of my own arguments.

4.8 Conclusion

I have argued that there exists a method for attributing mental states that can be applied to oneself and is independent of inner sense; this method is that which subserves confabulation (and, likely, mindreading). Various features of this method, along with other considerations, show that numerous first-personal deficits are not predicted by the inner sense theory, thus undermining numerous instantiations of the first step of the two-step argument. Moreover, because of the apparent normalcy of those with damaged inner sense, first-personal deficits and peculiarities that are genuinely predicted by the inner sense theory will be quite subtle and difficult to detect, thus significantly complicating the second step of the two-step argument. The existence of the method that subserves

---

speech (inner or outer) *is* an action, Carruthers regards the Frith account as applying to thought insertion and thought withdrawal (for thoughts realized in inner speech).

confabulation (and mindreading) seriously complicates matters, presenting difficulties at each of the two steps.

Let me emphasize, however, that I have considered only those arguments for or against the inner sense theory that derive a certain kind of prediction from the possibility of damage to the mechanism of inner sense. My arguments are silent with respect to the cogency of other related argument. For example, because inner sense, if it exists, cannot be used (on its own) to attribute mental states to others, problems analogous to the ones that I have offered in this paper do not obviously arise for a closely related two-step argument. Specifically, an individual suffering damage to his mindreading method, but not to his inner sense, *should* be expected to exhibit third-personal deficits in the absence of first-personal deficits; or at least nothing that I have said in this chapter suggests otherwise.

In conclusion, then, my aim in this chapter has been to demonstrate the *misuse* of a particular kind of argument used both for and against the inner sense theory. The alleged evidence cited in this kind of argument is often not evidence for or against the inner sense theory at all.

Chapter 5: The Acquaintance Theory

5.1 Introduction

My focus in this chapter is the acquaintance theory of introspective knowledge. Philosophers who have recently defended this theory include Laurence BonJour (2003), David Chalmers (2003), Richard Fumerton (1995), and Brie Gertler (2001, 2011a, 2012); Bertrand Russell (1912) was an early proponent of this theory. The goal in this chapter is primarily to describe and critically examine the theory. I examine Brie Gertler's (2011a) recent and thorough presentation of it, raising numerous concerns along the way (5.2-5.6).[68] In addition, I argue that the theory is unable to explain the privileged and peculiar access that we seem to have to our propositional attitudes (5.7).

5.2 Knowledge by Acquaintance – Two Requirements and a Core Epistemic Claim

The acquaintance theory of introspective knowledge claims that we are, or can be, acquainted with *some* of our own mental states. Moreover, it claims that this relation of acquaintance, plus other conditions, can be exploited to achieve especially secure self-knowledge. In this section I discuss two of the central requirements for achieving introspective knowledge by acquaintance; I also discuss the epistemic claim at the core of the theory.

5.2.1 Requirement One - Acquaintance

The acquaintance theory claims that at least some of our introspective knowledge involves our being acquainted with the mental states introspected. The acquaintance relation is a metaphysically direct relation. As Bertrand Russell describes it, "we have acquaintance with anything of which we are directly aware, without the intermediary of

[68] Unless otherwise specified, all references to Gertler in this chapter are to Gertler (2011).

any process of inference or any knowledge of truths" (1912, 73). Acquaintance is thus a relation that involves no intermediary between the subject and the object of acquaintance.

For this reason, perception does not afford us acquaintance with objects in the external world. Arguably, a subject perceives objects in the world only via representations of those objects, where these representations are the products of a very complicated *causal* process between the objects perceived and one's sense organs.[69] That there is a causal process that mediates between the objects of perception and perceptual experiences shows that perception is not metaphysically direct. Acquaintance is a non-causal relation. This is perhaps the most significant difference between the acquaintance and inner sense theories. While the latter likens our method for attaining privileged and peculiar self-knowledge to perception, the former explicitly distances itself from perception.

Importantly, where there is an intermediary, so too is there an opportunity for error. If my awareness of a tree is mediated via a complex causal process involving the tree itself, rays of light, my retina, etc., then there are various points at which something could go wrong, thus leading to an illusory experience of the tree. By contrast, because the relation of acquaintance lacks such intermediaries (i.e., is metaphysically direct), errors of this kind cannot occur.[70]

This difference in the kinds of error that are possible suggests a test for detecting acquaintance. This test, which is Cartesian in spirit, was used by Russell (1912) to

[69] This is not to presume the sense data theory. Perceiving the world via representations does not entail perceiving the world by perceiving sense data of the objects in the world.

[70] This is not to say that there is no room for error at all. Gertler repeatedly claims that the acquaintance theory does not entail that our introspective beliefs are infallible. More on this below.

determine whether one's awareness of something is direct/unmediated, and thereby underwritten by acquaintance. If one's awareness of x is underwritten by acquaintance, then there is no room for the kinds of error just discussed. For this reason, Russell claims, one should not be able to doubt that one is truly aware of x. On the other hand, if one's awareness of x is not underwritten by acquaintance, then there *is* room for error, and so there is room to doubt that one is truly aware of x.

Because one's perception of a tree is causally mediated, one can doubt that the tree is as it is represented in perception. One must simply imagine that there is an error somewhere in the causal process leading from the tree to the perceptual experience. Suppose, though, that one is acquainted with one's perceptual experience of the tree. Because one bears a non-causal relation to the perceptual experience, one cannot achieve doubt in this way.

This test strikes me as quite implausible. Surely one can mistakenly believe that one's awareness of x is *not* metaphysically direct. If so, then one might reasonably doubt that one is truly aware of x; one must simply imagine that something has gone wrong in whatever process one (mistakenly) thinks is intermediary between x and one's awareness of x. Determining whether one can doubt that one is truly aware of an object x thus seems to depend on one's belief about one's relation to x. But this relation is precisely what the test is supposed to reveal!

Perhaps, then, Russell should be read as claiming that if one's awareness of x is underwritten by acquaintance, then one cannot *reasonably* doubt it, where 'reasonably' involves not having mistaken beliefs about one's relation to x. But, again, this seems to

require knowledge of (or at least true belief about) the very thing that the test is after, namely, one's relation to x.

5.2.2 Requirement Two - Conceptualization

Introspective knowledge by acquaintance, in addition to requiring the metaphysically direct relation of acquaintance, also requires the deployment of concepts. As William James put the point, "[in self-knowledge the mental state] must be more than experienced; it must be remembered, reflected on, named, classed, known, related to other facts of the same order" (James 1884, 1). James is here claiming that in order to have introspective *knowledge* of a mental state, it is not enough to merely be acquainted with it. In addition, one must make a judgment to the effect that one is in that state. Because a judgment involves the deployment of concepts, one must conceptualize the presence of the mental state.

Interestingly, this point was denied by Russell. He distinguished two kinds of knowledge, 'knowledge of things' and 'knowledge of truths', and regarded only the latter to require the deployment of concepts. He believed that introspective knowledge was knowledge of things. Contemporary acquaintance theorists, including Gertler, deny this. They side with James on this issue.

According to the contemporary acquaintance theory, then, the picture is as follows. As Russell believes, introspective knowledge is sometimes grounded in the metaphysically direct relation of acquaintance; this relation sometimes holds between a subject and some of his or her mental states. However, contra Russell, the holding of this direct relation does not suffice for introspective knowledge. In addition, one must

conceptualize the object of acquaintance. These are the two necessary conditions at the heart of the acquaintance theory.

These conditions are not, however, jointly sufficient for attaining introspective knowledge by acquaintance. As should be obvious, introspective knowledge of any kind (including knowledge by acquaintance) also requires that the relevant conceptualization be correct and justified. Much more will be said about this point below.

5.2.3 The Core Epistemic Claim

The acquaintance theory's core epistemic claim is that the acquaintance relation allows our introspective judgments based on acquaintance to achieve an especially secure epistemic status. As Gertler puts the point:

> The core of contemporary acquaintance accounts derives from Russell's claim that one has metaphysically direct access to (some of) one's own mental states, and that this access provides for strongly justified, non-inferential judgments concerning those states. Any account of self-knowledge that accepts this claim may be plausibly regarded as an acquaintance account (94).

This claim connects the two requirements just discussed. According to this claim, the judgments required by the second requirement are able to achieve an especially strong justificatory status. That they can attain this status is made possible by virtue of the metaphysically direct acquaintance relation required by the first.

5.3 Knowledge by Acquaintance and Justification

How, though, does the metaphysical directness of the acquaintance relation make possible such an epistemic payoff? How does the relation of acquaintance serve to justify, or even help to justify, the required introspective judgments? According to Gertler, "[t]he

main task of acquaintance accounts is to explain how acquaintance with a mental state can *justify* the corresponding introspective judgment – e.g., that *pain is present*" (96).

5.3.1 A Potential Roadblock – Davidson's Challenge

Donald Davidson (1983) argues that this task cannot be achieved. To see why, it will be helpful to have a simple example on hand. Imagine that S is experiencing a dull pain in his left elbow. Further, imagine that S is acquainted with this pain and forms the introspective judgment that the pain is present. According to Davidson, this pain cannot help to justify the introspective judgment, given that the pain is non-propositional while the judgment is propositional. And this is the case for all sensations, not just pains.

Davidson here conceives of justification as a logical relation. Because he denies that a non-propositional state can stand in a logical relation to an introspective judgment (or anything else for that matter), he also denies that a non-propositional state can help to justify an introspective judgment. But notice that the acquaintance theory does not claim that a sensation, by itself, justifies an introspective judgment. Rather, the theory claims that one's *acquaintance* with a sensation justifies, or helps to justify, the relevant introspective judgment.[71] Does this point undermine Davidson's objection to the acquaintance theory?

It does not. Acquaintance is a kind of awareness (i.e., it is a direct awareness). But as Gertler points out, an awareness is non-propositional, "[f]or it is an *event*, and not the type of thing that could be true or false … *that the event occurred* may be true, but the event itself has no truth value" (98). Thus, although the acquaintance theory claims that

[71] As Gertler writes, "[o]n [the acquaintance] theory, the presence of pain helps to justify my introspective judgment that I am in pain. This justification occurs by way of *my awareness of* my pain" (71, my emphasis). Gertler is here speaking of direct awareness by acquaintance.

S's acquaintance with the dull pain, not the dull pain itself, is that which helps to justify S's introspective judgment, the justifier remains non-propositional. If, as Davidson claims, a judgment can be justified only by a propositional state, then S's acquaintance with his pain cannot help to justify his introspective judgment that the pain is present.[72]

In response, Gertler argues that Davidson's conception of justification is too demanding. If he were right that only propositional states can contribute to justification, then many introspective judgments that are seemingly justified are, in fact, unjustified. Take the judgment *I am now experiencing an itch*. Surely, such a judgment could be (and often is) justified. Suppose, then, that this judgment is justified. If Davidson is right, then this judgment is justified by some propositional state. Moreover, assuming it is not justified by itself, it must be justified by some *other* belief or judgment.[73]

Gertler is unable to identify such a belief or judgment. She claims that one's awareness of the itch *on its own* seems to justify the introspective judgment; no other belief or judgment is required. Of course if Davidson is correct, then this is mistaken; there must be some belief or judgment that justifies the introspective judgment. One

---

[72] Interestingly, Gertler claims that Davidson's constraint on justification is inconsistent with not only the acquaintance theory, as well as other accounts committed to internalism about justification, but also with various accounts committed to externalism about justification. For example, she claims that it is incompatible with process reliabilism, the view according to which a judgment is justified if and only if it is formed through a reliable process. She notes that, like being acquainted with a sensation, being formed by a reliable process is an *event*, and so is not a propositional state. I wonder, though, if this point is correct. Specifically, it seems that Davidson's point is concerned only with the 'reason-giving' sense of justification. That is, his claim seems to be that only propositional states can serve as reasons for judgments. But it is plausible that the process reliabilist is working with a 'non-reason-giving' sense of justification, and so it is plausible that Davidson's challenge does not apply to this account (or others like it).

[73] This is not to suggest that beliefs and judgments are the only propositional states. They are not. Propositional attitudes of all kinds are propositional states. The idea is simply that, among the class of propositional states, beliefs and judgments are the only ones capable of justifying other beliefs and judgments. Along these same lines, propositional attitudes that are not beliefs and judgments can stand in logical relations to beliefs and judgments. Their logical form does not disqualify them from justifying beliefs and judgments. Rather, it seems to me, they are disqualified due to their attitudinal component.

candidate that Gertler considers is a belief or judgment with the content 'in normal circumstances, if it seems to me that I have an itch then I have an itch' (99). Perhaps this connecting belief or judgment, in combination with one's awareness of the itch, serves to justify the introspective judgment.

Gertler denies this, for this suggestion does nothing to make the awareness of the itch propositional. Because the connecting belief or judgment cannot by itself justify the introspective judgment, the awareness of the itch seemingly needs to make a justificatory contribution. Because Davidson claims that it cannot make such a contribution, due to its being non-propositional, he cannot account for the justification of the introspective judgment in this way.[74]

Another possibility is that the awareness of the itch takes the form of a belief, namely, the belief with the content 'it seems to me that I have an itch'. In this way the awareness of the itch is "transformed" into the right kind of state, namely, a propositional state (with the right kind of attitudinal component, see footnote seventy-two). However, this belief can contribute to the justification of the introspective judgment only if it itself is justified.[75] But what justifies it? Gertler points out that, if Davidson is correct, it can be justified only by another propositional state, and thus *not* by the itch itself, nor by an awareness of the itch. At this point, though, whether such a propositional state is available is unclear.

---

[74] Gertler gives an additional reason for rejecting the possibility that the introspective judgment is justified, in part, by the connecting belief with the content 'in normal circumstances, if it seems to me that I have an itch then I have an itch'. She claims that while a young child can seemingly justifiably believe that he or she has an itch, it is unlikely that the child has any beliefs as complicated as the suggested connecting belief.

[75] Presumably, an unjustified propositional state cannot contribute to the justification of any propositional state.

Gertler thus concludes that there is no propositional state that justifies the introspective judgment *I am now experiencing an itch* (or that if there is such a state, it will face the same justificatory problem as the introspective judgment). Because, intuitively, this judgment is justified, she concludes that Davidson's justificatory requirement is too strong; she denies that only propositional states can justify judgments (introspective or not). This thus opens the door for the possibility that one's acquaintance with a sensation can help to justify an introspective judgment about that sensation.

I have one concern with Gertler's response to Davidson's challenge. She claims that "transforming" the awareness of the itch into a belief with the content 'it seems to me that I have an itch' does not help matters, given that this new belief is also in need of justification. The problem is that she gives no reason for thinking that it cannot receive the needed justification. Notice that pointing out that this line of reasoning will lead to an infinite regress of propositional justifiers is not sufficient. Whether it will is unclear. Or at least it is unclear *if* one has not ruled out the coherentist option.

In other words, this aspect of Gertler's criticism seems to reveal an implicit rejection of coherentism about justification. While she may be correct in this rejection, this option should not be ignored. Moreover, this is so even if coherentist justification is, for whatever reason, unavailable to the acquaintance theorist. After all, she is arguing that Davidson's constraint on justification is incorrect, not merely that it should be rejected by non-coherentists. For these reasons, Gertler's point is not decisive.

Nevertheless, suppose that Gertler's rejection of Davidson's constraint on justification succeeds. That is, suppose that acquaintance with a mental state is not barred from contributing to the justification of an introspective judgment about that state. There

of course remains the task of showing *how* it can make such a contribution. This task is the focus of the next section.

5.3.2 Explaining Justification by Acquaintance

The acquaintance theory is committed to internalism about justification.[76] As Gertler puts it, this is the view according to which "epistemic justification involves having an internal *reason* for one's belief, perhaps in the form of evidence. Some take internal reasons to be those within the mind … But the more standard construal of internalism takes them to be *accessible* reasons" (12). In contrast to this view, is externalism about justification: "epistemic externalists deny that knowledge requires accessible reasons or evidence. A true belief can count as knowledge so long as it is appropriately connected to the facts it concerns" (12).

Suppose that S is acquainted with mental state *m* at time *t*. According to contemporary acquaintance theorists, S's introspective judgment that he is in *m* at *t* is justified via acquaintance if and only if S grasps the correspondence between that judgment and *m*. That is, S's acquaintance with *m* contributes to the justification of S's introspective judgment by virtue of S's *g*rasping (via his acquaintance with *m*) that *m* corresponds with the introspective judgment. As per the acquaintance theory's emphasis on directness, this grasping must be metaphysically direct, meaning that *S* must be

[76] This is a bit of an oversimplification. While Gertler writes throughout the chapter that the acquaintance theory is epistemically internalist, she notes at one point that "the acquaintance theory is itself strictly neutral about the nature of justification" (107). She goes on, however, to write that "contemporary acquaintance theorists are generally committed to the internalist thesis that a knowing subject must *grasp* what it is that justifies her judgment" (107). So although the theory is not strictly committed to internalism about justification, in practice it is so committed.

acquainted with both objects of the correspondence and the fact that these objects correspond with one another.[77]

The picture is thus as follows:

> A subject *S* has introspective knowledge by acquaintance of a mental state *m* if and only if S directly grasps the correspondence between S's introspective judgment that *m* is present and *m* itself. This directness requires that:
>
> (i) S is acquainted with *m*;
>
> (ii) S is acquainted with S's introspective judgment that *m* is present; and
>
> (iii) S is acquainted with the correspondence between the introspective judgment and the mental state it is about.

The following quote from Richard Fumerton, a contemporary proponent of the acquaintance theory, nicely captures these three conditions:

> My suggestion is that one has a noninferentially justified belief that *P* when one has the thought that *P* and one is acquainted with the fact that *P*, the thought that *P*, *and* the relation of correspondence holding between the thought that *P* and the fact that *P* (Fumerton 1995, 75).

What exactly does it mean for the introspective judgment and the relevant mental state to correspond with one another? Gertler approvingly cites Lawrence BonJour (2003) on this issue. According to BonJour, "a foundational belief results when one directly sees or apprehends that one's experience satisfies the description of it offered by the content

[77] The acquaintance theory's commitment to internalism about justification places an additional constraint on the account. Specifically, in order for the grasping of the correspondence between the target mental state and the introspective judgment to justify that judgment, this grasping must be accessible to the subject. And it seems that there is no real difficulty here, for it is difficult to see how a subject's grasping of anything might be inaccessible to the subject.

of the belief (191). BonJour seems to be claiming that an introspective judgment is, or has as one of its components, a description of the target mental state. That mental state corresponds to this judgment iff it satisfies the relevant description (i.e., iff the relevant description is true of the target mental state). A subject will thus satisfy condition (iii) iff she is directly aware of *m*'s satisfying the descriptive content given in the introspective judgment.

Notice that meeting conditions (i)-(iii) guarantees the truth of the introspective judgment. This is because one can meet condition (iii) only if *m* satisfies the description of it given in the introspective judgment, thus making that judgment true. In addition, Gertler claims that meeting conditions (i)-(iii) suffices for that judgment's being strongly justified. When conditions (i)-(iii) are met, the subject *directly* grasps the correspondence between the introspective judgment and its truth-maker. In this way, the truth of the judgment is not accidental.

The metaphysical directness of the relevant grasping also contributes to the strong justification. As discussed above, direct awareness precludes certain kinds of errors. For this reason, the acquaintance theory claims that introspective judgments are more strongly justified than perceptual judgments about the external world. While the theory allows that one can be acquainted with both introspective and perceptual judgments, it denies that one can be acquainted with the truth-makers of one's perceptual judgments, and thus also with the correspondence between one's perceptual judgments and their truth-makers. The acquaintance theory thus offers an explanation of the intuitive epistemic asymmetry between introspective and perceptual judgments.

5.3.3 A Potential Problem

Before moving on, I want to raise a potential problem with the current proposal. Gertler notes numerous times that the proposal just sketched is an answer to Davidson's challenge. Specifically, she claims that it explains how a non-propositional mental state can contribute to the justification of an introspective judgment. According to Gertler:

> Davidson envisaged two possible ways that a mental state could conceivably justify a self-attributing judgment: either by causing the judgment or by standing in a "logical" relation to it. The acquaintance theory provides an alternative picture of justification. On this picture, an introspective judgment is justified by the subject's directly grasping the correspondence between that judgment and the mental state it concerns. The mental state thus contributes to justifying the judgment, but its contribution does not involve causing or entailing the judgment. Instead, it involves *corresponding* to the judgment, and thereby rendering the judgment accurate, in a way that can be directly grasped by the subject (102).

I grant that Gertler has provided an alternative picture of justification. However, to answer Davidson's challenge she must do more than simply provide an alternative picture. She must also make the case that this picture is plausible.

To this end, she owes us an explanation of exactly what the requisite grasping amounts to. Presumably, the grasping is *not* propositional; if it were, the proposal would be consistent with Davidson's constraint that only propositional states can contribute to the justification of propositional judgments. In several places, though, Gertler writes as though the grasping *is* propositional. For example, she writes that "it is the subject's grasping *that* her introspective judgment corresponds to her current experience [that justifies an introspective judgment]" (100, emphasis added).

However, for the reason just given, this is probably not what Gertler intends; the requisite grasping should be taken to be non-propositional. But then the following

question arises: what does it mean to non-propositionally grasp a correspondence between two states? Arguably, the grasping of such a correspondence requires *conceptualizing* the two corresponding states and judging that the conceptualizations agree in certain respects. Such conceptualization, though, would seem to make the grasping propositional, thus suggesting that conceptualization is *not* part of the grasping. But then we are left without an understanding of the nature of the requisite (non-propositional) grasping. That is, we are left with an incomplete answer to Davidson's challenge.

