Self-Knowledge in a Natural World

SELF-KNOWLEDGE IN A NATURAL WORLD

A Dissertation Presented

by

JEREMY CUSHING

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2012

Philosophy

SELF-KNOWLEDGE IN A NATURAL WORLD

A Dissertation Presented

by

JEREMY CUSHING

Approved as to style and content by:

_____
Hilary Kornblith, Chair

_____
Louise Antony, Member

_____
Joseph Levine, Member

_____
Jill de Villiers, Member

_____
Hilary Kornblith, Department Head
Philosophy

ACKNOWLEDGEMENTS

During my time working on this dissertation, I received help and support from more people than I can thank individually. I am grateful to all of the teachers, graduate students, undergraduates, and friends who have kept me committed to learning, but some deserve special thanks.

I am beyond grateful to my dissertation director, Hilary Kornblith. The influence of his ideas on my own has been enormous, but his personal impact on me has been even more meaningful. His passion and energy for philosophy, his professionalism, and his prodigious work ethic will, for the rest of my life, serve as ideals to strive for.

I am grateful to the other members of my committee, Joe Levine, Louise Antony, and Jill de Villiers, as well, for important conversation, feedback, and criticism.

I am grateful to have interacted, formally and informally, with so many members of the UMass philosophy community, and for being allowed to present virtually all the material in this dissertation in various dissertation seminars and graduate student colloquia. Among professors not on my committee, standouts were Lynne Baker, Phil Bricker, Fred Feldman, Gary Matthews, and Jonathan Schaffer.

Graduate students who were memorably helpful, professionally, philosophically, or personally include, Kristoffer Ahlstrom, Jake Bridge, Heidi Buetow, Sam Cowling, Dan Doviak, Jeff Dunn, Ed Ferrier, Lowell Friesen, Brandt van der Gaast, Chris Heathwood, Justin Klocksiem, Barak Krakuer, Meghan Masto, Kris McDaniel, Kirk Michaelian, Gabe Rabin, Jason Raibley, Creighton Rosenthal, Stephan Torre, and Kelly Trogdon.

I owe a special thanks to two friends that could have been philosophers, if they had had the stomach for it. Ebru Kardan and Alex Howell were each there for me when it mattered most.

I would also like to thank both my parents for their love and support. My mother, Susan, has inspired and fostered my intellectual curiosity throughout my life. My father, Chuck, has been an unwavering model of the sort of man I would like to be.

I reserve the largest thanks for my partner, Amy Ferrer. She has provided me with every kind of support since the day we met. Her love and faith in me have made all the difference.

ABSTRACT

SELF-KNOWLEDGE IN A NATURAL WORLD

FEBRUARY 2012

JEREMY CUSHING, B.A., UNIVERSITY OF DELAWARE

M.A., UNIVERSITY OF MASSACHUSETTS AMHERST

PH.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by Professor Hilary Kornblith

In this dissertation, I reconcile our knowledge of our own minds with philosophical naturalism. Philosophers traditionally hold that our knowledge of our own minds is especially direct and authoritative in comparison with other domains of knowledge. I introduce the subject in the first chapter.

In the second and third chapters, I address the idea that we know our own minds directly. If self-knowledge is direct, it must not be grounded on anything more epistemically basic. This creates a puzzle for all epistemologists. For the naturalist, the puzzle is especially tricky. To say that self-knowledge has no epistemic ground threatens the naturalist's ability to understand it as psychologically real. I argue that the idea that self-knowledge is direct is not well motivated and that models of direct self-knowledge have fundamental problems.

In the fourth and fifth chapters, I examine first-person authority. I distinguish between epistemic authority, or being in a better position than others to know, and non-epistemic authority, or being immune to challenge according to some conventional norm. I argue that we have only limited epistemic authority over our own minds. I then consider

whether there may be an interesting non-epistemic authority attached to the first-person perspective. This would locate first-person authority in connection with our responsibility for our own minds. I argue that this sort of authority may exist, but is unlikely to threaten naturalism without further anti-naturalist commitments in the philosophy of mind.