5.3.4 Summary

Putting this worry aside, the three conditions put forth by Gertler, and other contemporary acquaintance theorists, are fairly straightforward. If one is comfortable with the relation of acquaintance, then one will understand what it takes to acquire introspective knowledge by acquaintance. However, as Gertler admits, *satisfying* these conditions might be quite difficult. She writes that:

> Meeting conditions [(i)-(iii)] is cognitively quite demanding. It requires paying careful attention not only to the target mental state, but also to one's judgment about that state and to the relation between these. And while acquaintance theorists can allow that knowledge by acquaintance is relatively rare, they are committed to saying that we do achieve it, at least occasionally. As we will see shortly, some critics allege that satisfying these conditions is simply beyond our cognitive abilities (103).

Gertler considers two such criticisms. The first is the problem of the speckled hen, which threatens to undermine the satisfiability of condition (i). The second is the problem of conceptualization, which threatens to undermine the satisfiability of condition (iii). I shall consider these problems in turn.

5.4 The Problem of the Speckled Hen

The acquaintance theory is best suited for sensations and perceptual experiences. The reason for this is that for such states there seems to be no gap between appearance and reality. But where there is no appearance/reality gap, there is also no room for doubt; if there is no gap between the appearance of something and that something's reality, then that appearance cannot fall short of the reality.[78] If this is correct, then our awareness of such states pass Russell's test and are thereby states with which we are (or can be) acquainted.

That such states appear to lack a gap between appearance and reality seems to motivate the acquaintance theory and, specifically, the claim that we can satisfy condition (i). As Gertler puts the point, "[t]he lack of an appearance/reality gap suggests that sensations pass Russell's doubt test. More to the point, it suggests that your relation to your sensations can be especially secure, epistemically, and metaphysically direct" (95).

Accordingly, if it could be shown that there *is* an appearance/reality gap for sensations and perceptual experiences, then this would undermine the claim that we are (or can be) acquainted with such mental states. And showing this, Gertler claims, is exactly what the speckled hen problem threatens to do. She writes that "[t]he problem of the speckled hen threatens the acquaintance theory by suggesting that there is an appearance/reality gap even for sensations" (103).[79]

---

[78] Suppose there is no appearance/reality gap for itches. On this possibility, if it appears that you are currently experiencing an itch, then you are, in fact, currently experiencing an itch, and, conversely, if you are currently experiencing an itch, then it appears that you are currently experiencing an itch.

[79] While I have always associated the problem of the speckled hen with Chisholm, Gertler writes that "[a]ccording to Chisholm (1942), this problem was formulated by Gilbert Ryle, in discussion with A.J. Ayer about Ayer's (1940) sense datum theory.

Imagine that you see a speckled hen pass in front of you; you have a visual experience of a speckled hen. Further suppose that the hen has 48 speckles. Assuming that you are seeing the hen in optimal viewing conditions, your experience will involve the phenomenal property *48-speckledness*.[80,81] Now suppose that you attend to this experience through introspection, and became aware of it through acquaintance, thereby satisfying condition (i). In being acquainted with this experience you are acquainted with the various phenomenal properties involved in it, and so are acquainted with the phenomenal property *48-speckledness*.

However, despite your acquaintance with this phenomenal property, you will not be able to judge through introspection that the experience involves this property. Your discriminatory powers are simply too weak; they are unable to detect with precision numbers this great. Thus, although you are able to have experiences involving *48-speckledness*, you are unable to justifiably introspect that your experiences involve this property.

---

[80] The phenomenal properties of, e.g., your perceptual experience of a zebra constitute what it is like phenomenologically for you to see the zebra. Your visual experience involves the phenomenal properties *black*, *white*, *striped*, *black-and-white striped*, etc. Similarly, when I have an itch, my sensation involves the phenomenal property *itchiness*, and this property (as well as any other properties that might be involved in it) determines the phenomenology of my itch.

[81] A few clarificatory points are in order. First, a phenomenal property is a qualitative property of a phenomenal experience or sensation. It is not a property of any external object that the experience or sensation might be about. Or at least this is how Gertler understands it; some representationalists about qualia seem to identify phenomenal properties with the properties represented by an experience (see, e.g., Dretske 1995). Moreover, Gertler (2001) requires that phenomenal properties be *non-relational* properties of experiences. (This requirement on qualia is fairly standard.) This requirement is certainly incompatible with representationalism about qualia since representation is a relational matter. Thus, Gertler's understanding of phenomenal properties is incompatible with the most popular naturalistic account of qualia currently on offer. And although I do not wish get into the details here, the acquaintance theory as a whole, not just Gertler's version of it, seems to endorse this requirement. For example, Chalmers (2003), in his defense of a version of the acquaintance theory, appears to endorse a view of phenomenal properties very similar to this. He calls this view 'phenomenal realism'.

This case seems to show that an experience need not appear as it is. The experience with which you are acquainted involves the phenomenal property *48-speckledness*, yet it does not appear this way to you, as evidenced by your inability to justifiably judge that it involves the phenomenal property *48-speckledness*. The speckled hen case thus appears to show that there is an appearance/reality gap for our perceptual experiences. But given this gap, one's awareness of such an experience does not pass Russell's test, thereby showing that one is not acquainted with that experience, i.e., condition (i) is not satisfied. Gertler characterizes the moral of the speckled hen case as follows:

> The speckled hen example shows that the appearance of a phenomenal property sometimes falls short of its reality. The phenomenal reality is *48-speckledness*, but our inability to recognize it as *48-speckledness* shows that it does not *appear* this way to an introspective glance. This example thus threatens an idea that motivated the acquaintance theory: that the appearance of a sensation directly and completely reveals its reality (104).

5.4.1 Responding to the Problem of the Speckled Hen

Gertler argues that the speckled hen case does not show that there is an appearance/reality gap for sensations and perceptual experiences, and so does not threaten to undermine the acquaintance theory. Her argument is based on the idea that there are two senses of 'appearance', one phenomenal and the other epistemic:

> Something's *epistemically* appearing a certain way, to a subject, generally inclines the subject to believe that it *is* that way. Something's *phenomenally* appearing a certain way is a matter of the phenomenal properties that are involved in experiencing it (104).

If your experience of a hen involves the phenomenal property *48-speckeldness* (which it will, if you are viewing the hen in optimal conditions and it has 48 speckles), then the hen *phenomenally* appears to you to have 48 speckles. On the other hand, the hen

will epistemically appear to you to have 48 speckles only if you are inclined to believe that it has 48 speckles. As noted above, because our discriminatory capacities are limited as they are (and, perhaps, also because we recognize this fact about ourselves), a 48-speckled hen will generally not epistemically appear to a subject to have 48 speckles.

Thus, in the case of the speckled hen, there is no gap between the phenomenal appearance and the reality; the phenomenal appearance of the hen *does* track the reality (i.e., the hen's having 48 speckles). There is only a gap between the epistemic appearance of the hen and the reality. Importantly, this is not to say that the phenomenal appearance of the hen must have tracked this reality. If viewing conditions are not optimal, then one's experience of the hen might involve the phenomenal property *47-speckledness*, even though the hen has 48 speckles. Hens need not appear (phenomenally or epistemically) as they are. And this is true of all external objects.

Of course Gertler wants to say that sensations and perceptual experiences are unlike hens in this respect. She maintains that while there is an epistemic appearance/reality gap for such mental states, there is no phenomenal appearance/reality gap for such states. How a sensation or perceptual experience phenomenally appears to a subject matches that sensation or perceptual experience's reality. And how such a state is in reality matches how it phenomenally appears to a subject.

If this is true, then the speckled hen case does not undermine the satisfiability of condition (i), for it fails to show that there is a phenomenal appearance/reality gap for sensations and perceptual experiences. But, as we shall see in Section 5.5, Gertler does think that the speckled hen case poses a slightly different threat to the acquaintance theory, one that she also thinks can be overcome.

5.4.2 Evaluating Gertler's Response

I have two concerns with Gertler's response to the problem of the speckled hen. First, I wonder whether the distinction she draws gets her what she needs. Second, I wonder what sense can be made of speaking of a sensation or perceptual experience's phenomenal appearance. As will become apparent, these issues are connected.

To begin, notice that the distinction between phenomenal and epistemic appearances on its own is not enough to show that condition (i) is satisfiable. The distinction makes clear that the speckled hen case might only show that there is a gap between an experience's epistemic appearance and reality, leaving open whether there is a gap between an experience's phenomenal appearance and reality. But this is obviously different than showing that there is no gap between an experience's phenomenal appearance and reality. Without showing this, however, what reason is there to believe that we can be acquainted with sensations and perceptual experiences?

I am not claiming that Gertler takes herself to have established that there is no gap between an experience's phenomenal appearance and its reality. I am merely pointing out that the response to the speckled hen problem just rehearsed does not establish this claim. Indeed, Gertler seems to think that this claim is highly intuitive and should be accepted by default. After pointing out that the speckled hen case is meant to suggest that there *is* an appearance/reality gap for sensations and perceptual experiences, she notes that we are nevertheless left with the feeling that there is no such gap. According to Gertler:

> And yet it is hard to deny that, when it comes to sensations, appearance *is* reality. In other words, it is strongly intuitive that phenomenal features are as they appear, and appear as they are. An experience that appears itchy (that presents, to the subject, the characteristic *itchy* feeling) really is an itch. And any experience that really is an itch will appear itchy (104).

The situation, then, appears to be as follows. The distinction between phenomenal and epistemic appearances allows the acquaintance theorist to side-step the challenge posed by the speckled hen case. Having done this, she can return to the (allegedly) intuitively compelling view that there is no (phenomenal) appearance/reality gap for sensations and perceptual experiences.

My second concern has to do with what it means to speak of a sensation or perceptual experience's phenomenal appearance. A subject, S, sees the speckled hen in virtue of having a perceptual experience involving various phenomenal properties. These phenomenal properties determine how the hen phenomenally appears to S. That is, the phenomenal properties of the experience of the hen determine the phenomenal appearance of the hen.

In introspection, however, the object of awareness is not the hen, but is rather the perceptual experience of the hen. Supposing that we do not have experiences of our own perceptual experiences, it is difficult to see how the distinction between phenomenal and epistemic appearances applies to the objects of introspection. The definition of 'phenomenal appearance' given above suggests that a sensation or experience's phenomenal appearance is a matter of the phenomenal properties of the experience of *it*, the sensation or experience. But if we do not experience our own experiences, then this does not make sense.[82]

[82] Although Gertler does not explicitly deny that we have experiences of our own experiences in Gertler (2011), she does deny this in Gertler (2001). There she writes that "one does not *perceive* one's sensations" (316). Moreover, this denial is quite common among philosophers of mind. However, if Gertler does think that we have experiences of our own experiences, then the present concern no longer applies. In that case, the phenomenal appearance of an experience would simply be a matter of the phenomenal properties involved in the experience of the experience. Of course this is not to say that there are no problems with the claim that we have experiences of our own experiences.

On the other hand, there seems to be no problem speaking of a sensation or perceptual experience's 'epistemic appearance'. One can (and does) form judgments about the character of such states. For this reason, a sensation or perceptual experience's epistemic appearance is simply a matter of the judgments one is inclined to make about the character of the experience.[83]

Perhaps we should identify an experience's phenomenal appearance with the phenomenal appearance of its object.[84] For example, if my experience of the hen with 48 speckles involves the phenomenal property *48-speckledness*, and so the hen phenomenally appears to me to have 48 speckles, then my *experience* of the hen also phenomenally appears to me to have 48 speckles.[85] At any rate, this is how I shall understand an experience or sensation's phenomenal appearance in what follows.

As I noted above, there is a connection between the two concerns just discussed. Notice that on my suggestion for how to understand a sensation or experience's phenomenal appearance, a positive case can be made for the claim that there is no phenomenal appearance/reality gap for sensations and experiences. Specifically, if the phenomenal appearance of an experience of a speckled hen is identical to the phenomenal appearance of the hen itself, and if, as Gertler claims, the phenomenal appearance of the hen itself is simply a matter of the phenomenal properties (truly) involved in the experience of the hen, then the phenomenal appearance of an experience of a hen is

---

[83] Gertler claims (via personal correspondence) that the sensation or experience's epistemic appearance is that which disposes one to make the judgments one is disposed to make.

[84] Cases of hallucination pose a problem for this proposal. I am inclined, however, to think that hallucinations have intentional objects. Whether Gertler would be happy with this, I do not know. If she would, then she could adopt this proposal.

[85] This suggestion fits nicely with the alleged t*ransparency* of experience. On one understanding of the transparency thesis, when one tries to attend to the properties of one's experience, one inevitably ends up attending to the properties of the objects in the external world that the experience is about. I am not here endorsing this transparency claim, but am merely noting its affinity with the interpretation just given.

simply a matter of the phenomenal properties (truly) involved in the experience itself. But if this is the case, then there is no room for the phenomenal appearance of an experience to diverge from the reality of that experience; an experience must phenomenally appear as it is, and must be as it phenomenally appears.

5.5 The Problem of Conceptualization

If the speckled hen case does not show that there is a phenomenal appearance/reality gap for sensations and perceptual experiences, then what does it show and, specifically, what challenge does it pose to the acquaintance theory? It clearly shows that epistemic appearances may fall short of phenomenal reality. Although the phenomenal reality of the experience of then hen is such that *48-speckledness* is present, it does not epistemically appear this way to the subject; one is not inclined to believe that *48-speckledness* is involved in one's experience. Importantly, though, this is not a problem for the acquaintance theory. This is because the theory is not committed to any kind of self-intimation thesis, according to which all features of phenomenal reality are believed to be present by the subject.

The example does, however, highlight the fact that our powers of discrimination severely limit the scope of our introspective knowledge. We simply cannot know via introspection that *48-speckledness* is present; we are in a sense blind to such properties of our experiences.[86] More important than this limitation, however, is the fact that such discriminatory limitations can make it difficult, or even impossible, to know whether our introspective judgments are in line with their mental targets. If one introspectively judges

[86] I write 'in a sense' because, as discussed, we are nevertheless phenomenally aware of such features.

that one's experience involves *50-speckledness*, but yet one's discriminatory capabilities fall far short of such a large number, one cannot determine whether the judgment is true.

Thus, in addition to being in some sense blind to certain aspects of our sensations and experiences, we are capable of stepping beyond our discriminatory capacities, making epistemically irresponsible judgments. What is worse is that for such judgments we have no way of determining whether they are correct; our discriminatory capabilities are unable to settle the matter. To see how this poses a problem for the satisfiability of condition (iii), let's consider some examples.

Suppose you form the judgment that your experience involves *48-speckledness*, and also suppose that you are acquainted with this judgment. Assuming your discriminatory capabilities are typical, you are unable to directly grasp the correspondence (or lack thereof) between the experience and the introspective judgment. You cannot tell, without carefully counting,[87] that your experience involves the phenomenal property *48-speckledness*, as opposed to *46-speckledness* or *47-speckledness*, or even *35-speckledness*.[88] You are thus unable to grasp the correspondence (or lack thereof) between your experience and your introspective judgment about the experience. For this reason, you will fail to have introspective knowledge of this experience (even if the experience and judgment, in fact, correspond).

---

[87] As Gertler notes, counting relies on memory, which is a *causal* process, and thus acquaintance with a property cannot be underwritten by memory. Recall that acquaintance is a non-causal relation and that it is in virtue of this feature that judgments based on acquaintance are thought to be especially epistemically secure. For this reason, one cannot be acquainted with a phenomenal property such as *48-speckledness* via counting the phenomenal speckles involved in one's experience.

[88] Some people might be able to make such discriminations (Gertler mentions the autistic savant from the movie *Rainman*), but such a person would be highly atypical. Moreover, it seems reasonable that any such person would have experiences involving phenomenal properties for which her discriminatory capabilities are not equipped. If so, then the point made above applies to us all.

Admittedly, this is not true of all phenomenal properties. For example, suppose you have an experience involving *2-speckledness*. Because you are able to discriminate with precision numbers this low, you are able to determine that your experience involves the phenomenal property *2-speckledness*, rather than *1-speckledness* or *3-speckledness*. You are thus able to directly grasp whether your experience involving *2-speckledness* corresponds with your introspective judgment about that experience.

The problem of the speckled hen shows that a subject must appropriately limit her conceptualizations of her sensations and perceptual experiences; she must limit them to phenomenal properties that fall within her discriminatory capabilities. By not doing so, she risks putting herself in a position where she is unable to directly grasp the correspondence (or lack thereof) between her introspective judgment and the mental state it is about. So understood, the problem of the speckled hen poses a problem for the satisfiability of condition (iii). Gertler puts what remains of the challenge as follows:

> [The speckled hen case] shows that the extent of our introspective knowledge is limited by our powers of discrimination. The acquaintance theorist must explain how introspective subjects can respect those limits. In grasping the correspondence between a judgment and its truthmaker, how can one ensure that one's conceptualization of a phenomenal property as *pain* or *many-speckledness* does not outstrip one's powers of discrimination" (106).

Gertler refers to this problem as the 'problem of conceptualization'. The challenge is to explain how it is (or how it can be) that our introspective judgments do not go beyond our discriminatory powers, and thus to explain how we can, as condition (iii) requires, directly grasp the correspondence between our introspective judgments and the mental sates they are about.

5.5.1 A Sketch of a Response to the Problem of Conceptualization

Gertler's response to the problem of conceptualization has essentially two parts. The first consists of the following claim: "[e]pistemic appearances will not mislead a scrupulously cautious thinker who exercises adequate care in introspectively reflecting upon her current experience" (110). This claim constitutes *part* of an answer to the problem of conceptualization since, if true, it ensures that in certain cases one can trust the epistemic appearance of one's sensation or experience. That is, when adequate caution and care is practiced, the epistemic appearance of a sensation or perceptual experience will reflect its underlying phenomenal reality. The scrupulously cautious introspector need not worry that the epistemic appearance fails to correspond with the phenomenal reality, due to either a failure to respect one's discriminatory limitations, inattention, or some other cognitive deficiency.

This is not, however, a full response to the problem of conceptualization since, as just characterized, it makes no mention of one's conceptualization of, or judgment about, one's sensations or perceptual experiences. The previous claim about epistemic appearances must be combined with a claim connecting one's conceptualization of an experience with that experience's epistemic appearance. The sketch of a response to the problem of conceptualization, then, has the following two parts:

(A) The epistemic appearance of a sensation or perceptual experience cannot mislead a scrupulously cautious introspector who exercises adequate care.

(B) When conceptualizing a sensation or perceptual experience one can take advantage of (A), thus securing a correspondence between the

> introspective conceptualization/judgment and the mental state it is about.[89]

These two components, then, constitute a sketch of a response to the problem of conceptualization. It is just a sketch at this point, given that reasons are needed for accepting either component. According to Gertler, the required supplementation is to be found in an account of account of phenomenal concepts put forth (independently) by herself (2001) and David Chalmers (2003).

5.5.2 Filling in the Sketch - An Account of Phenomenal Concepts

According to Gertler, the account of phenomenal concepts draws upon the following claim about epistemic appearances:

> … [i]n the relevant cases, an experience's epistemic appearance plays a *dual role*. It is simultaneously an aspect of the experience's phenomenal reality and a component of the introspective judgment. The former role, which is metaphysical, ensures that the epistemic appearance fits the phenomenal reality. The latter role, which is epistemic, explains how the phenomenal property is conceptualized in introspective judgments (112, emphasis added).

As I see it, the alleged metaphysical role played by epistemic appearances is intended to explain the first component of the response to the problem of conceptualization. That is, it is meant to explain why it is that when scrupulous caution and adequate care are exercised, the way an experience epistemically appears reflects the sensation or experience's underlying phenomenal reality. The alleged epistemic role played by epistemic appearances is intended to explain the second component of this

[89] Note that my presentation of Gertler's solution to the problem of conceptualization does not nicely map onto her own presentation of the discussion. For one, she does not make explicit the second component of the solution that I have just identified. Nevertheless, I think my presentation captures the underlying logic of her response and perhaps improves upon her own presentation of it

response. That is, it is meant to explain how one's conceptualization of an experience connects up in the right sort of way with the epistemic appearance of the experience.

5.5.2.1 The Metaphysical Role of Epistemic Appearances

The first of the two aforementioned roles is metaphysical: in the relevant cases, a sensation or experience's epistemic appearance is an aspect of the mental state's phenomenal reality. This is supposed to explain how it is that epistemic appearances will not mislead the scrupulously cautious introspector. Gertler appeals here to our ability to "prune" epistemic appearances. Recall that x's epistemic appearance is a matter of the judgments that the subject (the one being appeared to) is inclined to make about x. But this, of course, will depend upon various factors concerning the subject. Gertler mentions two such factors: the subject's perspective and the subject's powers of discrimination.

As an example of the first, Gertler describes a subject who looks at a white wall with a blue light shown on it. Importantly, the subject knows about the blue light. The wall phenomenally appears to the subject to be blue; i.e., the subject's experience of the wall involves the phenomenal property *blueness*. Nevertheless, due to the subject's belief about the presence and orientation of the blue light, the wall does *not* epistemically appear to the subject to be blue. If the subject knows that the wall is white, despite its phenomenally appearing blue, then the wall will epistemically appear to the subject to be white. In this case, the subject's knowledge of the light gives him a perspective that differs from one who is unaware of the light. This perspectival difference generates a difference in epistemic appearance.

Differences in discriminatory powers can also account for difference in epistemic appearance. As an example, Gertler considers a red wine connoisseur who is able to

distinguish, not only pinot noirs from other kinds of red wines, but also Oregon pinot noirs from other kind of red wines (including non-Oregon pinot noirs). The connoisseur, when presented with a glass of Oregon pinot noir, is inclined to judge that it is an Oregon pinot noir. Of course she is also inclined to judge that it is a pinot noir, that it is a red wine, that it is a wine, that it is an alcoholic beverage, that it is a liquid, etc. The wine thus epistemically appears to the connoisseur to be all of these things.

But consider now a red wine drinker whose palette is not quite so sophisticated. Suppose he can distinguish pinot noirs from other kinds of red wines, but that he cannot distinguish Oregon pinot noirs from non-Oregon pinot noirs. To this subject, an Oregon pinot noir does *not* epistemically appear to be an Oregon pinot noir; he is not be inclined to judge that it is an Oregon pinot noir. But he is inclined to judge that it is a pinot noir, that it is a red wine, that it is a wine, that it is an alcoholic beverage, that it is a liquid, etc. For these two subjects, then, there is much overlap concerning how an Oregon pinot noir epistemically appears. Yet there is a difference, and this difference is due to differences in their discriminatory powers.

Moreover, Gertler claims that there are differences concerning epistemic appearances *within* a subject. These differences have to do with confidence levels. Although the wine connoisseur can reliably distinguish Oregon pinot noirs from non-Oregon pinot noirs, and can also reliably distinguish pinot noirs from other red wines, and reds from non-reds, and wines from non-wines, etc., she will presumably have different confidence levels with respect to these judgments. Assuming she is aware of these differences in confidence levels, she can, if exercising extreme caution, restrict her judgments about the wine to those for which her confidence levels are highest.

Suppose the connoisseur's life depends upon her making just *some* correct judgment about the wine. In this situation she will be inclined to judge that the wine is a red wine, and possibly even a pinot noir, but she will *not* be inclined to judge that it is an Oregon pinot noir. Her wanting to be especially careful in her judgments will disincline her from judging that the wine is an Oregon pinot noir. This cautious attitude thus affects how the wine epistemically appears to her (although it does not affect how the wine phenomenally appears to her). More importantly, though, this cautious attitude makes the epistemic appearance of the wine less likely to be mistaken. In this way, one can *prune* one's epistemic appearances, making them more likely to not mislead.

In these examples, red wine is the object of awareness. Because red wine is an external object, pruning cannot ensure that the epistemic appearance of the red wine does not mislead. Despite the connoisseur's caution, she may mistakenly judge the wine to be, e.g., a pinot noir when, in fact, it is a cabernet sauvignon. Consider, then, a *mental* object of awareness. Specifically, consider a subject who is acquainted with some perceptual experience. Gertler's claim is that the directness of one's relation to the experience, made possible by one's acquaintance with it, ensures that through scrupulous caution and adequate care the experience's phenomenal reality is all that contributes to the experience's epistemic appearance; the epistemic appearance is *exhausted by* the phenomenal reality. This is made possible by both the metaphysical directness of acquaintance and the introspector's exercise of scrupulous caution.

5.5.2.2 The Epistemic Role of Epistemic Appearances

The second of the two aforementioned roles is epistemic. This role is intended to explain how a conceptualization of one's own experience is formed in a way that takes

advantage of the fact that epistemic appearances will not mislead a scrupulously cautious introspector.

To this end, Gertler describes Keith Donnellan's (1966) well-known example concerning reference. Suppose, while at a party, one sees a man holding a martini glass. This man is Mr. Smith. If one does not know the man's identity, one might ask "who is the man drinking a martini?" Interestingly, whether or not the man in question is actually drinking a Martini, one's question is about that man. Reference to Mr. Smith is preserved despite one's erroneous description of Mr. Smith.

Gertler admits that Donnellan is here concerned with *linguistic* reference, but thinks that a similar point can be made about thought. Following Gertler, suppose that instead of asking a question about the identity of the man, one forms the following *thought* about the man: the man drinking the martini is nattily dressed. According to Gertler, one's thought is about Mr. Smith despite one's picking him out via a misleading epistemic appearance. She writes that "[h]e is the person who epistemically appears to me to be drinking a Martini, and while this appearance is misleading I can nonetheless use it to form a thought about Mr. Smith" (114).

Gertler claims that *attention* explains this ability to successfully refer despite one's employment of a misleading epistemic appearance. The epistemic appearance of Mr. Smith is misleading (he is not, in fact, drinking a Martini), but this misleading epistemic appearance is based on one's visual attention being directed at Mr. Smith. This fact about attention enables one to refer to Mr. Smith in thought, despite the fact that one's thought involves a description of Mr. Smith that he fails to satisfy. According to Gertler:

> To make this connection to attention more explicit, we should formulate my thought as *that man is nattily dressed*. Here*, that man* refers to Mr. Smith by virtue of the fact that Mr. Smith is the man who epistemically appears to me to be drinking a martini (114).

Returning to the topic of introspective knowledge by acquaintance, the idea is that one can pick out in thought the phenomenal properties involved in one's own sensations and experiences in a way that is essentially similar to the example involving Mr. Smith. According to Gertler, "[a]ttending to how your experience feels to you – the phenomenal quality it epistemically appears to exhibit – you can think of it as *this.* Whereas *that man* picks out Mr. Smith by visual attention (how he epistemically appears, to perception), *this* picks out the phenomenal property of your sensation by introspective attention – that is, by how it epistemically appears to introspection" (114). One can use the epistemic appearance of the experience, which will be exhausted by phenomenal reality when scrupulous caution is exercised, to secure reference in thought to that experience's phenomenal reality.

Essentially what Gertler wants here is for epistemic appearances to be components of introspective judgments. The way this is supposed to work, I take it, is that phenomenal concepts, which are uncontroversially components of introspective judgments, are supposed to be constituted by the epistemic appearances of the phenomenal properties to which they refer. In this way there is nothing more to these concepts than these epistemic appearances.

If this is correct, then an experience's epistemic appearance will literally be a component of the introspective judgment about that experience. In the above examples, '*that*' and '*this*' are concepts formed solely on the basis of attending to the appearances

of the properties in question; in the former case the attention is visual, while in the latter case the attention is introspective. These concepts are thus constituted by these epistemic appearances. Any subsequent judgment involving either of these concepts will thereby have as a component the relevant epistemic appearance.

5.5.2.3 Putting the Two Roles Together

In summary, through scrupulous caution and adequate care an introspector can make it the case that the epistemic appearance of a sensation or experience is exhausted by the sensation or experience's phenomenal reality; the way an experience epistemically appears to such an introspector is the way that it is in reality.[90] To use Gertler's terminology, an experience's epistemic appearance can be an *aspect* of the underlying phenomenal reality. This is the metaphysical role that epistemic appearances are alleged to play.

In addition, by introspectively attending to an experience's epistemic appearance one can refer in thought to that appearance. Through this sort of attention, one is able to form a phenomenal concept that is wholly constituted by the epistemic appearance. Any subsequent judgment about the phenomenal experience will thus literally have this epistemic appearance as a component. This is the *epistemic* role that epistemic appearances are alleged to play.

Combining these two roles, it follows that an introspective judgment can literally have as a component the underlying phenomenal reality of the experience it is about; this

[90] Interestingly, this does not hold in the reverse direction. It is not the case that a phenomenal experience epistemically appears to be all that it is. Recall that for some phenomenal properties (e.g., *48-speckledness*), we are unable to (justifiably) judge that the property is present. For this reason, a phenomenal experience with such a property cannot epistemically appear to have all the properties that it has. Or, to be more precise, any such epistemic appearance would be unjustified.

underlying phenomenal reality can constitute the way in which the introspector conceptualizes the targeted state. For this reason, such a judgment is guaranteed to be true.

This, then, constitutes a solution to the problem of conceptualization. An introspective judgment will correspond with the relevant phenomenal reality so long as (1) the introspector exercises scrupulous caution and adequate care, and (2) the phenomenal concept involved in the introspective judgment is formed in the manner described above. When this occurs, there is no worry that an introspector will step beyond one's discriminatory powers and form an inappropriate introspective judgment.[91]

5.5.2.4 A More Detailed Look at Phenomenal Concepts

As noted above, the account of phenomenal concepts explained in this section comes from Gertler (2001) and Chalmers (2003). I have tried in this section to keep the discussion of this account fairly simple. However, a more detailed description of their work might be beneficial. For this reason, I include this brief section discussing some of the details of these two works. The reader can skip ahead without loss of continuity, if desired.

Gertler (2001) claims that the referent of a phenomenal concept is fixed via pure demonstrative reference. Both demonstratives and non-demonstrative indexicals have descriptive components. These descriptive components determine the *types* of thing to which these terms can refer. According to Gertler:

---

[91] I earlier claimed that Gertler denies that our introspective judgments are infallible. This is not merely because such judgments need not be arrived at via acquaintance. She denies infallibility even if we restrict our attention to those introspective judgments the exploit one's acquaintance with the target mental state. The reason for this seems to be that the two conditions just listed must be satisfied in order to guarantee that one's introspective judgment corresponds with its target mental state.

> The descriptive element of the indexical 'I' entails that it refers to a person; the descriptive element of the indexical 'today' entails that it refers to a temporal region 24 hours long; the descriptive element of the demonstrative 'that tree' entails that it refers to a tree; etc. While 'here' is usually used indexically, it is sometimes used demonstratively, as when one points to a far-away location on a map and says 'we want to get here'. In both uses, the descriptive element of 'here' entails that it refers to a spatial region" (2001, 313).

In order to refer, a demonstrative requires an act of demonstration (in addition to its descriptive component). The combination of this descriptive component with both an act of demonstration and a context suffices for determining a referent. Non-demonstrative indexicals also have a descriptive component, yet this component (in combination with the context) is sufficient for determining a referent; no act of demonstration is required. Gertler claims that acquaintance with our own mental states allows for a kind of demonstrative reference involving *no* descriptive component. Reference is secured through introspective attention alone.[92] Gertler calls this type of demonstrative reference 'pure' demonstrative reference. Phenomenal concepts, on her account, are pure demonstratives.

According to Chalmers (2003), a *pure* phenomenal concept is one that picks out a phenomenal property directly, in terms of its intrinsic phenomenal nature. This is in contrast to phenomenal concepts that pick out phenomenal properties relationally. To give just one example of the latter, consider the concept that Chalmers refers to as the "community relational concept of phenomenal redness" (or red_C). This concept picks out "the phenomenal quality typically caused in normal subjects within [one's] community by paradigmatic red things" (2003, 224).

[92] As Ernest Sosa puts it, "[s]elective attention is the index finger of the mind" (2003, 279).

Here, the referent of the concept is picked out relationally; there is no mention of the intrinsic qualitative character of the phenomenal property in question. As Chalmers points out, though, we seem capable of picking out the referent of a phenomenal concept directly, in terms of the intrinsic phenomenal nature of the phenomenal property. Again, such a concept is a pure phenomenal concept.

Chalmers distinguishes pure phenomenal concepts that are direct from those that are indirect. The difference here is that a pure direct phenomenal concept is partly constituted by the underlying phenomenal reality, while a pure indirect phenomenal concept is not so constituted. According to Chalmers, the clearest case of a pure direct phenomenal concept is one where "a subject attends to the quality of an experience, and forms a concept wholly based on the attention to the quality, 'taking up' the quality into the concept" (2003, 235). Because a pure direct phenomenal concept is partly constituted by the instantiation of a phenomenal property, such a concept obviously requires the instantiation of a phenomenal property.

In contrast, a pure indirect phenomenal concept does not require the instantiation of a phenomenal property. Rather, all that is required is that the referent be picked out directly, via the referent's intrinsic phenomenal nature. One might, for example, form a concept of phenomenal redness directly, by attending to the intrinsic phenomenal nature of an instantiation of phenomenal redness, and by then using one's memory of this instantiation to secure the referent of the concept. One could then use this concept in the absence of any instantiation of phenomenal redness. Chalmers refers to this kind of pure indirect concept as a 'standing phenomenal concept'.

Chalmers posits the relation of acquaintance in order to explain our ability to form pure direct phenomenal concepts. Moreover, it is by employing direct pure phenomenal concepts in one's introspective judgments that one's introspective judgments are able to be especially epistemically secure.

Although there are some differences between these two accounts of phenomenal concepts, they both have in common a certain kind of *directness*. On both accounts, reference is fixed in the absence of any description of the relevant phenomenal property. According to Chalmers' account, a pure direct phenomenal concept refers to the phenomenal property that is its referent by virtue of the fact that the concept is partly constituted by an instantiation of that phenomenal property. As he puts it, when discussing the constitutive relation that a phenomenal property bears to a pure direct phenomenal concept, "we might picture this schematically by suggesting that the basis for a direct phenomenal concept contains within it a 'slot' for an instantiated quality, such that the quality that fills the slot constitutes the content" (2003, 243). Arguably, only mental properties can "fill" such a slot, or, as it was put earlier, can be "taken up" into a concept, and thus *phenomenal* concepts are the only pure direct concepts. Again, Chalmers' explains this uniqueness by positing the relation of acquaintance.

Similarly, according to Gertler's account, it is through introspective attention alone that the referent of a phenomenal concept is fixed. This makes any description of the phenomenal property unnecessary, and thus makes the referent that much more direct. This directness explains the epistemic security of introspective judgments involving pure demonstrative reference.

Finally, it occurs to me that the description-less nature of pure phenomenal concepts, as explicitly characterized by Gertler (2001) and implicitly by Chalmers (2003), has the appearance of conflicting with an earlier discussion. In Section 5.3.2 I noted that introspective knowledge by acquaintance requires that the introspector is aware of the correspondence between her introspective judgment and the mental state the judgment is about. There, it was suggested, following BonJour (2003), that the judgment offers a description of the relevant mental state and that the judgment corresponds with that state if and only if that state satisfies the description. The concern is that the description-less nature of pure phenomenal concepts is in tension with the claim that introspective judgments offer descriptions of the mental states they are about.

I do not know how serious this concern is. Perhaps these points are consistent after all, given that what is description-less is the way in which the referent of the phenomenal concept is secured. One refers to and conceptualizes a phenomenal property directly, without a description. Perhaps this is consistent with the subsequent introspective judgment not being description-less. The judgment describes a particular state of affairs, namely, one involving the particular phenomenal property referred to in a description-less way.

5.6 Attaining Introspective Knowledge by Acquaintance

Gertler's response to the problem of conceptualization, if it succeeds, guarantees that under certain conditions one's introspective judgment concerning a mental state one is acquainted with corresponds to that state's underlying phenomenal reality. But this is not enough to secure introspective knowledge. Recall that there are three conditions for introspective knowledge by acquaintance. The first two conditions require acquaintance

with both the targeted mental state and the introspective judgment concerning that state. The third condition requires acquaintance with their correspondence. While the solution to the problem of conceptualization secures this correspondence (under certain conditions), it does not secure one's awareness (by acquaintance) of this correspondence. How, then, according to Gertler, is one supposed to satisfy condition (iii)?

According to Gertler, one can satisfy this condition by recognizing that the epistemic appearance of the target experience plays the dual metaphysical and epistemic role explained in Section 5.5. One must recognize that, due to scrupulous caution and adequate care, the way that the experience appears is the way that it actually is. One must also recognize that this appearance is a component of one's introspective judgment (i.e., that one conceptualizes the target mental state via this appearance). As I noted above, if the epistemic appearance of one's experience does, in fact, play this dual role, then one's introspective judgment must correspond with the state's underlying phenomenal reality. The suggestion, then, is that by recognizing that the epistemic appearance plays this dual role, one can also recognize that the two must correspond, and thereby recognize that they *do* correspond.

I conclude this section by raising a concern with this suggestion. Specifically, how plausible is it that people are equipped to recognize the alleged dual role played by the epistemic appearances of their own experiences and sensations? Without philosophical training on this matter, such recognition would seem to be quite demanding. This is a problem for the acquaintance theory, if it maintains that introspective knowledge by acquaintance is something that is regularly achieved by non-

philosophers. If so, the acquaintance theorist must deny that the requisite recognition is too cognitively demanding. Whether she can plausibly maintain this is not clear.[93]

5.7 The Acquaintance Theory and the Propositional Attitudes

The above discussion has focused solely on knowledge of one's sensations and perceptual experiences. The reason for this is that such states are those with which we are most likely capable of being acquainted. And this, in turn, is because such states are those most likely to lack an appearance/reality gap.

Notice, however, that only conscious mental states can appear to a subject in any way. Consider my dispositional, non-conscious belief that Obama is president. This belief does not appear to me in any way, at least not until it is "activated" and becomes occurrent; five minutes ago there was simply nothing it was like for me to have this belief. I submit that most of the propositional attitudes that one has at given time are non-occurrent in this way. But then how can the acquaintance theory explain our alleged privileged and peculiar knowledge of these states? On the face of it, it cannot. One cannot be acquainted with a non-occurrent belief, desire, intention, etc.

Indeed, whether there even are such states as occurrent propositional attitudes is controversial. Peter Carruthers (2010), for example, argues quite convincingly, I think, that propositional attitudes have causal profiles that occurrent episodes of inner speech and mental imagery lack; consequently, those occurrent states, although perhaps related to the propositional attitudes, are not themselves propositional attitudes. If this is correct,

---

[93] I should note that, as Gertler points out, not all acquaintance theorists require that one be aware of the correspondence between one's introspective judgment and the mental state that the judgment is about. That is, some such theorists reject condition (iii), requiring only that there *exist* the relevant correspondence. She cites as an example Feldman (2006).

then the acquaintance theory seems unable to explain our privileged and peculiar access to *any* propositional attitudes.

Having noted Carruthers' concern, I wish now to put aside the issue of whether there are occurrent propositional attitudes. After all, even supposing that such attitudes exist, I assume that those who believe in privileged and peculiar access do not wish to restrict that access to occurrent, conscious states; I assume, that is, that such individuals regard our privileged and peculiar access as extending to at least some of our non-occurrent propositional attitudes. If this is correct, then the acquaintance theory will *at best* be only part of the story.

I must admit, however, that to conclude that the acquaintance theory cannot account for our privileged and peculiar access to our non-occurrent propositional attitudes on the grounds just described would be too quick. In a different context, Gertler (2011b) describes a method for knowing about one's non-occurrent dispositional beliefs that seems amenable to acquaintance. She describes Nick, an individual who has the superstitious belief that spilling salt brings bad luck, but who consistently judges that spilling salt does not bring bad luck. Gertler writes that:

> Suppose that … Nick investigates his beliefs about spilling salt by going through an imaginative exercise: he pictures some salt falling from a shaker in his hand. As he visualizes the grains dropping to the floor, he is full of foreboding, and feels a strong urge to pour salt over his shoulder. He concludes 'I guess I still believe that spilling salt brings bad luck' (2011b, 135-36).

She notes that this imaginative exercise amounts to a first-person simulation of the sort advocated by Robert Gordon (1986) and Alvin Goldman (1989).[94]

This example is relevant because Nick, as the example goes, comes to know about his non-occurrent belief that spilling salt brings bad luck via his awareness of two conscious and occurrent states, namely, the foreboding and the strong urge to throw salt over his shoulder. In this case, we might suppose, these states are indicative of the presence of the relevant superstitious belief. Awareness (perhaps by acquaintance) of these states can thus allow Nick to correctly infer that he has the superstitious belief in question.

Perhaps an acquaintance theorist should claim that our privileged and peculiar access to our non-occurrent propositional attitudes is like Nick's access to his superstitious belief: we know about their presence or absence in a privileged and peculiar way by being acquainted with conscious states indicating their presence or absence. So, contra what I wrote above, perhaps the acquaintance theory *can* be the whole story. I conclude this section by addressing this proposal.

First, is Nick's access to his superstitious belief really privileged and peculiar? Surely, I could not come to know that Nick has the superstitious belief in question by imagining myself spilling salt; indeed, my imagining *Nick* spilling salt would not even help, at least not unless I already know that he believes that spilling salt brings bad luck. Nick's knowledge of his belief thus appears to be peculiar.

---

[94] Gertler also quotes Wilson and Dunn (2004, 507) who claim that a study by Schultheiss and Brunstein (1999) provides evidence suggesting that such simulation can be effective in revealing one's propositional attitudes.

Is Nick's imaginative exercise epistemically superior to the methods that I have available for determining whether he has the superstition belief? Is the imaginative exercise more likely to lead to knowledge than the method available to me? This is not clear, but I can certainly appreciate that this might be the case. So I grant it for the sake of argument. There is, however, an additional component of privileged access (as defined in Section1.1). Whether Nick's method satisfies this component depends on whether his knowledge has a different kind of grounding than the kind of knowledge I can have of his superstitious belief. On the one hand, it seems that it does not, given that his knowledge is inferential; his seems to infer his belief from certain (mental) evidence. On the other hand, this evidence is only accessible to Nick. Does this fact make a difference here?

I am tempted to say no, for one can seemingly use mental evidence accessible only to oneself to come to know about the mental states of others; the simulation theory of mindreading (discussed in Section 4.5.4) explains how this could work. For these reasons, I am hesitant to grant Nick privileged knowledge of his superstitious belief. If this is correct, then Gertler's proposal can be set aside, for I am concerned only with privileged and peculiar knowledge of one's attitudes in this section.

Suppose, however, that I am wrong. That is, suppose that Nick's knowledge is both privileged and peculiar. At most, this shows only that Nick can attain privileged and peculiar knowledge of some of his beliefs via a method that potentially involves introspective knowledge by acquaintance. My concern is that for many propositional attitudes that are true of an individual at a particular time, and for which one has privileged and peculiar access to, there will be no exercise (imaginative or otherwise) that

will generate occurrent, conscious states indicating the presence or absence of those attitudes.

For example, I seem to currently know in a privileged and peculiar way that I intend to book a flight for an upcoming trip by the end of the week. My knowledge of this intention does not seem to have been arrived at via an inference from some occurrent, conscious mental state. Moreover, and more importantly, putting aside how I actually arrived at this knowledge, I have trouble seeing how I could have arrived at it in this way. What occurrent, conscious mental state or states would indicate to me the presence of this intention? And what would I need to do to elicit that state or states?

Perhaps I could come to know about this intention via an awareness of a visual image of me booking the flight. But unless a calendar or some other sign indicating the day of the imagined booking is present, how would I know that my intention is to book a flight soon, as opposed to simply booking a flight at some time or another? Also, how would I know that I am *intending* the imagined state, as opposed to entertaining, wishing, or dreading it?