In the final two chapters, I explore the possibility that the underlying disagreements between naturalists and anti-naturalists are about the nature of belief. I consider what failures of self-knowledge might demonstrate about the nature of belief. I show how, with the proper understanding of belief, a theory of self-knowledge can assuage some of these worries. Having adopted a conception of belief that makes sense for philosophy and empirical psychology, I outline a positive theory of self-knowledge and suggest directions for future research.

TABLE OF CONTENTS

CHAPTER 1

AN INTRODUCTION TO SELF-KNOWLEDGE

1. A Brief History of Self-Knowledge

This dissertation is an attempt to reconcile traditional philosophical concerns about self-knowledge with naturalistic approaches to philosophy. In it, I argue that these concerns can be handled, and in fact are best handled, by a naturalistic philosophy. Before beginning that task, it will be useful to consider a very short sketch of the history of thought about self-knowledge to better understand the subject of the dissertation.

Topics that fit comfortably under a label like 'self-knowledge' have been with western philosophy since the beginning. Supposedly the temple at Delphi contained a carved dictum of the Greek words for the phrase, "Know Thyself". While it is difficult to know exactly what such an old proverb meant to the people of that time, I think it shows something of the value that the ancient Greeks placed on a kind of self-knowledge. Roughly, my take on it is that the dictum was supposed to remind one to shield oneself against hypocrisy by monitoring one's own behaviors. Self-knowledge of this sort is very far from the topic of this dissertation, but I think there is an intellectual track one can follow from this old adage to the contemporary landscape of the issues surrounding self-knowledge.

The dictum is often associated with the founder of the western philosophical tradition, Socrates. In addition to the carved dictums, the temple at Delphi was also home to a legendary oracle. According to Plato's account of the life of Socrates, the oracle at Delphi proclaimed that no one in Athens was wiser than Socrates. Socrates, modest in show if not in truth, claimed not to understand the pronouncement. He said of himself

that he knew virtually nothing. He embarked on a project to find Athenians that were wiser than him. In testing the oracle's claim, he came to realize that while he didn't seem to know much at all, the people around him that claimed to have great knowledge also knew virtually nothing.[1] Yet, they were in some sense worse off than Socrates because they had no idea how ignorant they really were. He came to think that his wisdom, true wisdom, was in knowing what he did not know. Thus, for Socrates, wisdom was a kind of self-knowledge. He advocated more than just seeking knowledge. He stressed the importance of examining one's ideas to see if they should be held with confidence and ridiculed those that unthinkingly trusted their own understanding. In essence, he took the dictum to be indicating that one should examine one's own beliefs in order to guide against hypocrisy in one's claims to knowledge, perhaps further internalizing what was already a dictum aimed at encouraging self-conscious reflection.

In skipping ahead in the history of philosophy all the way to Descartes, I should note that I am committing a kind of gross oversight that rightly offends historians of philosophy. However, it is beyond the scope of this dissertation to trace out a full history of thought about self-knowledge. What is important and undeniable is that Descartes has clearly become the central figure on the subject of self-knowledge in analytic philosophy.[2] I see Descartes's skeptical project as one rooted in the sort of self-examination encouraged by Socrates. For Descartes, it was of the utmost importance that he examined the foundations of his beliefs to judge what he might legitimately be said to know. Descartes, like Socrates, is unwilling to recognize himself as knowing unless his

[1] See Plato's Apology in Plato 2002.