Or perhaps I could come to know about this intention via an awareness of me reciting in inner speech the sentence 'I intend to book that flight by the end of the week'. How, though, am I to elicit this episode of inner speech?[95] If it just happens to occur, then any knowledge of my intention that I might gain as a result of being aware of this episode of inner speech will be somewhat accidental. It will have depended on my happening to utter the relevant sentence in inner speech. If we suppose, as seems reasonable, that I could have known about this intention whenever I happened to consider the matter, this

[95] The same question can, and should, be asked of the previous example concerning visual imagery.

cannot be the correct story. Finally, it would be *ad hoc* to claim that whenever one happens to wonder whether one has a given propositional attitude, the inquiry generates an utterance in inner speech "saying" that one does have that attitude (if, in fact, one does have it).[96]

Thus, while I grant that occurrent, conscious mental states can indicate the presence of non-conscious, non-occurrent propositional attitudes, I deny that this can be the whole story. This proposal seems unable to account for some of our privileged and peculiar knowledge of our attitudes. Moreover, I have the sense that the majority of my beliefs about my propositional attitudes are arrived at *without* inference from conscious, occurrent states that I take to be indicative of the self-ascribed attitudes. Although this judgment is simply an expression of how things seem to me, it inclines me to think that the proposal considered in this section is, in addition to not being the whole story, very little of the story.

5.6 Conclusion

In this chapter I have critically assessed the acquaintance theory, as recently present by Brie Gertler. Perhaps most importantly, I have also argued that it cannot provide a full explanation of the privileged and peculiar access that we seem have to our propositional attitudes. If privileged and peculiar introspective knowledge by acquaintance is possible, it is limited either to conscious, occurrent mental states, or to conscious, occurrent mental states and *some* non-conscious, non-occurrent propositional attitudes.

---

[96] Carruthers (2010) points out, correctly, I think, that episodes of inner speech must be interpreted. For example, the imaged utterance 'I intend to book that flight by the end of the week' could be uttered sarcastically, in which case it would be incorrect to infer from it that I intend to book that flight by the end of the week. If Carruthers is correct, then this significantly complicates the proposal under consideration.

II. The Extrospective Approach

Chapter 6: Extrospection and Belief

6.1 Introduction

Recall the distinction drawn in Section 1.1 between accounts that attempt to explain privileged and peculiar self-knowledge in terms of an inwardly directed method and those that attempt to explain privileged and peculiar self-knowledge in terms of an outwardly directed method. Chapters two through five have focused on two inwardly directed accounts, namely, the inner sense and acquaintance theories; these are introspective accounts. In the next three chapters my focus is on extrospective accounts of privileged and peculiar self-knowledge.

The focus of this chapter is belief, perhaps the most central of all the propositional attitudes. I begin by describing the extrospective approach to self-knowledge in general (6.2). I then describe Alex Byrne's (2005) extrospective account of our privileged and peculiar access to our beliefs (6.3 and 6.4).[97] While I am largely sympathetic to Byrne's view, I argue that it rests on a contentious assumption concerning the relationship between judgment and belief (6.5). In response to this discussion, I develop and defend my own extrospective account that does away with this assumption (6.6). In addition, I offer a brief remark concerning the epistemology of the extrospective approach to self-knowledge (6.7).

[97] Byrne's account is, as far as I can tell, the most developed extrospective account of our knowledge of our beliefs; although Jordi Fernandez (2003) offers a similarly well developed view, it is limited to perceptual beliefs.

6.2 The Extrospective Approach to Self-Knowledge

I noted in Section 1.1 that accounts of self-knowledge have traditionally been in the introspective tradition. And this is understandable, for the mind is apparently *in* the head. Why, then, would one not attend inwardly to attain self-knowledge? Put another way, the (non-mental) world is quite independent from one's mind. How, then, could one possibly gain self-knowledge by attending to the (non-mental) world?[98]

Although this line of thinking is quite natural, some philosophers have found the following line of thinking equally, if not more, natural:

> [I]n making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward – upon the world. If someone asks me 'Do you think there is going to be a third world war?', I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' (Evans 1982, 225).

This quote, previously encountered in Section 1.1, is from Gareth Evans. It is routinely trotted out when discussing extrospective accounts of self-knowledge.

Whatever exactly Evans intended by this remark, many take from it the idea that, e.g., when coming to know whether you believe that *p*, you do not somehow peer inward, scanning your various mental states in search of the belief that *p*. Instead, you think about the potential belief's propositional content, *p*. More specifically, because of the nature of belief, you think about whether *p* is *true*. Because, on this approach, self-knowledge is attained by "looking through" one's mental states, attending to their worldly contents, accounts in the tradition are often referred to as 'transparency accounts'; I have called such accounts 'extrospective' to better capture the contrast with introspection. In addition

[98] I put 'non-mental' in parentheses, given that I take the mind to be part of the natural world. For the most part, though, I will omit this parenthetical remark in what follows.

to Byrne and Evans, philosophers who have recently defended extrospective accounts of self-knowledge include Fred Dretske (1994, 1995), Jordi Fernandez (2003, 2007, 2013), Robert Gordon (2007), Richard Moran (2001), and Michael Tye (1995, 2003).

Philosophers sympathetic to the extrospective approach find the idea that one's attention turns outward when attaining self-knowledge to be *descriptively accurate*. Such philosophers think that, in fact, people turn their attention outward when attaining self-knowledge. However, while this descriptive/psychological issue is interesting and evokes strong feelings on both sides of the debate, some find the issue utterly uninteresting, given the conviction that the extrospective approach is *epistemically* hopeless. Indeed, even those attracted to the approach experience such doubts. Such philosophers are faced with what has been called the 'paradox of transparency'.[99] Richard Moran expresses the point well:

> How can a question referring to a matter of empirical psychological fact about a particular person be legitimately answered without appeal to the evidence about that person, but rather by appeal to a quite independent body of evidence? (2003, 413).

Matthew Boyle (2011) also discusses this paradox, distinguishing two reactions to the claim that one can come to know a conclusion about one's mind via a premise about the world:

> A credulous reaction accepts that we do somehow draw a conclusion about our state of mind from a fact about the world ... For the philosopher who reacts credulously, the puzzle of transparency will consist in the fact that our normal knowledge of what we believe rests on a cognitive transition whose reasonableness is hard to understand, and the task will be to explain how the transition can be reasonable (226).

[99] It is sometimes referred to as the 'puzzle of transparency'; Byrne (2005) and Boyle (2011) refer to it in this way.

> An incredulous reaction holds that this cannot be right, and that the appearance that we make such a transition must reflect some misunderstanding … A philosopher who reacts incredulously … will hold that only a madman could draw such an inference, that it won't do to hold that normal self-aware believers are mad, and hence that the inferential approach to transparency cannot be correct (226-227).

Boyle is incredulous on this matter. I, on the other hand, have the credulous reaction. And so too does Byrne. In the next section I will describe his attempt to "explain how the transition [from world to mind] can be reasonable".

Before turning to that task, however, I wish to note something *positive* about extrospective accounts of self-knowledge. Because they claim that privileged and peculiar self-knowledge is attained by attending to, or thinking about, the world, such accounts need not, and do not, posit any special mechanisms (such as inner sense) or relations (such as acquaintance) intended to explain self-knowledge. The cognitive mechanisms and processes that allow us to think about the world suffice. This is an important feature of such accounts, and is one that many find quite attractive. To use Byrne's (2005) terminology, extrospective accounts are 'economical' as opposed to 'extravagant'.

6.3 Byrne's Account

Byrne offers an extrospective explanation of the alleged privileged and peculiar access that we have to our beliefs. The explanation is based on the notion of an 'epistemic rule'. According to Byrne, an epistemic rule, R, is a conditional of the following form (94):

R If conditions C obtain, believe that *p*.

One follows an epistemic rule just in case one comes to believe that *p* because one has recognized that the antecedent conditions C obtain.

Importantly, coming to believe that *p* because one has recognized that C does not require being aware (judging/believing/recognizing) that one has recognized that C. Following an epistemic rule does not require forming thoughts about what one recognizes; beings incapable of forming such thoughts are thus potentially capable of following epistemic rules. All that is required is that one's recognizing that C leads one to believe that *p*.

An epistemic rule is 'neutral' just in case its antecedent conditions, C, make no reference to the rule follower's mental states (ibid). Following a neutral epistemic rule thus does not require recognizing any mental facts. Notice that this point is distinct from the previous point that following an epistemic rule does not require awareness of one's recognition that C. Following an epistemic rule *always* involves recognizing that C; following an epistemic rule *sometimes* involves recognizing a mental fact (i.e., when C = a mental fact); following an epistemic rule *never* requires awareness that one has recognized that C. As we shall see, neutrality is central to Byrne's account.

Some epistemic rules (neutral or otherwise) are better than others. An epistemic rule is 'good' just in case following it tends to produce knowledge; an epistemic rule is 'bad' just in case following it tends to produce false and unjustified beliefs (ibid).

Byrne takes for granted that we are capable of following neutral epistemic rules, and that we often do follow such rules. He offers the following as an example (ibid):

DOORBELL If the doorbell rings, believe there is someone at the door.[100]

Byrne claims that DOORBELL is probably a good neutral epistemic rule. In contrast, he offers the following as an example of a bad neutral epistemic rule (ibid):

NEWS If the *Weekly World News* reports that *p*, believe that *p*.[101]

NEWS is also an example of a 'schematic' epistemic rule, given the presence of the variable *p*. A schematic epistemic rule is good to the extent that its instances are good, and bad to the extent that its instances are bad (ibid).

With this terminology in place, Byrne offers the following neutral and schematic epistemic rule for acquiring privileged and peculiar knowledge of one's own beliefs (95):

BEL If *p*, believe that you believe that *p*.

Given the above definition for following an epistemic rule, one follows BEL on a given occasion if and only if one forms the belief that one believes that *p* because one has recognized that *p*.

As I just noted, Byrne claims that BEL is neutral. Notice, however, that speaking of a schematic rule being neutral is problematic. If an instance of a schematic rule involves a condition, C, that concerns the state of one's mind, then that instance of the rule will not be neutral; one will be able to follow that instance only if one recognizes a fact about one's mental life. For this reason, Byrne must be implicitly restricting BEL so that *p* cannot concern the rule-follower's mind. An apparent instance of BEL that violates

[100] It seems clear that one can follow this rule without having to be aware that one has judged that the doorbell is ringing. One's judgment that the doorbell is ringing must simply cause one to come to believe that there is someone at the door. That this occurs in my own case seems highly plausible. That is, I doubt that my belief that there is someone at the door is caused by my first becoming aware that I have judged that the doorbell is ringing; I suspect the first-order state about the doorbell ringing is often all that is needed.

[101] The *Weekly World News* is a satirical tabloid specializing in outlandish headlines.

this restriction will thus not be an actual instance of BEL. With this stipulation in place, BEL *is* a neutral rule: following it does not require recognizing facts about one's own mind.

Consider, for example, one who comes to believe that he believes that the penny has landed heads because he has recognized that the penny has landed heads. This requires recognizing only the non-mental fact that the penny has landed heads. In this case, one comes to have a belief about one's belief via recognizing a non-mental fact. This is the way in which BEL is a neutral epistemic rule.

The importance of BEL's neutrality cannot be overstated. If it were not neutral, then following it would presuppose the capacity for self-knowledge, and thus it could not explain that capacity.[102] More importantly, though, to give up on neutrality is to give up on that which is most distinctive and interesting about extrospective accounts of self-knowledge, namely, the claim that self-knowledge is attained by turning one's attention outward to the (non-mental) world. I thus regard the account's neutrality as essential to it.

I will now introduce a few more pieces of terminology that are central to the epistemology of Byrne's account. An epistemic rule, R, is 'self-verifying' just in case following R guarantees the truth of the resulting belief. Byrne claims that BEL is self-verifying. This claim is based on his contention that recognizing that *p* suffices for coming to believe that *p*. I will grant this point for now. If S comes to believe that he

[102] Although given that BEL is intended only to account for our privileged and peculiar access to our *belief*s, the account would not be viciously circular so long as the antecedent conditions made no reference to belief.

believes that *p* because he has recognized that *p*, then, because recognizing that *p* leads to believing that *p*, S's second-order belief is guaranteed to be true (96).[103]

In fact, Byrne goes a bit further, claiming that merely trying, but failing, to follow BEL guarantees the truth of the resulting second-order belief. One merely tries, but fails, to follow BEL on a given occasion just in case one comes to believe that one believes that *p* because one has incorrectly judged (and so has not recognized) that *p*. In what follows, I will refer to this particular kind of failure as 'merely trying to follow BEL'.

Incidentally, this shows that, for Byrne, one's judging that *p* (correctly or incorrectly) is that which makes BEL so truth-conducive. Correctly judging that *p* (i.e., recognizing that *p*) involves coming to believe that *p*. But so too does incorrectly judging (and so not recognizing) that *p*. BEL is thus 'strongly self-verifying'.[104] This is so in virtue of the alleged connection between judgment and belief, which I will discuss in detail in Section 6.3.[105]

If Byrne is correct, then BEL is a hyper-reliable rule. Because he assumes a reliabilist account of justification, he concludes that BEL is knowledge-conducive: following (or merely trying to follow) BEL will tend to yield knowledge. Notice, however, that this leaves open whether BEL can accommodate the alleged privileged and peculiar access that we have to our beliefs. Knowledge is one thing, privileged and peculiar knowledge is another. Further argument is thus required to show that BEL can accommodate privileged and peculiar access.

---

[103] Byrne notes that, strictly speaking, this is incorrect. Because there will be a short time lag between one's recognizing that *p* and one's coming to believe that one believes that *p*, it is possible that one will no longer believe that *p* by the time one has formed the higher-order belief. I regard this as a very minor issue and will not address it in what follows.

[104] The term 'strongly self-verifying' comes from Byrne (2011).

[105] Byrne (2005) writes in a footnote that "… judging is the act that results in the state of belief" (102).

To this end, Byrne appeals to the epistemic notion of 'safety'. A true belief is safe just in case it could not easily have been false, and safety is taken by many to be a necessary condition for knowledge.[106] The details of Byrne's argument need not concern us here. His main point is simply that certain kinds of error that are possible with respect to our rules for attributing beliefs to others are not possible with respect to BEL. Consequently, true beliefs produced by following (or merely trying to follow) BEL are epistemically safer than those produced by following such rules. Byrne concludes from this that BEL can accommodate the alleged privileged access that we have to our beliefs.

Notice that BEL cannot be used to gain knowledge of others' beliefs. Concluding that another person believes that *p* on the basis of one's own recognition that *p* is obviously not a truth-conducive method of reasoning.[107] If following (or merely trying to follow) BEL is our first-personal method for attaining knowledge of our own beliefs, then this method is different in kind from our method for coming to know the beliefs of others. Byrne thus concludes that BEL can accommodate the alleged peculiar access that we have to our beliefs.

6.4 Two Projects: Psychological and Epistemological

I should point out that even if following (or merely trying to follow) BEL yields privileged and peculiar knowledge of one's beliefs, this is not to say that Byrne's account *explains* the privileged and peculiar access that we seem to have to our own beliefs. For that, Byrne's account must be descriptively true of us. That is, BEL must be the first-

---

[106] Following Byrne, I have offered Williamson's (2001) formulation of safety. But see also Sosa (1999).
[107] Perhaps a better third-person analogue is: if S says that *p*, believe that S believes that *p*. This, though, seems like a bad rule, given that people often say things they do not take to be true. To avoid this problem would seem to require judging that another's utterance is sincere, in which case the reasoning is no longer neutral.

personal method that normal human beings use to acquire first-personal knowledge of their beliefs. This highlights the fact that there are at least two distinct, but related, projects to keep in mind when considering self-knowledge: an epistemological project and a psychological project.

Assume that we often attain privileged and peculiar knowledge of (some of) our own mental states, including our beliefs. One project is to articulate a first-personal method for forming beliefs about one's mental states such that those beliefs qualify as privileged and peculiar knowledge; this is primarily an epistemological project. A second project is to make a compelling case that some particular first-personal method is, in fact, the first-personal method that we use to attain privileged and peculiar self-knowledge; this is primarily a psychological project.[108]

As I noted above, the projects are not completely independent of one another. Psychological plausibility is a constraint on those engaging in the epistemological project. A first-personal method capable of yielding privileged and peculiar self-knowledge will be of interest only if there is a decent chance that it is our actual first-personal method for attaining such knowledge. And of course one can engage in the psychological project only if one already has some particular first-personal method in mind.[109]

---

[108] This is not to say that philosophers cannot contribute to this project, for I think that they can.

[109] This is correct, given how I have defined the two projects. However, one could attempt to establish that humans use an extrospective method, rather than an introspective method, without thereby arguing for a particular extrospective account. Here, one would be engaged in *a* psychological project without having a particular first-personal method in mind.

While Byrne engages a bit in both projects, most of his attention is directed at the epistemological project. And this is exclusively what I have focused on in this section. I will set aside the psychological project in this chapter, returning to it in Chapter 9.

6.5 Judgment and Belief - A Problem for Byrne's Account

I explained above that Byrne takes BEL to be strongly self-verifying. This is because he thinks that judging that *p* (correctly or incorrectly) involves coming to believe that *p*. But is this claim about the connection between judgment and belief true? This question has recently received quite a bit of attention from philosophers.[110] The kind of case at the center of this debate is one where an individual allegedly judges that *p*, but fails to exhibit many of the dispositions normally associated with believing that *p*. Christopher Peacocke (1999) describes a case of this kind:

> Someone may judge that undergraduate degrees from countries other than her own are of an equal standard to her own, and excellent reasons may be operative in her assertions to that effect. All the same, it may be quite clear, in decisions she makes on hiring, or in making recommendations, that she does not really have this belief at all (242–243).

Suppose this individual comes to believe that she believes that undergraduate degrees from other countries and her own are of an equal standard as a result of her judging that degrees from other countries and her own are of an equal standard. On this assumption, she has followed (or merely tried to follow) BEL. However, if Peacocke's analysis of this case is correct, her second-order belief is false; although she judges that undergraduate degrees from other countries and her own are of an equal standard, she does not believe this. In self-ascribing this belief, the woman would be "relying on the

[110] See, e.g, Peacocke (1999), Zimmerman (2006), Gendler (2008a and b), Schwitzgebel (2010), Gertler (2011b), and Paul (2012).

holding of the normal relations between judgment and belief which are not guaranteed to hold" (243). In short, Peacocke claims that the relevant connection between judgment and belief is contingent.

Eric Schwitzgebel (2010) discusses similar cases. He writes, for example, of the academic, Juliet, who proclaims that all races are of equal intelligence. While Juliet is "prepared to argue coherently, sincerely, and vehemently for equality of intelligence and has argued the point repeatedly in the past", she is nevertheless "systematically racist in most of her spontaneous reactions, her unguarded behavior, and her judgments about particular cases" (532). Juliet, despite her sincere judgments to the contrary, is an implicit racist. Arguably, if Juliet were to come to believe that she believes that all races are of equal intelligence as a result of her judging that this is true, the resulting second-order belief would be false. Arguably, then, following (or merely trying to follow) BEL *can* lead to false second-order beliefs, contra Byrne's claim that BEL is strongly self-verifying.

Schwitzgebel's position, like Peacocke's, denies that judging that *p* suffices for coming to believe that *p*. They differ, however, in that Schwitzgebel takes cases of this kind to support a dispositional account of belief (2002).[111] According to his particular kind of dispositionalism, and over-simplifying a bit, believing that *p* is nothing more than having an appropriate number of the dispositions that are stereotypical of believing that *p*. Importantly, these dispositions can be behavioral, cognitive, or phenomenal, thus distinguishing the view from a crude form a behaviorism. Crucial to the account is Schwitzgebel's denial that any single disposition, including the disposition to judge that

[111] For discussion and defense of dispositionalism (expanded beyond 'belief') see Schwitzgebel (2001, 2002, 2012), Hunter (2009, 2011), and Steinberg (2010, 2011, unpublished).

p, is either necessary or sufficient for believing that *p*. Believing that *p*, on this view, is a matter of degree, depending on how many of the relevant dispositions one possesses.[112]

Of course there are philosophers who, like Byrne, maintain that judging that *p* suffices for coming to believe that *p*. Aaron Zimmerman (2006), for example, finds "unassailable" the claim that "if a subject judges that *p* at *t* that subject believes that *p* at *t*" (365). Zimmerman addresses Peacocke's example, offering several alternative interpretations that do not violate the alleged "unassailable" fact. On Zimmerman's preferred interpretation, the woman gains the belief that undergraduate degrees from other countries and her own are of an equal standard upon making the relevant (sincere) judgment, but loses this belief in circumstances where she is not attending to her occurrent judgment. Tamar Szabó Gendler (2008a and b) appears also to endorse the claim that judgment suffices for belief. At the very least, she can deal with the kinds of cases considered above without denying the link between judgment and belief. She can claim that the individuals from the two cases described above have the relevant beliefs, but have 'aliefs' that drive their behavior in the opposite direction.

I tend to side with Peacocke and Schwitzgebel in this debate. But justifying their position is not a goal of this chapter. Rather, my more modest aim is merely to make clear that there is significant disagreement over whether judging that *p* suffices for believing that *p*. That Byrne's account rests on such a contentious claim is certainly less

[112] Notice that one can hold that dispositions are important for belief while denying the dispositionalist's claim that having such-and-such dispositions *constitutes* one's believing that *p*. A functionalist or identity theorist, for example can agree that dispositions are important to (but not constitutive of) belief. Moreover, they can claim on that basis that one who lacks most of the dispositions *associated* with believing that *p* is unlikely to genuinely believe that *p*. This claim, I think, is something of a truism about belief, and is one that any philosopher should accept, regardless of his or her view about what belief *is*.

than ideal. This raises the question of whether an account based on BEL can succeed without assuming that judgment suffices for belief. I think that it can and I will develop just such an account in the next section.

6.6 A New BEL-Based Account of Knowledge of Belief

I noted in Section 6.4 that I am putting aside the psychological project in this chapter. My aim in what remains is to show how a BEL-based account can accommodate privileged and peculiar knowledge of our beliefs. That is, my aim is to provide an account on which following (or merely trying to follow) BEL yields privileged and peculiar knowledge of one's beliefs. Unlike Byrne, however, I will not assume that judging that *p* suffices for coming to believe that *p*. On my account, BEL is not self-verifying (and so is not strongly self-verifying). My account is consistent with the connection between judgment and belief being much *looser* than is supposed by Byrne. This flexibility is a virtue of the account.

By dropping the assumption that BEL is strongly self-verifying, I am forced to supplement BEL with a different claim (or set of claims) in order to show that it can yield privileged and peculiar self-knowledge. Doing so is the task of this section. My account supplements BEL with three claims. The first concerns a tendency of individuals who believe that *p*. The second concerns a tendency of individuals who do not believe that *p*. The third concerns how BEL is used.

6.6.1 [Believe, Judge] and [~Believe, ~Judge]

The first two claims of my BEL-based account are as follows:

> [Believe, Judge] Those who believe that *p* are such that they will tend to judge that *p*, upon considering whether *p*, if they neither (i) make use of

new (external) information, nor (ii) make new use of old (internal) information, when considering whether *p*.

[~Believe, ~Judge] Those who do not believe that *p* are such that they will tend not to judge that *p*, upon considering whether *p*, if they neither (i) make use of new (external) information, nor (ii) make new use of old (internal) information, when considering whether *p*.

These claims assert that when conditions (i) and (ii) hold, an individual's judgment concerning a proposition, *p*, will tend to align with his or her belief attitude towards *p* held just prior to considering whether *p*.

Two important points about these claims are worth noting. First, the direction of [Believe, Judge] is from belief to judgment, not judgment to belief. Similarly, the direction of [~Believe, ~Judge] is from lack of belief to lack of judgment, not lack of judgment to lack of belief. For these reasons, the issues raised in the previous section do not straightforwardly apply to the plausibility of these claims.[113] Second, the claim discussed in the previous section is a sufficiency claim, namely, the claim that judging that *p* suffices for coming to believe that *p*. In contrast, [Believe, Judge] and [~Believe, ~Judge] assert tendencies.

6.6.2 The Need for Conditions (i) and (ii)

Why are conditions (i) and (ii) included in these two claims? To answer this question, it will be helpful to have at my disposal a simple case. Gareth is a normal adult who happens to believe that Hillary Clinton will be the next president. Moreover, Gareth does not believe that Joe Biden will be the next president.

[113] I will have more to say about this in Section 6.6.4.

Imagine that Gareth considers whether the proposition ‘Hillary Clinton will be the next president’ is true. Further, imagine that just as Gareth begins his consideration, he sees on the news that Hillary Clinton has been arrested for armed robbery. Surely, this would cause Gareth to either withhold judgment about whether Hillary Clinton will be the next president, or to judge that this proposition is false; he almost certainly would not judge that it is true. Here, *new (external) information* informs Gareth’s consideration of the proposition, causing him to make a judgment that does not align with his previous belief attitude towards that proposition. And the same kind of case could be easily constructed for Gareth’s lack of belief that Joe Biden will be the next president. Imagine, for example, that just as he begins to consider that proposition he learns that Hillary Clinton has retired from politics. Thus are the reasons for including condition (i) in both [Believe, Judge] and [~Believe, ~Judge].

New (external) information is not the only reason that one’s judgment concerning whether *p* might not align with one’s previous belief attitude towards *p*. Making *new* use of old (internal) information (i.e., information had prior to considering whether *p*) is another reason. Imagine, for example, that Gareth believes that (i) President Obama will endorse Joe Biden if the latter decides to run, (ii) Joe Biden will decide to run, and (iii) whomever President Obama endorses will be the next president. The contents of these three beliefs entail the proposition that ‘Joe Biden will be the next president’. Given that Gareth believes that Hillary Clinton will be the next president, these other beliefs have (for whatever reason) failed to extinguish that belief. Perhaps he just recently formed the first of the three beliefs described, and has not yet had a chance to reflect on its relevance to his belief that Hillary Clinton will be the next president.

The important point here is that if Gareth were to consider whether Hillary Clinton will be the next president, the relevance of these other beliefs might become apparent to him. If so, then they might prevent him from judging that Hillary Clinton will be the next president. As before, a similar kind of case could easily be constructed for Gareth's lack of belief that Joe Biden will be the next president. For these reasons, condition (ii) is included in both [Believe, Judge] and [~Believe, ~Judge].

6.6.3 The Sufficiency of Conditions (i) and (ii)

I will now argue that when conditions (i) and (ii) hold, one's judgment about whether a proposition, *p*, is true will tend to align with one's belief attitude towards *p* held just prior to considering whether *p*; the holding of these conditions *suffices* for this tendency. I will proceed by considering two kinds of belief: explicit belief and implicit belief.

Suppose, as seems plausible, that Gareth's belief that Hillary Clinton will be the next president is an explicit standing belief in something like the following sense: (a) Gareth has in the past endorsed the relevant content (i.e., 'Hillary Clinton will be the next president'), (b) Gareth has stored the endorsed content in memory, and (c) Gareth can readily recall the endorsed content.[114] Next, suppose that Gareth (for whatever reason) begins to consider whether Hillary Clinton will be the next president. On the face of it, Gareth's consideration of this proposition should "trigger" his stored endorsement of it, leading him to judge that Hillary Clinton will be the next president. Of course acquiring a new piece of (external) information, or making new use of old (internal) information

[114] Gertler (2011b) discusses this kind of belief; she calls it '(ordinary) dispositional belief'. The above three conditions are from her discussion. I, along with Gertler, take these conditions to be jointly sufficient for having an explicit standing belief. The ensuing discussion of implicit belief is also influenced by Gertler (2011b).

could interfere with this. But in the absence of these potential interferences, Gareth's explicit standing belief should lead him to judge that Hillary Clinton will be the next president. This suggests, then, that [Believe, Judge] is true of explicit standing beliefs of the sort just described.

But we also seem to have non-explicit beliefs. Is [Believe, Judge] plausible for them as well? To see that it is, consider Gareth's implicit belief that Hillary Clinton has ten toes. Suppose he has this belief by virtue of satisfying the following conditions: (a') he has not previously endorsed the relevant content (i.e., 'Hillary Clinton has ten toes'), and so (trivially) has not stored the endorsed content in memory, but (b') if he were to consider the content, he would assent to it without acquiring new evidence concerning the matter.[115] Similarly, and using these same criteria, most adults implicitly believe that there are no bicycles on the moon, that the number 307 is less than the number 313, etc.[116]

On this sense of belief, the case for [Believe, Judge] is quite strong. If we assume (as seems plausible) that assenting to a proposition, *p*, is the same as (or involves) judging that *p*, then to implicitly believe that *p* is, in part, to be such that, if one were to consider whether *p*, one would judge that *p* without acquiring new evidence concerning

---

[115] As before, I understand (a') and (b') to be jointly sufficient for having an implicit belief. The clause in (b') concerning new evidence is meant to distinguish one's having an implicit belief that *p* from one's being merely disposed to believe that *p*. One who does not believe that there are exactly five people in the room (explicitly or implicitly) might nevertheless be disposed to believe that there are exactly five people in the room by virtue of being able to see and count all five people in the room.

[116] Some philosophers believe that implicitly believing that *p* requires the additional condition that one's assent be *quick*. If S is such that she would assent to *p* (without acquiring new evidence) only after a lengthy period of deliberation (say, two weeks time), then arguably S does not implicitly believe that *p*. I have dropped this condition in order to simplify the discussion that follows. I believe this simplification is harmless, given that in this paper I am concerned with cases where one's judgment (or lack thereof) concerning *p* will be issued within an amount of time that would satisfy the requirement for quickness (whatever that requirement happens to be).

whether *p*. But if this is true, then [Believe, Judge] holds with respect to implicit beliefs of the kind just described. If Gareth implicitly believes that Hillary Clinton has ten toes, then he should judge that Hillary Clinton has ten toes, upon considering whether this is the case, without acquiring new evidence. So if conditions (i) and (ii) hold, he should judge that Hillary Clinton has ten toes.

Next, consider [~Believe, ~Judge]. Recall that Gareth does not believe that Joe Biden will be the next president.[117] Suppose that Gareth, at time *t*, begins to consider the proposition 'Joe Biden will be the next president'. Either Gareth has previously endorsed this proposition or he has not. If he has not, then (trivially) he has also not stored the endorsed content in memory. Gareth thus satisfies condition (a') at *t*. If Gareth were to judge that this proposition is true without acquiring new evidence, then he would also satisfy condition (b') at *t*. But then, contrary to the initial hypothesis, Gareth did, at *t*, (implicitly) believe that Joe Biden will be the next president. So he can judge that Joe Biden will be president only if this judgment is based on new evidence. If it is not, i.e., if condition (i) holds, then Gareth cannot judge that Joe Biden will be the next president. So if conditions (i) and (ii) hold, he should be expected to not judge that Joe Biden will be the next president, which is what [~Believe, ~Judge] claims.

Now consider the possibility that Gareth *has* previously endorsed the proposition that 'Joe Biden will be the next president'. Given that he does not explicitly believe this proposition, this endorsement is either not stored in memory, or is not easily accessible. Consequently, Gareth's previous endorsement would seemingly have no influence on his consideration of the matter. But then because Gareth does not believe that Joe Biden will

[117] Note that this is different from disbelieving the proposition, i.e., believing that Joe Biden will not be the next president.

be the next president, it is difficult to see why he would judge that this is so, *if* conditions (i) and (ii) hold. I take these considerations to make a strong case for [~Believe, ~Judge].

This concludes my arguments for [Believe, Judge] and [~Believe, ~Judge]. Before turning to other matters, however, I wish to consider an objection to these claims that is suggested by the discussion from Section 6.5.

6.6.4 Peacocke and Schwitzgebel's Cases Revisited

The cases discussed in Section 6.5 suggest that one who sincerely judges that *p* might not believe that *p*. Although [Believe, Judge] and [~Believe, ~Judge] move in the other direction, they may nevertheless appear to be threatened by such cases. I suggested above that Juliet, from Schwitzgebel's case, does not believe that all races are of equal intelligence. While not believing that *p* does not entail believing that not-*p*, it is of course possible that Juliet *dis*believes that all races are of equal intelligence.

Suppose, then, that this possibility is actualized: Juliet believes that *not* all races are of equal intelligence. If she were to consider whether not all races are of equal intelligence, and if conditions (i) and (ii) were to hold, would she judge that this is the case? The answer seems to be no, for recall that she routinely and sincerely asserts, argues for, and defends the claim that all races are of equal intelligence. [Believe, Judge] thus appears to be threatened. Moreover, supposing that she does not believe that all races are of equal intelligence, [~Believe, ~Judge] also appears to be threatened. Despite not believing that all races are of equal intelligence, she would likely judge that this proposition is true, even when conditions (i) and (ii) hold.

I believe that this threat is only apparent. First, and as already noted, not believing that *p* is different than believing that not-*p*. Peacocke and Schwitzgebel argue that the individuals in their examples lack belief in the relevant propositions. Arguing that these individuals disbelieve the relevant propositions requires further argument and is, in my opinion, not obvious.

Second, the discussion is Section 6.5 concerned a sufficiency claim, namely, the claim that judging that *p* suffices for believing that *p*. I have already noted that [Believe, Judge] and [~Believe, ~Judge] merely assert tendencies. Casting doubt upon the sufficiency claim requires only a single case, namely, a case where one judges that *p* without coming to believe that *p*. To cast doubt upon [Believe, Judge] and [~Believe, ~Judge] requires much more. With respect to [Believe, Judge], for example, it is not enough to present a case where conditions (i) and (ii) hold, but where one believes that *p* without judging that *p* (upon considering whether *p*). One must also argue that such cases occur frequently enough so as to undermine the asserted tendency.

For these reasons, I do not believe that the cases discussed in Section 6.5 threaten to undermine my account. At the very least, showing that they do threaten my account is no small task.

6.6.5 The Significance of [Believe, Judge] and [~Believe, ~Judge]

[Believe, Judge] and [~Believe, ~Judge] are significant because, if true, following (or merely trying to follow) BEL is a highly reliable method for forming beliefs about one's beliefs, *if* conditions (i) and (ii) hold. When these conditions hold, [Believe, Judge] and [~Believe, ~Judge] tell us that one's judgment concerning a proposition will tend to align with one's belief attitude towards that proposition held just prior to considering the

proposition. If S believes that *p* just prior to considering whether *p*, S will tend to judge that *p*; if S does not believe that *p* just prior to considering whether *p*, S will tend not to judge that *p*.

This shows that if S believes that *p*, and uses BEL to determine whether she believes that *p*, S will tend to come to believe that she believes that *p*, if the two conditions hold. Because there is no reason to suppose that judging that *p* will cause one to lose one's belief that *p*, the resulting second-order belief should be true. Similarly, if S does not believe that *p*, and uses BEL to determine whether she believes that *p*, S will tend to not come to believe that she believes that *p*, if the two conditions hold. Because there is no reason to suppose that not judging that *p* will cause one to gain the belief that *p*, S's not coming to believe that she believes that *p* is appropriate. BEL is thus a reliable belief-forming process when appropriately used.

The third claim of my BEL-based account asserts that BEL is (for the most part) used in the appropriate way.

> [USE_B]Those using BEL make use of new (external) information, or make new use of old (internal) information, when considering whether *p*, relatively infrequently.

[USE_B] is consistent with cases where one using BEL either makes use of new (external) information, or makes new use of old (internal) information, when considering whether *p*. It only claims that such cases are infrequent relative to cases where one using BEL does not make use of new (external) information, nor makes new use of old (internal) information, when considering whether *p*. For reasons I will get to below, I will defend [USE_B] in the next chapter.

If BEL is our first-personal method for forming beliefs about our beliefs, then [Believe, Judge], [~Believe, ~Judge], and [USE_B] show that this method is a reliable one: it most often produces true beliefs. Assuming a reliabilist account of justification, these beliefs are justified and, if true, should qualify as knowledge. My account endorses this reliabilist epistemology.[118]

However, as I have noted before, this is not to say that following (or merely trying to follow) BEL yields privileged and peculiar self-knowledge. To establish this, further argument is needed. Consider, first, peculiar access. Because nothing that I have said in this section undermines Byrne's argument that BEL cannot be used to attain knowledge of others' beliefs, I take it that my account can accommodate peculiar self-knowledge. What about privileged self-knowledge?

First, notice that BEL is not an inferential method, at least not in the sense that the resulting beliefs are based on evidence. The mere proposition, *p*, is certainly not evidence for one's believing that *p*. Second, the judgment that leads to the belief that one believes that *p* need not be true. On my account, believing that one believes that *p* as a result of incorrectly judging that *p* does not make the second-order belief any less likely to be true. For these reasons, the first-personal method that I am proposing seems to be more reliable than the method or methods used to form beliefs about others' beliefs; using BEL is a much less risky endeavor (epistemically speaking).

In addition, I take the two points just articulated to show that BEL yields knowledge that is *different in kind* from the knowledge one has of others' beliefs. Surely, my beliefs about the beliefs of others are highly inferential. Moreover, such beliefs are

[118] I say more about the epistemology of my account (and Byrne's) in Section 6.7.

negatively affected, epistemically speaking, if arrived at via false beliefs. I thus conclude that my account can accommodate privileged self-knowledge.

I have already defended [Believe, Judge] and [~Believe, ~Judge]. While I believe that [USE_B] can also be defended, I will postpone doing so until the next chapter. There, I will discuss an important objection to extrospective accounts. The reason for this is twofold. First, the objection, although not aimed at my account, can be seen as targeting my account's commitment to [USE_B]. Second, the objection places an important constraint on *how* I argue for [USE_B]. My response to this objection will serve as an argument for [USE_B], thereby completing the description of, and argument for, my account.

6.7 A Note on Extrospection and Epistemology

Both Byrne's account and my own adopt a reliabilist epistemology. Moreover, the form of reliabilism is what Dretske has disparagingly referred to as 'austere reliabilism'. The kind of non-austere reliabilism that Dretske prefers can be seen in the following passage:

> I am a reliabilist about knowledge, but the kind of reliability I embrace is that relating to the reasons, grounds, or evidence one has for the proposition that one believes, not simply reliability (with or without supporting reasons) of one's belief in the propositions" (Dretske, 2012, 51, n. 3).

Dretske would object to my account (and Byrne's) on the grounds that one using BEL has no reason to believe that he believes that *p*. A proposition about the (non-mental) world is (most often) not a reason to believe that one believes anything, let alone that proposition. On my account, the belief that one believes that *p* counts as knowledge simply by virtue of being produced by a reliable process; a reason is not required. My

account thus adopts the kind of process reliabilism defended by Goldman (1979, 1986).[119]

I suspect that any extrospective account of self-knowledge is going to fail to live up to Dretske's demand for reasons. This is because, as explained above when discussing the paradox of transparency, such accounts claim that one attains knowledge of one's mind by attending to the quite independent matter of the state of the world. Given this independence, that which leads to the belief about one's mind cannot possibly *on its own* serve as a reason for that belief.

But perhaps there are connecting beliefs that one might justifiably hold which when combined with judgments about the world yield reasons in support of one's beliefs about one's mind. Dretske's early work on self-knowledge (1994, 1995, and 1999) appears to hold that there are such beliefs; his later work (2003a, 2003b, 2012a, 2012b), however, unequivocally denies that there exist the requisite connections.[120] I am inclined to agree with the later Dretske on this matter.

This shift in Dretske's thinking caused him to move from defending an extrospective account of self-knowledge (1994, 1995, 1999) to arguing for a kind of skepticism about self-knowledge (2003a, 2003b, 2012a, 2012b). According to this skepticism, while we are able to know about the contents of our minds in a privileged and peculiar way, we have no first-personal way of knowing that we have minds.[121] He was led to this skepticism because of (i) the role that reasons play in his epistemology, (ii) his

---

[119] Armstrong (1973) and, earlier, Ramsey (1931) can also be seen as defending an austere kind of reliabilism.

[120] See Lycan (1999) and Aydede (2003) for arguments against such connecting beliefs; respectively, they take Dretske (1999) and Dretske (1995) as their targets.

[121] Dretske (2012a and 2012b) calls this skepticism 'conciliatory', given that it grants that we have knowledge of our minds via a non-first-personal method.

ardent opposition to inner sense, and (iii) his rejection of the aforementioned connecting belief.

I share Dretske's opposition to inner sense. I thus avoid his skepticism by rejecting his epistemology. That is, I deny that knowledge requires conclusive reasons.[122] I do not find objectionable my account's commitment to austere reliabilism.

6.8 Conclusion

In this chapter I have introduced both the extrospective approach to self-knowledge and the so-called paradox of transparency. In addition to explaining Byrne's extrospective account of our knowledge of belief, I have argued that it rests on a contentious assumption concerning the relationship between judgment and belief. In response, I have put forth my own extrospective account that avoids the aforementioned assumption. I complete the defense of my account in the following chapter.

[122] See Dretske (1971) for his account of conclusive reasons.

Chapter 7: Gertler's Objection and a Defense of [USE_B]

7.1 Introduction

Brie Gertler (2011b) offers an important objection to extrospective accounts of self-knowledge. While her argument is primarily aimed at Byrne's particular extrospective account (described in Section 6.3), it applies to my own as well. In addition, it places a constraint on how I am able to defend [USE_B]. I begin by laying out Gertler's objection, making clear its connection to my own account (7.2). I then offer a defense of [USE_B] that is consistent with the aforementioned constraint (7.3). Doing so completes my defense of the three main claims of my account. In the bulk of what remains, I argue that [USE_B] thwarts Gertler's objection to my account (7.4).

7.2 Gertler's Objection

Gertler's objection has two components. Central to the first is a concern about the influence of *new* information on one who uses BEL. In the previous chapter (Section 6.6), I described certain ways in which such information can influence the connection between belief (or lack thereof) and judgment (or lack thereof). This component of the objection should thus be quite familiar. More important than this, though, is the second component of her objection. Here, she identifies an important restriction regarding the way in which the influence of new information can be dealt with. I divide the remainder of this section into two parts. First, I describe Gertler's objection as applied to Byrne's account (7.2.1). Second, I describe how the objection applies to my own account (7.2.2).

7.2.1 Applied to Byrne's Account

In order to see exactly what Gertler's objection is, it will be helpful to consider two (familiar) cases. First, consider an individual, S, who does not believe that *p*. Next,

imagine that S turns to BEL in order to determine whether she believes that *p*. If S, upon considering whether *p*, looks outward and takes in new information that she takes to be relevant to whether *p*, then S might judge that *p* is the case. If she does, then, following BEL, she will form the belief that she believes that *p*. This, Gertler thinks, is a significant problem.

If, as Byrne maintains, BEL is self-verifying, then this second-order belief will be true. This shows that Gertler's complaint is not that BEL can lead to false beliefs. Instead, her concern is that using BEL might *change* one's beliefs, generating a belief that is then self-ascribed. She denies that our first-personal method for coming to know what we believe has this property. Apparently, for Gertler, the method, M, (whatever its nature) that accounts for our privileged and peculiar access to our beliefs should have something like the following feature:

> Feature 1: if one does not believe that *p* when one turns to M in order to determine whether one believes that *p*, M should not lead one to believe that one believes that *p*.[123]

The problem, then, is that BEL, does not have this feature. One who turns to BEL without believing that *p* might end up believing that she believes that *p* after following BEL.

A similar problem arises with respect to the case where an individual, S, does believe that *p*. As before, imagine that S turns to BEL in order to determine whether she believes that *p*. If S, upon considering whether *p*, looks outward and takes in new

[123] When offering a different objection to Byrne's account, Gertler claims that an adequate account of our special access to our own beliefs must explain, not only privileged and peculiar access (as defined by Byrne), but also the following (alleged) fact: "[w]here t_1 and t_2 are separated only by a moment in which the subject uses some procedure to determine whether she believes that *p* … [i]f (at t_1) I do not believe that *p*, and I happen to wonder whether I believe that *p*, I will not (at t_2) self-attribute the belief that *p*" (128). This is obviously quite similar to Feature 1.

information that she takes to be relevant to whether *p*, then S might *not* judge that *p*. If she does not, then she will not form the belief that she believes that *p* (at least not in virtue of following (or trying to follow) BEL).