[2] It is somewhat difficult to find a clear statement in Descartes writings of the doctrines he is later taken to have held. However, the clear starting point for discussion of Descartes and self-knowledge is the Meditations. See Descartes 1993.

beliefs can stand up to scrutiny. However, unlike the Platonic Socrates, who scrutinizes the beliefs of others, Descartes plays his own gadfly and scrutinizes his own beliefs. He finds reason to doubt many things that most would pre-reflectively regard as paradigms of knowledge. He doubts whether he is at that moment awake rather than dreaming and thereby comes to doubt all of his beliefs based on his sensory experience. Interestingly, as Descartes finds reason to doubt more and more, it is ultimately his own mind that he judges to be immune to doubt. For Descartes, a certain kind of self-knowledge is the only thing that can withstand the internal scrutiny suggested by the Socratic ideal.

To be clear, Descartes is not rejecting the Socratic idea that most people do not know the things they take themselves to know. Rather, Descartes is locating what he takes to be one domain in which everyone really does know what they take themselves to know. For Descartes, the fact that one exists as a thinking thing is indubitable. Moreover, Descartes seems to have ultimately held that our knowledge of our own mind was infallible and perfectly transparent. At the very least, this has been an incredibly common interpretation of Descartes's views: that he thought we know perfectly well everything that occurs in our minds. Ultimately, this insight led Descartes to advance a view about the metaphysics of mind that was widely influential. Cartesian Dualism, which held mind and body to be metaphysically distinct substances, was partially justified by appeal to the apparently different modes of access we have to physical bodies and to our own minds. The Cartesian view of mind is one in which you are presented with your thoughts in such a way that whenever you have a thought, you know exactly which thought you have.

While not many contemporary philosophers advocate a Cartesian view of mind, many share with Descartes the intuition that there is something fundamentally special

about our access to our own minds. For these philosophers, self-knowledge is special. For many, consideration of self-knowledge is thought to reveal something distinctive about the mental that, even if it does not show the mental to be a metaphysically distinct substance, does make reductive analysis of mind difficult or impossible and creates problems for naturalistic accounts of mind and knowledge.

Another important figure in the history here is John Locke. His understanding of how we know our mind was as a kind of internal perception. In contemporary discussion, Locke's views are best applied to the topic of consciousness. The perceptual metaphor makes the most sense when the questions are surrounding our knowledge of our conscious states. This will not be the focus of this dissertation, so Locke himself is not much discussed. However the idea that self-knowledge involves a kind of inner sense, or introspection, remains quite influential throughout topics in self-knowledge. So Locke is an important forerunner to virtually all contemporary discussion of self-knowledge and I would be remiss if I did not mention him here.[3]

The beginning of the overthrow of the Cartesian epistemology of mind was in the late nineteenth and early twentieth centuries. While there are many important figures in the early days of psychology as a discipline, it is safe to say that no one did more to popularize the idea of the unconscious mind than Sigmund Freud. Freud's conception of the unconscious mind was one in which certain mental events occurred without the awareness of the person whose mind they occurred in. In addition, this unconscious mind sometimes worked to deceive a person about that person's own mind. These ideas amounted to a rejection of Cartesian infallibility and transparency. By spreading

[3] See Locke 1975.

acceptance of the unconscious mind, Freud generated (or re-generated) a problem of self-knowledge. In the Cartesian conception of mind, self-knowledge comes easily. It can scarcely be said to 'come' at all. After Freud, it began to look as though it was sometimes very difficult to get self-knowledge. A whole industry of people employing analysts as assistants in acquiring self-knowledge has since emerged.

While Freudians attacked the infallibility of the Cartesian picture of mind, behaviorists rejected the separate metaphysics of mind and body. While the behaviorist program of analyzing mental terms as descriptive shorthand for behaviors failed, Cartesian dualism has remained out of favor. Gilbert Ryle, a philosophical forerunner of the psychological behaviorists, proposed a behaviorist account of self-knowledge.[4] While the failure of behaviorism did not completely reinvigorate the Cartesian view of mind, the backlash against Ryle's particular account of self-knowledge may have partly reinvigorated the Cartesian view of our access to our own mind. Ryle held that we know our own minds the same way that we know the minds of others, by inference from observable behavior. Most contemporary philosophers view this suggestion as wildly implausible. In many contemporary discussions of self-knowledge, Ryle is assumed to be wrong with little or no argument and the project of giving a philosophical account of self-knowledge is taken to be explaining how our self-knowledge is especially secure despite not being grounded in any inference. The differences between our knowledge of our own mind and other minds are spelled out in Cartesian terms. This has been a central question in the recent literature on the epistemology of self-knowledge: How can we explicate the