Again, there is a sense in which BEL has not performed as it should in this case. Gertler seems to think that if a method, M, is to explain the special access that we have to our beliefs, it should have something like the following feature:

> Feature 2: if one believes that *p* when one turns to M in order to determine whether one believes that *p*, M should lead one to believe that one believes that *p*.

As before, the problem is that BEL does not have this feature. One who turns to BEL while believing that *p* might end up not believing that she believes that *p* after following BEL.

Of course there is a way to amend Byrne's account so that it avoids this objection. As Gertler notes, the account can simply require the follower of BEL to restrict herself, when considering whether *p*, to information that she had in mind when first turning to BEL; she must not allow new information to inform her consideration of whether *p*. If S does not believe that *p* when turning to BEL, this restriction will prevent S from judging that *p* on the basis of new information. Similarly, if S does believe that *p* when turning to BEL, this restriction will prevent S from not judging that *p* on the basis of new information.

Gertler persuasively argues that to restrict oneself in this way is to violate neutrality. Requiring S to restrict herself, when considering whether *p*, to information had in mind prior to turning to BEL is to require S to look "inward"; S must distinguish

information that she had in mind when turning to BEL from new information acquired after turning to BEL. The proposed amendment can thus be adopted only if the method's neutrality is abandoned. However, as noted in the previous chapter, neutrality is a core commitment of extrospective accounts of self-knowledge; indeed, it is their distinctive feature. For this reason, an extrospective account of self-knowledge cannot deal with Gertler's objection in this way.

I suspect some readers will at this point wonder whether the alleged problems that Gertler has identified are legitimate problems for an extrospective account. After all, in each of the two cases described above, S's state of mind seems to change as a result of the new information taken in. That S's beliefs about her beliefs reflect these changes is a good thing, one might suppose.

While I am somewhat sympathetic to this reaction, I can also see matters from Gertler's perspective. Features 1 and 2, when viewed on their own, seem fairly reasonable. Intuitively, our method for attaining privileged and peculiar self-knowledge should have these features. That an account based on BEL apparently does not have these features should thus be of some concern.

7.2.2 Applied to My Account

The issue of neutrality connects with my own BEL-based account described in the previous chapter (Section 6.6). My account's commitment to [USE_B] might appear to be in tension with neutrality. Recall that, according to [USE_B], those using BEL make use of new (external) information, or make new use of old (internal) information, when considering whether *p*, relatively infrequently. The problem is that [USE_B] might appear plausible only if those using BEL take measures, when considering whether *p*, to restrict

themselves to old (internal) information and to avoid making new use of old (internal) information. Because either kind of maneuver requires thinking about internal matters, my account cannot defend [USE_B] in this way. To rebut this objection, I must show that [USE_B] can be defended in a way that is compatible with neutrality.

Notice, though, that this objection is distinct from the one described in Section 7.2.1. Even if I succeed in defending [USE_B] in a way that is consistent with neutrality, my account, like Byrne's, lacks Features 1 and 2. Fortunately, I think that my defense of [USE_B] shows that my account has two features that, although falling short of Features 1 and 2, approximate them enough so as to thwart Gertler's objection. I make this case in Section 7.4.

7.3 Defending [USE_B]

Given that neutrality is an essential feature of an extrospective account of self-knowledge, I must defend [USE_B] in a way that is consistent with neutrality. In Section 7.3.1 I give such a defense for the first half of [USE_B], namely, the claim that those using BEL make use of new (external) information, when considering whether *p*, relatively infrequently. Then, in Section 7.3.2, I give such a defense for the second half of [USE_B], namely, the claim that those using BEL make new use of old (internal) information, when considering whether *p*, relatively infrequently.

7.3.1 Making Use of New (External) Information

My defense of the first half of [USE_B] consists of three stages. Although the third stage is, I think, the most important to my defense, each of the prior stages paves the way for the next.

Stage 1 - The Absence of Relevant Information in One's Immediate Environment

To begin, recall the passage from Evans quoted in the previous chapter that serves as the inspiration for Byrne's account (and extrospective accounts in general). Evans claims that "in making a self-ascription, one's eyes are, so to speak, *or occasionally literally*, directed outward – upon the world" (225, my emphasis). One must caution against making too much of the emphasized part of this quotation. After all, in many cases where an individual, S, might wonder whether he believes that *p*, there will be no information in S's immediate environment that S takes to be relevant to whether *p*.

Taking S to be myself right now (as I sit typing in my office), there is presently no information in my immediate environment that I take to be relevant to the truth of the propositions 'there will be a third world war', 'water boils at 100 degrees Celsius', 'Obama is a good president', 'Tiger Woods will win another major', etc. When considering these propositions, a literal look outward to the world would thus be of no use; such a look would give me information only about the walls in my office, the objects on my desk, the words on my computer monitor, etc.[124]

Importantly, Evans' insight applies in cases of this kind. On both Byrne's account and my own, the sense in which one looks outward (as opposed to inward) to determine whether one believes that *p* is as follows: rather than turning one's attention to the question of whether one believes that *p* (a question concerning one's mind), one turns one's attention to the question of whether *p* (a question concerning the worldly content of

[124] Of course I could obtain information that I take to relevant to the propositions referenced above by searching the internet, (perhaps) by consulting some of the books in my office, etc. There is a sense, then, in which information relevant to these propositions *is* available in my immediate environment. Importantly, though, this is not the sense that I have in mind. The information just mentioned is available to me only if I undertake certain actions, such as accessing the internet or opening a book. This information is not available in one's 'immediate environment', as I am using this term. In Section 7.3.1.1 I consider the relevance of such information-gathering activities to my case for [USE_B].

a belief that one may or may not have). This latter question can be answered when there is no information in one's immediate environment that one takes to be relevant to whether *p*.

To see the importance of these points, imagine an individual, S, who turns to BEL in order to determine whether she believes that *p*. If there is no information in S's immediate environment that she takes to be relevant to whether *p*, then we should expect that new information will *not* inform her consideration of whether *p*; we should expect her consideration to be informed only by information that she possessed when she turned to BEL. S will thus avoid the risk that new (external) information will inform her judgment of whether *p*. Importantly, this risk will be averted, not because S restricts herself to information had in mind when turning to BEL, but rather because of the absence of information in her immediate environment that she takes to be relevant to whether *p*. Neutrality is thus preserved.

Stage 2 - Non-Psychological Interpretations of 'Do You/I Believe that P?'

The preceding stage shows only that for many cases where one might use BEL, new information will not inform one's consideration of whether *p* and neutrality will be preserved. This of course leaves open the possibility that many cases are *not* like this. Consider the following case. Suppose that Billy asks Suzy, while both look up at a darkening sky, whether she believes (or thinks) that there is going to be a storm. In this case, there is clearly information available in Suzy's immediate environment that she takes to be relevant to the proposition in question; she is already looking up at the sky. For this reason, if Suzy were to use BEL, her consideration of whether it will storm

would almost certainly be informed by new information that she takes to be relevant to whether it will storm.

In thinking about this case (and others like it), we must be careful to keep in mind an important fact about the use of sentences with mentalistic terms. Seemingly, Billy asks Suzy about her mind; he asks her whether she *believes* there is going to be a storm. But this is not the only interpretation of the case, nor is it obviously the most natural. A question of the form 'do you believe that *p*?' (posed to S), can be used in at least two ways. First, it might be used to ask whether S has a particular belief at the time the question is posed; let's call this the 'psychological interpretation'. Second, it might be used to ask about *p* itself; let's call this the 'non-psychological interpretation'. On the first interpretation, Billy is genuinely interested in whether Suzy, at that moment, has the belief that there will be a storm. On the second interpretation, Billy is merely interested in the possibility of there being a storm.[125]

Importantly, on the non-psychological interpretation, Billy is *not* making a request for self-knowledge. Rather, he is merely doing what is quite ordinary, namely, talking with another individual about the (non-mental) world. On this interpretation, Billy's question is equivalent to the question (posed to Suzy) 'will it storm?' But to answer this kind of question does not require self-knowledge; it does not require using one's method (whatever its nature) for attaining privileged and peculiar knowledge of one's beliefs.

---

[125] That people use sentences with mentalistic terms even though they do not intend to speak of the states/event/properties to which these terms refer is, I think, uncontroversial. For example, saying that "the restaurant is open until 10:00, *I think*" need not be interpreted as a report of self-knowledge (specifically, a report of what one is occurrently thinking). Often when one says "*p*, I think" (or something similar), one merely intends to express one's being quite confident, although less than certain, that *p* is true. Here, what is being communicated would seem to be a probability claim, namely, something like $0.9 < \Pr(p) < 1$. So long as self-knowledge is not required in order to make and express this kind of judgment, there is a use of the sentence '*p*, I think' that, despite its surface grammar, neither requires self-knowledge, nor is intended to report on the state of one's mind.

Consider, for example, answering affirmatively to the question of whether Barack Obama is president. While, surely, this answer is the result of mental states and operations, it need not be the result of one's being aware of such states and operations.

I suppose one might object that even on the second, allegedly non-psychological interpretation, Billy's question is, at least in part, about Suzy's mind. Although his primary concern, we might suppose, is the possibility of there being a storm, he is also interested in Suzy's opinion on the matter. He poses his question to *Suzy*, after all. However, as I just pointed out, by asking Suzy about the storm, Billy is doing something quite unremarkable, namely, talking with another individual about the (non-mental) world. This kind of engagement between people is ubiquitous. Surely, though, we do not want to claim that metacognition is this pervasive. So although we can admit that there is a sense in which Billy's question, on the non-psychological interpretation, partly concerns Suzy's mind, we should deny that in answering Billy's question, Suzy ought to exercise her method for attaining privileged and peculiar knowledge of her beliefs. But so long as this is denied, the case (on the non-psychological interpretation) is irrelevant to the issue at hand.

Without additional details, there is no correct interpretation of Billy's question to Suzy. I do, however, believe that the non-psychological interpretation is the more natural of the two. On the psychological interpretation, the case seems a bit odd, and not the sort of request one typically makes of another. The main point, though, is simply that care is required when thinking about the first half of [USE_B]. There are many cases that may superficially appear to be of the kind that this half of [USE_B] alleges is relatively infrequent, but which, upon closer examination, are best interpreted in the non-

psychological way described above. But so interpreted, such cases are irrelevant to my defense of [USE_B].

Stage 3 - The Instability of an Introspective, yet Relevantly Informed, Context

Consistent with the previous two stages of my argument are cases where S is genuinely interested in whether she believes that *p* – due to either a question from herself or another – and where there is information in S's immediate environment that she takes to be relevant to whether *p*. While the first half of [USE_B] is compatible with such cases, it requires that they be relatively infrequent. We are now in a position to see why this is likely the case.

The reason, I think, has to do with the *norms* governing belief. For the most part, we want our beliefs to be true; truth is their aim.[126] When information that one takes to be relevant to whether *p* is available in one's immediate environment, one is seemingly in a strong position to determine whether *p* is true, and thus also to ensure that one has the appropriate belief attitude towards *p*. When such information is immediately available, any interest that one might have as to whether one believes that *p* should thus be expected to *shift* to the question of whether *p* is true. Because this latter interest concerns the (non-mental) world, the case no longer concerns self-knowledge.

Consider, again, the case of Billy and Suzy. Suzy has immediately available to her a wealth of information that she (correctly) takes to be relevant to whether there will be a storm. She is thus in a strong position to determine whether there is going to be a storm. Assuming that Suzy aims to have true beliefs, if and when she ever considers whether she

[126] Exceptions include cases where beliefs are held for pragmatic reasons. One might believe, for example, that one's spouse is faithful, not because one has good reason for thinking that this is so, but rather because one is comforted by the belief.

believes that there is going to be a storm – either due to a question posed by herself or another – she should be expected to take advantage of her strong position: she should evaluate the available information, possibly integrating it with other information that she might also have in mind. Failing to do so would demonstrate an indifference towards her (possible) belief's truth. If Suzy lacks this sort of indifference, then any interest in whether she believes that there is going to be a storm will very quickly give way to an interest in the possibility of the storm itself.

The crucial claim, then, is that the immediate availability of information that one takes to be relevant to whether *p* makes for an *unstable environment* with respect to interest in whether one believes that *p*: if and when S becomes interested in whether she believes that *p* – again, due to either a question from S or another – this psychological interest will almost immediately give way to a non-psychological interest, *if* information that S takes to be relevant to whether *p* is available in S's immediate environment. But when one's interest is non-psychological, i.e., when one is interested in whether *p* (as opposed to whether one believes that *p*), one should not be expected to use BEL. After all, BEL is a method for attaining self-knowledge, and so is not suited for an inquiry concerning the non-mental world.

There is thus reason to think that when one does use BEL to determine whether one believes that *p*, there will most often be no information in one's immediate environment that one takes to be relevant to whether *p*. Because of this, cases where one

using BEL makes use of new (external) information when considering whether *p* will be relatively infrequent. This, though, is just what the first half of [USE_B] alleges.[127]

7.3.1.1 An Objection

Before moving on to my defense of the second half of [USE_B], I would like to consider an objection to my argument for its first half. Implicit in my argument is the claim that one using BEL will make use of new (external) information only if that information is available in her immediate environment. Given that there is nothing to stop a user of BEL from seeking out new information, why think that this does not happen frequently enough so as to undermine the first half of [USE_B]? The objection I have in mind claims that there is no good reason.

In response notice that when information that one takes to be relevant to whether *p* is not available in one's immediate environment, acquiring such information will require something more than simply re-directing one's eyes, turning one's head, adjusting the aim of one's ear, etc. One will need to seek out and open a book, get on-line and perform a search, make a phone call and ask a question, etc.

To engage in such information-gathering activities in response to an inquiry concerning one's mind would be extremely odd. People simply do not do this. This is likely because doing so would be to treat the question non-introspectively; one would be concerned not with whether one believes that *p*, but rather with whether one *should*

[127] While Shoemaker (1996) and Moran (2001), among others, have discussed the connection between whether one regards *p* to be true and whether one believes that *p*, my point in this sub-section is unique. I have argued that the presence in one's immediate environment of information that one takes to be relevant to whether *p* will typically extinguish any interest that one might have had in whether one believes that *p*. In effect, my claim is that the presence of such information, combined with an interest in one's beliefs being true, practically ensures that one will not be concerned with what one believes *as such*. If such a concern were present, it would quickly shift to a purely first-order concern, where the use of one's method for privileged and peculiar self-knowledge is inappropriate.

believe that *p*. Given this (contingent) fact, one who uses BEL will be limited, when considering whether *p*, to new (external) information that is already present in her immediate environment. But as I argued above, such information will most often not be relevant to whether *p*.

7.3.2 Making New Use of Old (Internal) Information

In Section 6.6.2, I described the following possibility. Gareth, who believes that Hillary Clinton will be the next president, also believes that (i) President Obama will endorse Joe Biden if the latter decides to run, (ii) Joe Biden will decide to run, and (iii) whomever President Obama endorses will be the next president. Because Gareth believes that Hillary Clinton will be the next president, these other beliefs have failed to extinguish that belief. If Gareth were to consider whether Hillary Clinton will be the next president, these other beliefs might prevent him from judging that this is the case. That is, he might make new use of old (internal) information. The second half of [USE_B] claims that those using BEL will make new use of old (internal) information relatively infrequently. Why think that this is true?

To begin, notice that the problematic kind of case is one where an individual, S, has old (internal) information that is at odds with his current belief attitude towards *p*. Gareth, despite believing that Hillary Clinton will be the next president, has available to him (internal) information to the contrary. This feature of the old (internal) information is very important. Given this feature, the information has for whatever reason failed to have the appropriate effect on one's belief attitude towards *p*. I can think of three possibilities for why this might occur.

First, the relevant information might not be readily accessible to S. Perhaps S would easily see the information's relevance to *p*, but is simply unable to (easily) access that information. In this kind of case, if the individual were to use BEL, he would almost certainly not make new use of that information when considering whether *p*. This first possibility thus does not pose a problem for the second half of [USE_B].

A second possibility is that the information's relevance to *p* might not be readily apparent to S. Perhaps S could see that the information tells against (or in favor of) *p* only if he were to very carefully consider both *p* and the old (internal) information. The example involving Gareth is not like this, given that the relevance of beliefs (i)-(iii) to whether Hillary Clinton will be the next president is fairly obvious. A better example would be one where the old (internal) information's relevance to the proposition is less obvious.

In this kind of case, how likely is it that the individual would make new use of that information when using BEL? This is difficult to say in the abstract. But I think one can safely say that the less obvious the information's relevance to *p* is to S, the less likely that S will make new use of that information when using BEL. After all, BEL is a method for attaining self-knowledge. Because one using BEL is primarily interested in whether she believes that *p*, as opposed to whether *p*, one using BEL is unlikely to uncover some unobvious and previously unknown connection between some old (internal) information and *p*.

The third possibility I have in mind is one on which S has information that is both easily accessible and obviously relevant to whether *p*, but where S has simply not had a chance to reflect on the information's relevance to *p*. If S were to turn to BEL and

consider whether *p*, he would immediately come to see the relevance of that information and would make new use of it: if he previously believed that *p*, he would not judge that *p* (and would possibly cease to believe that *p*); if he previously lacked the belief that *p*, he would judge that *p* (and would possibly come to believe that *p*).

Cases of this kind will not threaten [USE_B] if they occur relatively infrequently, which I suspect they do. In these cases the information's relevance to *p* is obvious to the individual; upon considering whether *p*, she will immediately see the information's relevance. I admit that one can acquire such information at a time without changing one's belief attitude towards *p* at or (around) that time. However, I do think that more often than not the information will have the relevant impact at (or around) the time that it was acquired, *not* later on when one eventually gets around to considering whether *p*. There is thus reason to suspect that cases of this kind will be rare.

Having considered these three possibilities, I conclude that the second half of [USE_B] is quite plausible. Those who use BEL will make new use of old (internal) information relatively infrequently. The reason for this will vary from case to case.

7.3.3 Summary

This concludes my defense of [USE_B], the third and final claim of my account. Importantly, my defense is compatible with neutrality. In no way does it require that those using BEL restrict themselves to certain kinds of information, or to certain uses of information. The user of BEL need not consider such internal matters.

7.4 Revisiting Gertler's Objection

Recall Gertler's objection to Byrne's account. She correctly points out that on his account our first-personal method for attaining knowledge of our beliefs does not have the following features:

- Feature 1: if one does not believe that *p* when one turns to M in order to determine whether one believes that *p*, M should not lead one to believe that one believes that *p*.
    - A method, M, lacking this feature makes possible the following mistake: one turns to M not believing that *p* and yet ends up believing that she believes that *p* after using M.
- Feature 2: if one believes that *p* when one turns to M in order to determine whether one believes that *p*, M should lead one to believe that one believes that *p*.
    - A method, M, lacking this feature makes possible the following mistake: one turns to M believing that *p* and yet ends up not believing that she believes that *p* after using M.

Because Gertler thinks that our first-personal method for attaining knowledge of our beliefs has these two features, and so does not make possible these two kinds of mistakes, she concludes that Byrne's account is inadequate.

My account, like Byrne's, does not have these features. Notice, however, that my account does have two features that closely approximate them:

- Feature 1': if one does not believe that *p* when one turns to M in order to determine whether one believes that *p*, M should *relatively frequently* not lead one to believe that one believes that *p*.[128]
- Feature 2': if one believes that *p* when one turns to M in order to determine whether one believes that *p*, M should *relatively frequently* lead one to believe that one believes that *p*.[129]

I submit that these features of my account thwart Gertler's objection. But I also recognize that some will disagree with this contention. What reasons, then, might one have for disagreeing? That is, what reasons might one have for thinking that our first-personal method for attaining privileged and peculiar knowledge of our beliefs must have the stronger pair of features? I can see just two.

First, it might be claimed that the mistakes identified by Gertler never actually occur. Because they would occur on my account (even if relatively infrequently) my account is inadequate. The problem with this objection is that there appears to be no way of defending the claim on which this objection is based, namely, that the mistakes identified by Gertler never occur. That these mistakes do not necessarily involve *false* beliefs about one's beliefs would appear to make them all the more difficult to detect. The so-called mistakes are ones where an individual's belief about whether he believes

[128] Given [~Believe, ~Judge], one who does not believe that *p* will tend not to judge that *p*, if conditions (i) and (ii) hold. Given [USE_B], these conditions most often do hold when one uses BEL. Consequently, one who does not believe that *p* will tend not to judge that *p* when turning to BEL, and so will tend not to come to believe that one believes that *p* as a result of using BEL.

[129] Given [Believe, Judge], one who believes that *p* will tend to judge that *p*, if conditions (i) and (ii) hold. Given [USE_B], these conditions most often do hold when one uses BEL. Consequently, one who believes that *p* will tend to judge that *p* when turning to BEL, and so will tend to come to believe that she believes that *p* as a result of using BEL.

that *p* does not line up with his belief attitude towards *p* prior to using BEL; the second-order belief might very well be true. I thus conclude that this first reason is unpersuasive.

Second, it might be claimed that it is implausible that our first-personal method for acquiring privileged and peculiar knowledge of our own beliefs can, *when properly used*, lead to the two kinds of mistakes identified by Gertler. Because, on my account, BEL can, when properly used, lead to these mistakes, my account is inadequate. The problem with this objection, however, is that it is unjustifiably demanding. A reason is needed for why the relevant method, when properly used, should be immune to the kinds of mistakes identified by Gertler. In the absence of such a reason, this suggestion fails.

7.5 Conclusion

In this chapter I have explained Gertler's objection to extrospective accounts of self-knowledge. This objection, in addition to applying to my own account, places a constraint on how I am able to defend [USE_B]. In response, I have offered a defense of [USE_B] that is compatible with neutrality. I have also argued that [USE_B] thwarts Gertler's objection, as applied to my account.

Chapter 8: Extending My Account Beyond Belief

8.1 Introduction

In this chapter I examine whether my account can be extended to propositional attitudes other than belief. Are their analogues to BEL, [Believe, Judge], [~Believe, ~Judge], and [USE_B] that can accommodate our alleged privileged and peculiar access to non-beliefs? In the bulk of this chapter I examine this question with respect to desire (8.2), arguing that the extension is promising. I also remark on the form of this extension (8.3) and then briefly consider the extension of the account to intention (8.4).

8.2 Desire

Can my account be extended to accommodate the privileged and peculiar access that we seem to have to our desires? The first question to consider is what is the desire-analogue to BEL? The following obviously will not work:

DES* If *p*, believe that you desire that *p*.

This does not work, given that there is no reliable connection between one taking some proposition to be true and one desiring that proposition. For example, I take the proposition 'the earth's climate is warming' to be true, but I do not desire that it be true; in fact, I desire that it not be true. While truth and belief go together, truth and desire do not.

The natural response to this problem is to move from truth to desirability, where 'being desirable' is taken to mean something like 'worthy of being desired' (as opposed to 'able to be desired'). One immediate problem with this proposal is that one can seemingly find something desirable (in this sense) without thereby desiring it. Take, for example, eating a healthy diet. On the face of it, most people would judge the proposition

'I eat a healthy diet' to be desirable. However, the statistics on obesity suggest that most people do not desire this proposition, for if they did, then, presumably, a much larger number of people would be eating a healthy diet.[130] For now, though, put aside this objection; I will return to it at a later time.

Despite the problem just noted, desirability vis-à-vis desire seems to be the natural analogue to truth vis-à-vis BEL. This leads to the following epistemic rule:

DES If *p* is desirable, believe that you desire that *p*.

Importantly, to judge that something is worthy of being desired is to make a judgment about the non-mental world. While this judgment contains a mental term, namely, 'desire', it does not require judging that anyone (including oneself) has a particular desire. For this reason, DES is a *neutral* epistemic rule, and so is an acceptable basis for an extrospective account of our knowledge of desire.

As with BEL, one follows DES on a given occasion if and only if one comes to believe that one desires that *p* because one has recognized that *p* is desirable. One merely tries to follow[131] DES on a given occasion if and only if one comes to believe that one desires that *p* because one has incorrectly judged (and so not recognized) that *p* is desirable.

Consider, next, the analogues to [Believe, Judge] and [~Believe, ~Judge] appropriate for DES:

- [Desire, Judge] Those who desire that *p* are such that they will tend to judge that *p* is desirable, upon considering whether *p* is desirable, if they

[130] This assumes that most such people have the knowledge and means to eat a healthy diet.
[131] Recall that 'merely trying to follow' an epistemic rule is a technical notion. I defined this in Section 6.3.

neither (i) make use of new (external) information, nor (ii) make new use of old (internal) information, when considering whether *p* is desirable.

- [~Desire, ~Judge] Those who do not desire that *p* are such that they will tend not to judge that *p* is desirable, upon considering whether *p* is desirable, if they neither (i) make use of new (external) information, nor (ii) make new use of old (internal) information, when considering whether *p* is desirable.

The reasons given in Section 6.6.2 for the inclusion of conditions (i) and (ii) in [Believe, Judge] and [~Believe, ~Judge] apply *mutatis mutandis* to [Desire, Judge] and [~Desire, ~Judge].

Finally, the analogue to [USE_B] appropriate for DES is:

- [USE_D] Those using DES make use of new (external) information, or make new use of old (internal) information, when considering whether *p* is desirable, relatively infrequently.

As before, the claim is that such cases are infrequent relative to uses of DES where neither new (external) information, nor old (internal) information influence one's consideration of whether *p* is desirable.

Fortunately, my case for [USE_B] applies *mutatis mutandis* to [USE_D]. Below are the three stages of my argument for the first half of [USE_B] appropriately adjusted for desire. These adjustments do not seem to affect the plausibility of each stage:

- (*Stage 1*) In many situations where one wonders whether one desires that *p*, no information relevant to whether *p* is desirable will be available in one's immediate environment. Thus, there is no risk that such information

will inform one's consideration of whether *p* is desirable. In such cases, the first half of [USE_D] will not be threatened.

- (*Stage 2*) A question of the form 'Do you desire/want that *p*?' (posed to S), can be used in at least two ways: (i) to ask whether S has a particular desire at the time the question is posed and (ii) to ask S about *p*'s desirability. When a question of that form is asked in the presence of information relevant to whether *p* is desirable, the appropriate interpretation *might be* the non-psychological one. If so, then whether such information informs one's consideration of whether *p* is desirable is irrelevant to the plausibility of the first half of [USE_D].
- (*Stage 3*) For the most part, we aim to desire that which is desirable; we aim to avoid desiring that which is undesirable. For this reason, the immediate availability of information that one takes to be relevant to whether *p* is desirable makes for an *unstable environment* with respect to interest in whether one desires that *p*: when such information is immediately available, if and when S becomes interested in whether she desires that *p,* this psychological interest will almost immediately give way to a non-psychological interest. But when it does, the first half of [USE_D] is not threatened.

My arguments for the second half of [USE_B] also apply *mutatis mutandis* to the second half of [USE_D]. However, because this application is quite straightforward, I have chosen to leave it as an exercise for the reader.

I thus conclude that [USE_D] is justified. This leaves only [Desire, Judge] and [~Desire, ~Judge] for me to defend. This is the task of the following subsection.

8.2.1 Defending [Desire, Judge] and [~Desire, ~Judge]

The problematic case for [Desire, Judge] is one where an individual, S, desires that *p* and yet does not judge that *p* is desirable; Ben might desire to eat a double cheese burger with fries for lunch every day, but we can suppose that he would not judge this to be desirable, i.e., worthy of being desired. [Desire, Judge] will be undermined if cases of this kind are not relatively infrequent. Similarly, the problematic case for [~Desire, ~Judge] is one where S does not desire that *p* and yet judges that *p* is desirable; Ben might not desire to eat a healthy diet despite being such that he would judge that eating a healthy diet is desirable. [~Desire, ~Judge] will be undermined if cases of this kind are not relatively infrequent.

My initial response to these problematic cases is to claim that although they might be frequent, they are likely infrequent relative to cases where one's desire (or lack thereof) that *p* lines up with one's judgment (or lack thereof) that *p* is desirable. It is easy to get distracted by these problematic cases. To counter this, consider the following cases:

- (a) Jim desires to get his college degree and is also such that he will judge that getting his college degree is desirable, upon considering the matter.
- (b) Jane desires to publish in a top journal in her field and is also such that she will judge that publishing in a top journal in her field is desirable, upon considering the matter.

- (c) Luke desires good health for his children and is also such that he will judge that good health for his children is desirable, upon considering the matter.
- (d) Lucy desires finding an honest mechanic for her car and is also such that she will judge that finding an honest mechanic for her car is desirable, upon considering the matter.
- (e) Adam does not desire getting laid off at work and is also such that he will not judge that getting laid of at work is desirable, upon considering the matter.
- (f) Alice does not desire to get caught speeding and is also such that she will not judge that getting caught speeding is desirable, upon considering the matter.[132]
- (g) Bill does not desire mayonnaise on his sandwich and is also such that he will not judge that mayonnaise on his sandwich is desirable, upon considering the matter.
- (h) Betty does not desire purchasing ethically raised meat over factory-farmed meat and is also such that she will not judge that purchasing ethically raised meat over factory-farmed meat is desirable, upon considering the matter; indeed, she will judge that purchasing factory-farmed meat over ethically raised meat is desirable, upon considering the matter.

[132] Lucy does not think that speed limits are crucial to public safety. She thinks such limits are an example of the worst kind of government overreach and takes great joy in speeding without getting caught.

Again, I have listed so many examples to make a point. There seem to be *numerous* desires that a given person *has* that happen to line up with that person's judgments about desirability (a-d). And there appear to be *numerous* desires that a person *lacks* that line up with a person's judgments about desirability (e-h). If such desires (or absences of desires) significantly outnumber desires (or absences of desires) that do not line up with judgments about desirability, then [Desire, Judge] and [~Desire, ~Judge] are true.

While I suspect the antecedent of this conditional is true, I do not know how to prove it. My strategy in what follows is thus to play defense. In this spirit, I will now make two cautionary points. First, consider case (h) from above. I constructed this example to illustrate an important point. Suppose that purchasing ethically raised meat over factory-farmed meat is, in fact, desirable. That is, suppose that Betty is wrong to not judge that this proposition is desirable.

Importantly, this does not matter as far as the truth of [Desire, Judge] and [~Desire, ~Judge] is concerned. Despite Betty's mistaken attitude towards this proposition, this mistaken attitude is *her* (mistaken) attitude. Because, as is obvious, *her* attitudes determine the kinds of judgments she is disposed to make at a particular time, she should be expected not to judge that purchasing ethically raised meat over factory-farmed meat is desirable.

Representing Betty's case schematically will be helpful: (i) Betty lacks the desire that *p*; (ii) Betty does not find *p* desirable; (iii) because of this, she should not be expected to judge that *p* is desirable, upon considering whether *p* is desirable; and (iv) all of this is the case despite the fact that *p is* desirable. This illustrates, what was perhaps already

obvious, that the cases that challenge [Desire, Judge] and [~Desire, ~Judge] are not cases where one has an irrational attitude towards a proposition's desirability. Rather, they are case where one's attitude towards a proposition's desirability (rational or irrational) fails to line up with one's desire attitude towards that proposition.

I suspect some find implausible an extrospective account of knowledge of desire because they overlook this fact. They mistakenly assume that judgments about desirability will typically be true, and they then (correctly) infer that this spells trouble for an extrospective account, given that many people seem not to desire much of what is desirable (and seem to desire much of what is undesirable).

My second cautionary point is that one must be careful to distinguish physical urges from desires. Recall that I am concerned with the propositional attitudes; I am thus conceiving of 'desire' as a propositional attitude. Surely, though, this should be distinguished from a physical urge, such a strong urge for nicotine.[133] The relevance of this distinction is easily demonstrated.

Consider the smoker who desires to smoke, but who nevertheless judges that smoking is undesirable (and so does not judge that smoking is desirable). This is the kind of case that [Desire, Judge] claims is relatively infrequent. While I grant that the kind of individual just described is possible, I suspect that in a substantial number of cases that are seemingly of this kind, the desire that is alleged to be present is, in fact, *not* present. Rather, something like a physical urge is in its place. The individual genuinely lacks the desire to smoke, but smokes due to a physical addiction. At the very least, this is a

[133] I am not saying that a physical urge for *p* is incompatible with a desire concerning *p*. I accept that they can, and perhaps often do, coincide.

possibility. Failure to appreciate this point can cause one to overestimate the threat to [Desire, Judge].

A similar point can be made with respect to [~Desire, ~Judge]. Consider the unhealthy eater who does not desire to eat healthy food, but who nevertheless judges that eating healthy food is desirable. Again, I grant that the kind of individual just described is possible. However, I suspect that in a substantial number of cases that are seemingly of this kind, the desire that is alleged to be absent is, in fact, present. The individual has the desire to eat healthy food, but has an even stronger physical urge to eat unhealthy food.[134] Failure to appreciate this possibility can cause one to overestimate the threat to [~Desire, ~Judge].

Two lessons can be extracted from the last two paragraphs. First, when S's behavior seems to indicate the presence of a desire that *p*, and when S nevertheless refuses to judge that *p* is desirable, S may, in fact, lack the desire that *p*, having in its place a physical urge for *p*. Realizing this possibility promises to neutralize many of the cases that are allegedly at odds with [Desire, Judge]. Second, when S's behavior seems to indicate the absence of a desire that *p*, and when S nevertheless judges that *p* is desirable, S may, in fact, have the desire that *p*, yet have a competing (and stronger) urge for not-*p*. Realizing this promises to neutralize many of the cases that are allegedly at odds with [~Desire, ~Judge].

[134] Suppose that instead of having an "even stronger physical urge to eat unhealthy food", he has an even stronger *desire* to eat unhealthy food. If so, then this case is of the kind that [Desire, Judge] claims is relatively infrequent. This is because the individual desires eating unhealthy food, but yet does not judge that eating unhealthy food is desirable.

8.2.2 The Significance of [Desire, Judge] and [~Desire, ~Judge]

[Desire, Judge] and [~Desire, ~Judge], when combined with [USE_D], show that following (or merely trying to follow) DES is a reliable method for forming beliefs about one's desires. If S desires that *p*, and uses DES to determine whether she desires that *p*, S will tend to come to believe that she desires that *p*. Because there is no reason to suppose that judging that *p* is desirable will cause one to lose one's desire that *p*, the resulting second-order belief should be true. Similarly, if S does not desire that *p*, and uses DES to determine whether she desires that *p*, S will tend to not come to believe that she desires that *p*. Because there is no reason to suppose that not judging that *p* is desirable will cause one to gain the desire that *p*, S's not coming to believe that she desires that *p* is appropriate. DES is thus a reliable belief-forming process.

Given my account's commitment to reliabilism about justification, it counts as justified those beliefs attained by following (or merely trying to follow) DES; when true, such beliefs count as knowledge. I argued in Section 6.6.5 that knowledge attained by using BEL is privileged and peculiar. The reasoning provided there applies *mutatis mutandis* to the knowledge attained by using DES. The epistemic rule DES can thus accommodate privileged and peculiar knowledge of one's desires.

8.2.3 An Objection to the Account Just Sketched

One who accepts my defense of [Desire, Judge] and [~Desire, ~Judge] might nevertheless deny that DES is our first-personal method for attaining privileged and peculiar knowledge of our desires. The objection I have mind claims that DES is *explanatorily inadequate*: DES cannot account for some instances of privileged and peculiar knowledge of desires. Assuming that humans have at most one first-personal

method for attaining privileged and peculiar self-knowledge, these alleged instances of privileged and peculiar knowledge show that DES is not our method.

The kinds of cases at issue will be familiar. The smoker who does not judge that smoking is desirable can (allegedly) nevertheless know in a privileged and peculiar way that he desires to smoke. Similarly, the unhealthy eater who does not judge that eating unhealthy food is desirable can (allegedly) nevertheless know in a privileged and peculiar way that he desires to eat unhealthy food. In these cases, one (allegedly) has privileged and peculiar knowledge of a desire that *p*, despite not judging that *p* is desirable. Because one can use DES to attain knowledge that one desires that *p* only if one judges that *p* is desirable, DES cannot be the method used in these cases.

I have two responses to this objection. The first response questions whether the relevant cases involve knowledge. The second response grants that the cases involve knowledge, but questions whether the knowledge is privileged and peculiar. The success of either response undermines the objection.

First, while it might seem that the smoker who does not judge that smoking is desirable nevertheless desires to smoke, he might not; I noted this point at the end of Section 8.2.1. In place of this alleged desire might be the physical urge to smoke. If so, then, contrary to the objection, the smoker does not desire to smoke and so does not know that he desires to smoke. Given this possibility, whether the kind of case described in the objection ever occurs is unclear.

My second response grants for the sake of argument that the smoker *does* desire to smoke and that the unhealthy eater *does* desire to eat unhealthy food. The objection alleges that the smoker and the unhealthy eater have privileged and peculiar knowledge

of their respective desires, despite not judging that the objects of their desires are desirable. What I want to suggest is that the alleged privileged and peculiar nature of this knowledge might be illusory. The knowledge might be arrived at via one's mindreading method.

Consider how much evidence (behavioral and otherwise) these two individuals likely have about their smoking and eating habits. The smoker knows how many packs a week he purchases, he knows how often he steps outside for a cigarette, he (presumably) knows that his friends and family have often spoken of his seemingly strong *desire* for cigarettes and cigars, etc. Given all this evidence, for the smoker to conclude that he desires smoking would be perfectly natural (not to mention rational).

My point is that an extrospective account of knowledge of desire is able to handle cases of alleged privileged and peculiar knowledge of one's desire not attained via using DES; it is able to claim that such self-ascriptions are the products, not of DES, but of an interpretive method of the sort used to attribute mental states to others. Moreover, this move is not *ad hoc*. There is an abundance of evidence that humans engage in such self-interpretation; indeed, some of this evidence was discussed in Chapter 4. No account of privileged and peculiar self-knowledge should deny that individuals might self-attribute mental states on the basis of behavioral and situational evidence about oneself.

Finally, notice that this response cannot be rejected on the grounds that the smoker's knowledge that he desires to smoke *feels* direct, non-interpretive and first-personal. One must keep in mind the discussion from Chapter 4. In that chapter I noted that confabulated self-ascriptions often feel this way, despite being products of an interpretive method. Although, as we are supposing, the smoker's belief is not

confabulated (because it is true), it might nevertheless be the result of the method that subserves confabulation, namely, one's mindreading method.

These two responses do not prove that one who does not judge that *p* is desirable cannot have privileged and peculiar knowledge of her desire that *p*. But I think they significantly under-cut the force of the objection. At the very least, they provide two alternatives that must be ruled out before rejecting the claim that DES is our first-personal method for attaining privileged and peculiar knowledge of our desires.

8.3 A Recipe for Extending My Account

The discussion in the previous section should give one a sense for how my account for knowledge of belief is to be extended to the other propositional attitudes. For a given attitude type, T, there are two tasks: (i) find the appropriate (neutral) antecedent for the epistemic rule, 'if C, then believe that one Ts that *p*' and (ii) defend the claims [T, Judge] and [~T, ~Judge]. The antecedents of BEL and DES are '*p* is true' and '*p* is desirable', respectively. These antecedents were chosen because of the relationship between belief and truth, on the one hand, and desire and desirability, on the other. An extrospective account requires such relationships.

8.4 Intention

One potentially troublesome case in this respect is intention. What should the antecedent of the epistemic rule for intention be? Given that intention is connected to action, the following epistemic rule suggests itself:

INT If I will *Φ*, believe that I intend to *Φ*.

Unfortunately, this suggestion appears deficient. To see why, consider [Intend, Judge] and [~Intend, ~Judge] on this proposal:

- [Intend, Judge] Those who intend to Φ are such that they will tend to judge that they will Φ, upon considering whether they will Φ, if they neither (i) make use of new (external) information, nor (ii) make new use of old (internal) information, when considering whether they will Φ.
- [~Intend, ~Judge] Those who do not intend to Φ are such that they will tend not to judge that they will Φ, upon considering whether they will Φ, if they neither (i) make use of new (external) information, nor (ii) make new use of old (internal) information, when considering whether they will Φ.

While [Intend, Judge] seems quite promising, the same cannot be said of [~Intend, ~Judge]. Consider, for example, a commuter who is running late to work. She judges that she will be late to work, but she does not intend this. If she were to follow INT, she would incorrectly self-ascribe the intention to be late to work. This kind of case is both commonplace and of the kind that [~Intend, ~Judge] claims is relatively infrequent. The desired extension to intention seems only to go half way.

Byrne (2011) attempts to get around this problem by introducing a defeating condition. The condition he has in mind is inspired by G.E.M Anscombe's comment that "[t]he class of intentional actions is a sub-class of [the class of things known without observation]" (1957). The commuter from the previous example knows that she will be late to work *only* through observation. If she had not read the clock, or noticed the slow traffic, she would not have come to know (or simply believe) that she would be late to

work. If Anscombe is correct, this shows that the commuter does not intend to be late for work.[135] And this is correct.

Although Byrne does not speak of 'epistemic rules' in his (2011), what he does say straightforwardly applies to INT. Byrne's suggestion is that one should not use INT if one believes that one's belief that one will *Φ* is based on good evidence. According to Byrne, "I believe that I will wear down my sneakers, but I also believe that I believe this because (and only because) I have good evidence for it" (2011, 218). This, he claims, is why he does not come to believe that he intends to wear down his sneakers (despite knowing that this is what he will do).

The following revised version of INT builds in this defeasibility condition:

INT* If I will *Φ* and if I believe that I will *Φ* on the basis of no evidence, then believe that I intend to *Φ*.[136]

INT* is not neutral, given that the second antecedent is mentalistic; it refers to both belief and evidence. In response to this alleged problem, Byrne offers two points. First, he claims that "from the first-person point of view, an enquiry into one's evidence is (near enough) extensionally equivalent to an enquiry into one's beliefs" (2011, 218).[137] This point, which I find plausible, leads Byrne to conclude that BEL can be used to recognize whether the second antecedent is satisfied. His account of our knowledge of intention thus rests, in part, on his account of our knowledge of belief; he acknowledges this fact.

---

[135] As Byrne points out, this diagnoses why the runner knows, but does not intend, that she will wear down her shoes during tomorrow's marathon (Bratman, 1984) and why the tactical bomber knows, but does not intend, that she will kill many innocent civilians during her air raid (Bennett, 1981).

[136] Recall that he takes judgment to be sufficient for belief. Recognizing that one will *Φ* suffices for one's believing that one will *Φ*. This is why INT* references 'my belief'.

[137] "[O]ne takes *P* to be part of one's evidence just in case one believes that one believes that *P*" (218-19).

Changing INT's antecedent requires also changing [Intend, Judge] and [~Intend, ~Judge]:

- [Intend, Judge] Those who intend to *Φ* are such that they will tend to *both* judge that they will *Φ* and judge that they believe that they will *Φ* on the basis of no evidence, upon considering these matters, if they neither (i) make use of new (external) information, nor (ii) make new use of old (internal) information, when considering these matters.
- [~Intend, ~Judge] Those who do not intend to *Φ* are such that they will tend not to *both* judge that they will *Φ* and judge that they believe that they will *Φ* on the basis of no evidence, upon considering these matters, if they neither (i) make use of new (external) information, nor (ii) make new use of old (internal) information, when considering these matters.