[4] Ryle 1949.

notion of an authoritative and direct access to our own minds which does not depend on a picture of the mind as infallibly transparently conscious?

My summary of the history of thought on self-knowledge has been admittedly simplistic and brief. However, the point of this summary is merely to set the scene for contemporary discussion. To complete the scene setting, I must mention one more philosopher: Ludwig Wittgenstein. Wittgenstein's influence on thought about self-knowledge, like his influence throughout philosophy, is wide ranging and complex. On the one hand, some think of Wittgenstein as a kind of behaviorist about the mind, much in line with Ryle. Anthony Kenny, a prominent scholar in the philosophy of history and former student of Ryle, claims that Wittgenstein's posthumously published *Philosophical Investigations* contained many of the ideas in found in Ryle's work, "presented with far greater subtlety and profundity," and hints that Ryle may have gotten many of his ideas from Wittgenstein.[5]

However much Ryle took from Wittgenstein, there is some apparent disagreement between the two with respect to self-knowledge. While Ryle treated self-knowledge on a par with knowledge of the external world, Wittgenstein's writings have been taken by some to support some notion of a special first-person authority. Wittgenstein held that we could not make certain kinds of mistakes about our own mind that we can make about the minds of others. It is not clear, however, how to understand the kind of authority that Wittgenstein advocated. It is not clear, for instance, whether or not one should count Wittgenstein as thinking we have secure self-knowledge or as incapable of knowledge of our own minds at all. Because of the difficulties in interpreting Wittgenstein, I will not be

[5] Kenny 2007, p. 64.

directly addressing his views. However, his influence is profound on many of the philosophers that I do discuss.

2. Contemporary Discussion of Self-Knowledge

In contemporary analytic philosophy, there are two dominant threads of discussion surrounding self-knowledge. The first draws much influence from Wittgenstein and takes very seriously some aspects of the picture of the mental inherited from Descartes. The second thread is one contained in interdisciplinary work done by empirical psychologists, philosophers, and cognitive scientists. These two threads have operated largely separately. This dissertation represents an effort to bridge some of the gaps between the two threads. In the remainder of this section, I will summarize the philosophical literature comprising these two threads.

Wittgenstein's thought on self-knowledge is directly engaged by Crispin Wright and John McDowell in an exchange over interpreting Wittgenstein.[6] Wright argues that Wittgenstein's ultimate position on self-knowledge, and on essentially all philosophical discussion, is to reject the need for explanation of features of common discourse. In the case of self-knowledge, the features apparently in need of explanation are attributed to present tense first-person assertions of one's own mental states, which Wright and others call 'avowals'. Avowals are supposedly treated as authoritative, groundless, and transparent. Authority is presumption of truth for any apparently sincere avowal. Transparency is a presumption in favor of a person's ability to avowal any particular states that he or she has. Groundlessness is the inappropriateness of asking a person to

[6] See Wright 1998 and McDowell 1998.

provide reasons for his or her avowals. Wright thinks that Wittgenstein accepted that our linguistic practices do contain these features while rejecting the adequacy of or even the need for philosophical explanations of them. It is hard to pin down a positive view of Wright's, but it is important to note the pervasiveness of this picture of self-knowledge as authoritative, transparent, and groundless. While some of the literature sheds the terminology of 'avowals' and avoids the explicitly linguistic focus of Wittgenstein, this picture is something of a default for much discussion of self-knowledge in philosophy.