Notice that the previous problem is solved. Although the runner who does not intend to wear out her shoes will judge that she will wear out her shoes, upon considering the matter, she will *not* also judge that she believes that she will wear out her shoes on the basis of no evidence. There is no longer a problem with [~Intend, ~Judge].

Also, [Intend, Judge] remains plausible. One who intends to *Φ* is certainly likely to judge that she will *Φ*, upon considering whether she will *Φ*. In addition, although one who intends to *Φ* might have evidence that she will *Φ*, she will likely judge that she believes that she will *Φ* on the basis of no evidence, upon considering whether she believes that she will *Φ* on the basis of no evidence.

One potential issue with this proposal is that one following INT* must make a judgment concerning the *basis of* one's belief that one will *Φ*. Byrne seems to think this

is unproblematic, but I am not sure. While BEL can give one knowledge of the basis of a particular belief, it seems unable to give one knowledge that the basis of that belief is the *basis of* that belief. But perhaps one can know this by some other means.

I conclude that the extension of my account to intention, although still a work in progress, is hopeful. The versions of [Intend, Judge] and [~Intend, ~Judge] that go along with INT* are plausible. The way in which the remaining details of the extension can be handled should be apparent from the previous two sections.[138]

8.5 Conclusion

In this chapter I have considered whether the account detailed in the previous two chapters can be extended to propositional attitudes other than belief. Various aspects of the account transfer to non-beliefs with relative ease. On the other hand, for any such extension there are issues that arise that are specific to the attitude type in question. Nevertheless, this chapter provides reason to believe that such issues can be adequately dealt with and thus that an extrospective approach to accounting for our apparent privileged and peculiar access to the propositional attitudes is hopeful.

[138] Sarah Paul (2012) offers a much different account of our knowledge of intention. Oversimplifying the account, she argues that one knows one's intentions via an awareness of one's *decisions*. Because the commuter did not decide to run late, she should not be expected to believe that she intends to be late (despite knowing that she will be late). This account thus avoids the problem that necessitated the move to INT*. My account's commitment to neutrality is incompatible with this option, however.

Chapter 9: Concluding Remarks

In Chapter 6 (Section 6.4) I distinguished two projects that one faces when investigating privileged and peculiar access. The epistemological project is to articulate a first-personal method for forming beliefs about one's mental states such that those beliefs qualify as privileged and peculiar knowledge. The psychological project is to make a compelling case that some particular first-personal method is, in fact, the first-personal method that we use to attain privileged and peculiar self-knowledge.

The previous three chapters have focused almost exclusively on the epistemological project. I have argued that particular kinds of reasoning can yield privileged and peculiar knowledge of one's beliefs, desires, and intentions. That this reasoning can produce such knowledge is made possible by various connections that exist between the presence or absence of these attitudes and dispositions to make certain judgments. While this reasoning produces knowledge of one's mind, it nevertheless proceeds via the consideration of non-mental matters. The reasoning is outwardly directed in this sense.

In Chapter 6 I noted that many philosophers find the idea that we attend outwardly, as opposed to inwardly, when considering what we believe, desire, intend, etc. to be descriptively accurate. For such philosophers, it seems true that one turns one's attention outward when considering whether one has these attitudes. While I share this sentiment, I unfortunately do not know how to change the mind of someone who does not. In some cases I think it is sufficient to simply make one aware of the possibility; upon reflecting on the matter, one is able to appreciate the pull of the idea. Of course in other cases this is insufficient.

I suspect such resistance is primarily grounded in the suspicion that if this is how we form beliefs about our minds, then such beliefs will surely not amount to knowledge and perhaps will typically be false. The best way to combat such resistance is to make the case that extrospection can accommodate not only self-knowledge, but privileged and peculiar self-knowledge. That is, the best way to fight such resistance is to engage in the epistemological project. Supposing this is correct, I hope that the previous three chapters soften the resistance to the idea that extrospection is descriptively accurate of us and so *explains* our privileged and peculiar self-knowledge.

Finally, recall the feature of the extrospective approach that many find so attractive, namely, its economy. It does not posit any inwardly directed mechanisms or metaphysically special relations. According to this approach, we are able to attain privileged and peculiar self-knowledge simply by virtue of our ability to think and reason about the world. Although this is by no means a conclusive reason to select the extrospective account defended in the previous three chapters over an introspective account, it is perhaps an additional consideration in favor of making this selection.

References

Alston, W. (1971), "Varieties of Privileged Access", *American Philosophical Quarterly*, 8, 223-41.

Anscombe, G. E. M. (1957), *Intention*, Ithaca, NY: Cornell University Press.

Austin, J. L. (1953), "How to Talk - Some Simple Ways", *Proceedings of the Aristotelian Society*, 53, 227-46.

Armstrong, D.M. (1968/1993), *A Materialist Theory of Mind*, New York: Humanities Press.

Armstrong, D. M. (1973), *Belief, Truth and Knowledge*, Cambridge: Cambridge University Press.

Armstrong, D.M. (1981), *The Nature of Mind and Other Essays*, Ithaca, NY: Cornell University Press.

Aydede, M. (2003), "Is Introspection Inferential?", in B. Gertler (ed.), *Privileged Access: Philosophical Accounts of Self Knowledge*, Burlington, VT: Ashgate Publishing Company.

Ayer, A.J. (1940), *The Foundations of Empirical Knowledge*, London: Macmillan.

Bennett, J. (1981), "Morality and Consequences", *The Tanner Lectures on Human Values*, 2, 45–116.

Bratman, M. E. (1984), "Two Faces of Intention", *Philosophical Review*, 93, 375–405.

Boghossian, P. (1989), "Content and Self-Knowledge", *Philosophical Topics*, 17, 5-26.

BonJour, L. and E. Sosa (2003), *Epistemic Justification: Internalism versus Externalism, Foundations versus Virtues*, Malden, MA: Blackwell Publishing Ltd.

Boyle, M. (2011), "Transparent Self-Knowledge", *Proceedings of the Aristotelian Society Supplementary Volume*, LXXXV, 223-41.

Brasil-Neto, J., A. Pascual-Leone, J. Valls-Sole, L. Cohen, and M. Halett (1992), "Focal Transcranial Magnetic Stimulation and Response Bias in a Forced Choice Task", *Journal of Neurology, Neurosurgery, and Psychiatry*, 55, 964-66.

Brown, J. (1995), "The Incompatibility of Individualism and Privileged Access", *Analysis*, 55, 149-56.

Burge, T. (1979), "Individualism and the Mental", *Midwest Studies in Philosophy*, 4, 73-122.

Burge, T. (1988), "Individualism and Self-Knowledge", *The Journal of Philosophy*, 85, 649-63.

Byrne, A. (2005), "Introspection", *Philosophical Topics*, 33, 79-104.

Byrne, A. (2011), "Transparency, Belief, Intention", *Proceedings of the Aristotelian Society Supplementary*, 85(1), 201-21.

Carruthers, P. (2009), "How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition", *Behavioral and Brain Sciences*, 32, 121-182.

Carruthers, P. (2010), "Introspection: Divided and Partly Eliminated", *Philosophy and Phenomenological Research*, LXXX(1), 76-111.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, Oxford: Oxford University Press.

Chalmers, D. (2003), "The Content and Epistemology of Phenomenal Belief", in Q. Smith and A. Jokic (eds.), *Consciousness: New Philosophical Essays*, Oxford: Oxford university Press.

Chisholm, R. (1942), "The Problem of the Speckled Hen", *Mind*, 51, 368-73.
Corcoran, R., G. Mercer, and C.D. Frith (1995), "Schizophrenia, Symptomatology and Social Inference: Investigating Theory of Mind in People with Schizophrenia", *Schizophrenia Research*, 17, 5-13.
Craig, A.D. (2002), "How Do You Feel? Interoception: The Sense of the Physiological Condition of the Body", *Nature Reviews Neuroscience*, 3, 655-66.
Davidson, D. (1983), "A Coherence Theory of Truth and Knowledge", in D. Henrich (ed.), *Kant oder Hegel?*, Stuttgart: Klett-Cotta.
Donnellan, K.S. (1966), "Reference and Definite Descriptions", *Philosophical Review*, 75, 281-304.
Dretske, F. (1971), "Conclusive Reasons", *Australasian Journal of Philosophy*, 49(1), 1–22.
Dretske, F. (1994), "Introspection", *Proceeding of the Aristotelian Society*, 94, 263-78.
Dretske, F. (1995), *Naturalizing the Mind*, Cambridge, MA: MIT Press.
Dretske, F. (1999), "The Mind's Awareness of Itself", *Philosophical Studies*, 95, 103-24.
Dretske, F. (2003a), "How Do You Know You are Not a Zombie?", in B. Gertler (ed.), *Privileged Access: Philosophical Accounts of Self Knowledge*, Burlington, VT: Ashgate Publishing Company.
Dretske, F. (2003b), "Externalism and Self-Knowledge", in S. Nuccetelli (ed.), *New Essays on Semantic Externalism and Self-Knowledge*, Cambridge, MA: MIT Press, Bradford Books.
Dretske, F. (2012a), "Awareness and Authority: Skeptical Doubts about Self-Knowledge", in D. Smithies and D. Stoljar (eds.), *Introspection and Consciousness*, Oxford: Oxford University Press.
Dretske, F. (2012b), "I Think I Think, Therefore I am – I Think: Skeptical Doubts about Self-Knowledge, in J. Liu and J. Perry (eds.), *Consciousness and the Self: New Essays*, NY: Cambridge University Press.
Eagly, A. and S. Chaiken (1993), *The Psychology of Attitudes*, Fort Worth, TX: Harcourt, Brace, and Jovanovich.
Engelbert, M. and P. Carruthers (2010), "Introspection", *WIREs Cognitive Science*, 1, 245-53.
Evans, G. (1982), *The Varieties of Reference*, J. McDowell (ed.), Oxford: Oxford University Press.
Feldman, R. (2006), "BonJour and Sosa on Internalism, Externalism, and Basic Beliefs", *Philosophical Studies*, 131, 713-728.
Fernandez, J. (2003), "Privileged Access Naturalized", *The Philosophical Quarterly*, 53, 352-72.
Fernandez, J. (2007), "Desire and Self-Knowledge", *Australasian Journal of Philosophy*, 85(4), 517-36.
Fernandez, J. (2013), *Transparent Minds: A Study of Self-Knowledge*, Oxford: Oxford University Press.
Fodor, J.A. (1975), *The Language of Thought*, Scranton, PA: Crowell.
Fodor, J.A. (1987), "Mental Representation: An Introduction", in N. Rescher (ed.), *Scientific Inquiry in Philosophical Perspective*, N.Y.: University Press of America.

Freund, H., B. Sabel, and O.W. Witte (eds.) (1997), *Advances in Neurology*: *Volume 73 Blain Plasticity*, Philadelphia, PA: Lippincott-Ravin Publishers.
Frith, C.D. (1992), *The Cognitive Neuropsychology of Schizophrenia*, Hillsdale, N.J.: Lawrence Erlbaum and Associates.
Frith, C.D. (1994), "Theory of Mind in Schizophrenia", in A. David and J. Cutting (eds.), *The Neuropsychology of Schizophrenia*, 147-162, Hillsdale, N.J.: Lawrence Erlbaum and Associates.
Frith, C.D. (2012), "Explaining Delusions of Control: The Comparator Model 20 Years On", *Consciousness and Cognition*, 21, 52–54.
Frith, C.D., S-J Blakemore, and D. Wolpert (2000a), "Explaining the Symptoms of Schizophrenia: Abnormalities in the Awareness of Action", *Brain Research Reviews*, 31, 357-63.
Frith, C.D., B. Wolpert, and D.M. Wolpert (2000b), "Abnormalities in the Awareness and Control of Action", *Philosophical Transactions of the Royal Society Biological Sciences*, 355, 1771-1788.
Fumerton, R. (1995), *Metaepistemology and Skepticism*, Lanham: Rowman and Littlefield.
Gazzaniga, M. (1995), "Consciousness and the Cerebral Hemisphere", in M. Gazzaniga (ed.) *The Cognitive Neurosciences*, 1391-1400, Cambridge, MA: MIT Press.
Gazzaniga, M. (2000), "Cerebral Specialization and Interhemispheric Communication: Does the Corpus Callosum Enable the Human Condition?", *Brain*, 123, 1293-1326.
Gendler, T. S. (2008a), "Alief and Belief", *Journal of Philosophy*, 105, 634–63.
Gendler, T. S. (2008b), "Alief in Action (and Reaction)", *Mind and Language*, 23, 552–85.
Gertler, B. (2001), "Introspecting Phenomenal States", *Philosophy and Phenomenological Research*, 63, 305-28.
Gertler, B. (ed) (2003), *Privileged Access – Philosophical Accounts of Self-Knowledge*, Burlington, VT: Ashgate Publishing Company.
Gertler, B. (2011a), *Self-Knowledge*, Oxford: Routledge.
Gertler, B. (2011b), "Self-Knowledge and the Transparency of Belief", in Hatzimoysis (ed.), *Self-Knowledge*, Oxford: Oxford University Press.
Gertler, B. (2012), "Renewed Acquaintance", in D. Smithies and D. Stoljar (eds.), *Introspection and Consciousness*, Oxford: Oxford University Press.
Goldman, A. I. (1979), "What Is Justified Belief?", in G. Pappas (ed.), *Justification and Knowledge*, Dordrecht: Reidel.
Goldman, A. I. (1986), *Epistemology and Cognition*, Cambridge, MA: Harvard University Press.
Goldman, A.I. (1993), "The Psychology of Folk Psychology", *Behavioral and Brain Sciences* 16, 15-28.
Goldman, A.I. (1989), "Interpretation Psychologized", *Mind and Language*, 4, 161-185.
Goldman, A.I. (2000), "The Mentalizing Folk", in D. Sperber (ed.), *Metarepresentations*, Oxford: Oxford University Press.
Goldman, A.I. (2006), *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*, Oxford: Oxford University Press.

Gopnik. A. (1993), "How we know our minds: The illusion of first-person knowledge of intentionality", *Behavioral and Brain Sciences*, 16, 1-14.
Gopnik, A. and H. M. Wellman (1994), "The theory theory", in L. Hirschfeld and S. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture*, Cambridge: Cambridge University Press.
Gordon, R. (1986), "Folk Psychology as Simulation", *Mind and Language*, 1, 158-171.
Gordon, R. (1995), "Simulation Without Introspection or Inference from Me to You", in M. Davies and T. Stone (eds.), *Mental Simulation*, Cambridge, MA: Blackwell.
Gordon, R. (2007), "Ascent Routines for Propositional Attitudes", *Synthese*, 159, 151-65.
Harman, G. (1999), "Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error", *Proceedings of the Aristotelian Society*, 99, 315-31.
Harman, G. (2009), "Scepticism About Character Traits", *Journal of Ethics*, 13, 235–42.
Heil, J. (1988), "Privileged Access', *Mind*, 42, 238-51.
Hunter, D. (2009), "Beliefs and Dispositions", *Journal of Philosophical Research*, 34, 243-262.
Hunter, D. (2011), "Alienated Belief", *Dialectica*, 65, 221-240.
James, W. (1884), "On Some Omissions of Introspective Psychology", *Mind*, 33, 1-11.
Levin, H. and J. Grafman (eds.) (2000), *Cerebral Reorganization of Function After Brain Damage*, Oxford: Oxford University Press.
Locke, J. (1690/1975), *An Essay Concerning Human Understanding*, P.H. Nidditch (ed.), Oxford: Clarendon Press.
Lycan, W.G. (1987). *Consciousness*, Cambridge, MA: Bradford Books/MIT Press.
Lycan, W.G. (1996). *Consciousness and Experience*, Cambridge, MA: Bradford Books/MIT Press.
Lycan, W.G. (1999), "Dretske On the Mind's Awareness of Itself", *Philosophical Studies*, 95(1-2), 125-133.
McKinsey, M. (1991), "Anti-Individualism and Privileged Access", *Analysis*, 51, 9-16.
McLaughlin, B.P. and M. Tye (1998), "Is Content-Externalism Compatible with Privileged Access?", *Philosophical Review*, 107(3), 349-80.
Mellor C. (1970), "First Rank Symptoms of Schizophrenia", *British Journal of Psychology*, 117, 15-23.
Moran, R. (2001), *Authority and Estrangement – An Essay on Self-Knowledge*, Princeton, NJ: Princeton University Press.
Moran, R. (2003), "Responses to O'Brien and Shoemaker", *European Journal of Philosophy*, 11, 402-19.
Nichols, S. and S.P. Stich (2003), *Mindreading: An Integrated Account of Pretense, Self-Awareness, and Understanding of Other Minds*, Oxford: Oxford University Press.
Nisbett, R. and T. Wilson (1977), "Telling More than We Can Know", *Psychological Review*, 84, 231-95.
Paul, S.K. (2012), "How We Know What We Intend", *Philosophical Studies*, 161(2): 327-46.
Peacocke, C. (1999), *Being Known*, Oxford: Oxford University Press.

Pickup, G.J. and C.D. Frith (2001), "Theory of Mind Impairments in Schizophrenia: Symptomatology, Severity and Specificity", *Psychological Medicine*, 31, 207-220.
Putnam, H. (1975), "The Meaning of Meaning", *Philosophical Papers, Vol. II: Mind, Language, and Reality*, Cambridge: Cambridge University Press.
Ramsey, F. P. (1931), "Knowledge", in R. B. Braithwaite (ed.), *The Foundations of Mathematics and Other Essays*, New York: Harcourt Brace.
Rey, G. (1997), *Contemporary Philosophy of Mind*, Cambridge, MA: Blackwell Publishers Inc.
Ryle, G. (1949), *The Concept of Mind*, Chicago: The University of Chicago Press.
Russell, B. (1912), *Problems of Philosophy*, NY: Henry Holt and Company.
Ryle, G. (1949). *The Concept of Mind*. Chicago: The University of Chicago Press.
Schultheiss, O.C. and J.C. Brunstein (1999), "Goal Imagery: Bridging the Gap between Implicit Motives and Explicit Goals", *Journal of Personality,* 61, 1-38.
Schwitzgebel, E. (2001), "In-Between Believing", *Philosophical Quarterly*, 51, 76-82.
Schwitzgebel, E. (2002), "A Phenomenal, Dispositional Account of Belief", *Nous*, 36, 249-275.
Schwitzgebel, E. (2010), "Acting Contrary to Our Professed Beliefs, or the Gulf Between Occurrent Judgment and Dispositional Belief", *Pacific Philosophical Quarterly*, 91, 531-53.
Schwitzgebel, E. (2012), "Knowing Your Own Beliefs", in D. Hunter (ed.), *Belief and Agency*, Calgary: Calgary University Press.
Shapiro, L. (2000), "Multiple Realizations", *Journal of Philosophy*, 97, 635-654.
Shapiro, L. (2004), *The Mind Incarnate*, Cambridge, MA: MIT Press, Bradford Books.
Shoemaker, S. (1988), "On Knowing One's Own Mind", *Philosophical Perspectives, 2, Epistemology*, 183-209.
Shoemaker, S. (1994a), "Self-Knowledge and 'Inner Sense': Lecture I: The Object Perception Model", *Philosophy and Phenomenological Research*, LIV, 249-269.
Shoemaker, S. (1994b), "Self-Knowledge and 'Inner Sense': Lecture II: The Broad Perceptual Model", *Philosophy and Phenomenological Research*, LIV, 271-290.
Shoemaker, S. (1996), *The First-Person Perspective and Other Essays*, Cambridge: Cambridge University Press.
Sosa, E. (1999), "How to Defeat Opposition to Moore", *Philosophical Perspectives*, 13, 141-53.
Sosa, E. (2003), "Privileged Access", in Q. Smith and A. Jokic (eds.), *Consciousness: New Philosophical Essays*, Oxford: Oxford University Press.
Sprong, M, P. Schothorst, E. Vos, J. Hox, and H. Van Engeland (2007), "Theory of Mind in Schizophrenia: Meta-Analysis", *British Journal of Psychiatry*, 191, 5-13.
Steinberg, J.R. (2010), "Dispositions and Subjunctives", *Philosophical Studies*, 148, 323-341.
Steinberg, J. R. (2011), "Dispositions, Moral Judgments, and What We're Motivated to Do", *Canadian Journal of Philosophy—Supplement 35: Belief and Agency*, 1-24.
Steinberg, J.R. Unpublished. "Dispositionalism and Intentional Mental States", manuscript currently under submission.

Tye, M. (1995), *Ten Problems of Consciousness*, Cambridge, MA: MIT Press, Bradford Books.
Tye, M. (2003), "Representationalism and the Transparency of Experience", in B. Gertler (ed.), *Privileged Access: Philosophical Accounts of Self Knowledge*, Burlington, VT: Ashgate Publishing Company.
Wegner, D. and T. Wheatley (1999), "Apparent Mental Causation: Sources of the Experience of the Will", *American Psychologist*, 54, 480-91.
Wiffen, B. and A. David (2009), "Metacognition, Mindreading, and Insight in Schizophrenia", *Behavioral and Brain Sciences*, 32, 161-62.
Williamson, T. (2001), *Knowledge and Its Limits*, Oxford: Oxford University Press.
Wilson, T. and E.W. Dunn (2004), "Self-Knowledge: Its Limits, Value, and Potential for Improvement", *Annual Review of Psychology*, 55, 493-518.
Wittgenstein, L. (1953), *The Philosophical Investigations*, trans. G.E.M. Anscombe, Oxford: Blackwell.
Wright, C. (1989), "Wittgenstein's Later Philosophy of Mind: Sensation, Privacy, and Intention", *The Journal of Philosophy*, 86, 622-34.
Zawidzki, T. (2008), "The Function of Folk Psychology: Mind Reading or Mind Shaping?", *Philosophical Explorations*, 11(3), 193–210.
Zimmerman, A. (2006), "The Nature of Belief", *Philosophical Studies*, 128, 337-79.

Figure 1. The Monitoring Mechanism Account of Attitudinal Detection (adapted from Nichols and Stich 2003, 162)