Another very useful feature of Wright's essay is that it provides a detailed examination of a kind of expressivism about avowals that is sometimes attributed to Wittgenstein. Roughly, the idea is that avowals are contentless expressions of our mental states and not the kinds of contentful assertions that they appear to be. This sort of expressivism has a variety of problems, but modified versions of it have been recently defended, most notably by Dorit Bar-On.[7] In her view, avowals are products of expressive acts that still manage to have the semantic content that we would ordinarily expect them to have. I find much to like in her picture of the semantics of avowals, although I am ultimately skeptical about the relation of her project to genuine self-knowledge.

Since Wittgenstein, one of the most prominent voices on self-knowledge has been Donald Davidson.[8] Davidson sought to explain first-person authority via a presumption that a speaker cannot misinterpret his or her own language and thus knows what he or she believes in virtue of knowing which propositions he or she holds true. Davidson was also an active participant in the debate surrounding the compatibility of privileged access and

[7] See Bar-On and Long 2001 and Bar-On 2004.
[8] See Davidson 1984 and Davidson 1987.

content externalism, which occupied many philosophers, such as Tyler Burge and Paul Boghossian, among others, during the late nineteen eighties and nineties.[9] Roughly, the issue there is how we know our own beliefs given that the content of those beliefs might be fixed by external factors. These worries grew from an example of Hilary Putnam's often referred to as Twin Earth.[10] Putnam argued that our word water refers to H_2O and that, if one were transported to a near replica of Earth in which there was a substance, XYZ, which functioned just like water does on earth, one's term 'water' would fail to refer to XYZ and one would speak falsely in saying things like, "This glass is full of water." If Putnam is right, then there is an interesting question about our knowledge of our own mental content. How, exactly, could I tell whether my beliefs are about water (H_2O) or twin-water (XYZ) without a detailed survey of the environment? This puzzle has received much attention in the philosophical literature, but I will be largely silent on it. In general, I seek to avoid issues of mental content and focus on the processes underlying belief formation and the nature of belief and meta-belief without worrying about content.

Another extremely important philosopher on the subject of self-knowledge is Sydney Shoemaker.[11] His view is unique and has been influential. Shoemaker has backed the idea that self-knowledge carries special authority. For Shoemaker, this special authority comes from the fact that mental states in rational creatures constitute knowledge of them. Shoemaker argues against the idea that there is any special process or procedure by which we acquire self-knowledge. Rather, it comes along for free with our first order

[9] For Burge, see Burge 1979, 1988, and 1996 all of which were reprinted in Ludlow 1998. For Boghossian, see Boghossian 1989. Reprinted in Boghossian 2008 with many other relevant papers.
[10] See Putnam 1981, especially chapter 2.
[11] See Shoemaker 1988, 1990, 1994a, 1994b, 1994c, and 1995 all reprinted in Shoemaker 1996.

states because of the sorts of creatures that we are. I think that Shoemaker's account does remarkably well at capturing the sorts of intuitions that philosophers concerned with self-knowledge have had. Unfortunately, I think it is ultimately somewhat self-undermining. The right conclusion to draw from Shoemaker's arguments, I argue in the dissertation, is that self-knowledge is not as valuable or pervasive as we take it to be.

Richard Moran has a view heavily influenced by Shoemaker.[12] Moran's view is never spelled out in explicit detail, but it agrees with Shoemaker's criticism of perceptual models of self-knowledge. The view that Moran favors is one in which self-knowledge is tied to our ability to make up our minds. Moran is fond of claiming that there is a difference between knowledge based on discovery and knowledge based on decision. Moran attempts to use this distinction to present a sketch of a theory of self-knowledge that is inherently first-personal. If nothing else, Moran's work provides a valuable service in bringing issues of agency and doxastic voluntarism to discussions about self-knowledge.

A much different version of a constitutive account has been developed by Akeel Bilgrami.[13] Bilgrami's account is derived from a normative notion of intentionality proposed by Strawson.[14] Since Bilgrami proposes that a subject must be capable of epistemic criticism in order to count as having belief, Bilgrami ends up requiring self-knowledge in order to have belief as well. Being subject to epistemic criticism, Bilgrami thinks, requires knowledge of one's mental states. In Bilgrami's account, self-knowledge is constituted by being the sort of thing that is capable of epistemic criticism. That is,

[12] See Moran 1997 and 2002.
[13] See Bilgrami 1998 and 2006.
[14] See Strawson 1974.

self-knowledge involves the very same dispositions as having beliefs or being subject to epistemic criticism. Bilgrami's view represents what I take to be the furthest movement away from a naturalized worldview within western philosophy.

However, there are philosophers with views of self-knowledge that are working with very naturalistic ideas. Naturalistic theories of self-knowledge have been heavily influenced by the work of David Armstrong.[15] Armstrong took cues from John Locke in proposing that self-knowledge resulted from an inner sense. Different theorists have made use of the perceptual metaphor to different degrees. Some of the more literal attempts to view self-knowledge as a perceptual mechanism were sharply, and I think correctly, criticized by Sydney Shoemaker.[16] I should note something of a split in the literature here. While Armstrong, like Locke, was concerned to articulate a global model of introspection, later thinkers have often focused either on knowledge of or access to conscious mental states or focused on knowledge of intentional attitudes like belief and desire. The perceptual metaphor perhaps works best on the consciousness side of this split.[17] However, the idea that one has to do something akin to perception in order to access one's own intentional states has remained influential. Throughout the dissertation, I discuss perceptual models of self-knowledge, which make up the great majority of naturalistically friendly models of self-knowledge.

In the last two decades, there has been increasing scientific interest in our knowledge of the minds of others. It has become increasingly clear that humans have a dedicated system or systems for interpreting the minds of others. These abilities come

[15] See Armstrong 1968, especially chapter15.
[16] See Shoemaker 1994a and 1994b.
[17] See Lycan 1995.

online in clusters and most normal children have developed them by age six. For a time, debate over how these abilities are realized in the brain was dominated by two camps: theory-theory and simulation.[18] The theory-theory camp claimed that we have an innate theory of mind that, once switched on, result in unconscious inferential processes similar to scientific theorizing. Alison Gopnik has been a leading voice for the theory-theorists.[19] The simulation theorists held that we know the minds of others by simulating them. In essence, the simulation theorists held that we know the minds of others by pretending to be in the situations of others or imagining ourselves in their shoes.

Gopnik squared off in a debate with philosopher Alvin Goldman in the early nineties.[20] Goldman is a simulation theorist who has done a large amount of work on self-knowledge.[21] Everyone can concede that it is sometimes helpful to imagine ourselves in the place of others and use our knowledge of what we would do to predict what others would do in the same situation. The main idea behind simulation is that we have an implicit mechanism that does what we sometimes do explicitly. Obviously, simulation theorists are committed to some sort of reliable self-knowledge in order to get the simulation to work properly. Goldman himself is committed to something like an inner-sense view in which we have an internal mechanism sensitive to the phenomenal properties of our mental states.

[18] There are three good anthologies collecting important papers from this debate, Davies and Stone 1995a, Davies and Stone 1995b, and Carrathurs and Smith 1996.
[19] Gopnik has published several books arguing for the theory-theory from the standpoint of developmental psychology. These include, Gopnik and Meltzoff 1997 and Gopnik 1999.
[20] See Gopnik 1993 and Goldman 1993.
[21] See Goldman 1995, 2006 (especially chapter 9), and 2008.

Another important recent view was developed by Steven Stitch and Shaun Nichols.[22] They argue for a fairly simplistic mental mechanism that can unconsciously detect our beliefs and encodes, under the right conditions, a new belief with the content, "I believe that X" where X is the content of the original belief. Like many other theorists in this thread of the literature, Stitch and Nichols are primarily concerned with modeling how we know the minds of others and only model self-knowledge as a stepping stone to that goal. Yet, it is becoming an increasingly important stepping stone because many are moving toward a hybrid account merging elements from Simulation and Theory-Theory.[23] As long as simulation plays any role in knowing the minds of others, self-knowledge will as well. Presumably, a subject knows the result of his or her simulation in roughly the same way that he or she knows his own mind normally.

Self-knowledge has slowly been emerging as a significant field of philosophical inquiry. There are a few important anthologies collecting papers on self-knowledge.[24] Additionally, after the principle material of this dissertation was written, a survey textbook was released, which overlaps with some material in the dissertation, including providing a similar overview of the history of philosophical discussion of self knowledge.[25]

3. The Problem of Self-Knowledge

A useful, though simplistic metaphorical way to understand the field is as a battle between Descartes and Ryle. As noted above, Descartes and Ryle occupy somewhat

[22] See Stitch and Nichols 2003.
[23] This is true of Stich and Nichols 2003, Goldman 2006, Carruthers 1996, among others.
[24] See Cassam 1994, Gertler 2003, Hatzimoysis 2011, which each cover a range of topics. Also see Ludlow and Martin 1998 and Nuccetelli 2003 on the topic of self-knowledge and semantic externalism.
[25] See Gertler 2011.

extreme ends of the spectrum on self-knowledge. On the one hand, we have Cartesian Dualism, complete with infallible and transparent access to all our mental contents. On the other, we have a behaviorist physicalism placing self-knowledge on a par with knowledge of the external world. For Ryle, self-knowledge is had via inference from observed evidence and subject to the same possibilities of error as all other kinds of knowledge. In this dissertation, I advocate a position much closer to Ryle than to Descartes.

However, while it is somewhat informative that I am casting Ryle in the role of hero and Descartes in the role of villain, this metaphor could be misleading in several ways. First, as noted above, it is not clear that it is really Descartes's views about self-knowledge that Ryle and others have attacked. The Descartes evoked here may well be a straw man. However, there are very real philosophers that have seen something of value in the ideas of transparency and infallibility, who use the intuitive appeal of these views as starting points for their own modified Cartesian theses. Descartes is inescapably one of the central figures, and, since I will be arguing against all philosophers on the Cartesian side, he is inevitably something of a villain here.

Holding up Descartes as a villain is probably less misleading than making Ryle out to be a hero. For one thing, though Descartes's mental metaphysics is largely defunct, the metaphysics that Ryle advocated has been as thoroughly rejected. Ryle was a behaviorist, which means, roughly, that he thought that all talk of the mental was descriptive shorthand for observable behaviors and behavioral dispositions. Even though most contemporary philosophers of mind are materialists virtually none would be inclined toward the strict behaviorism of Ryle. For contemporary materialists, the

Cartesian doctrines that Ryle rails against are seen as intuitive and worthy of consideration; to be rejected only on the basis of good argument or empirical evidence. Behaviorism, by contrast, is viewed as a non-starter.

To the extent that Ryle is the hero, it is specifically for his analysis of self-knowledge, though that too is largely out of favor with philosophers. Ryle held, roughly, that we know our own minds in the same way that we know the minds of others. As a behaviorist, he thought that mental talk was short for behaviors and behavioral dispositions. Knowing about minds, whether another's mind or one's own, was a matter of inferring from evidence of past behaviors what future behaviors that minded creature would perform or what behavioral dispositions that creature had. Ryle thought that we know our own minds via inferential mechanisms that took behaviors as evidential inputs and the resulting mental talk could be analyzed in terms of behavioral predictions. He thought that, to the extent that we know our own minds better than we know others, the asymmetry could be explained by the extra evidence we have about our own minds in virtue of constant exposure. I other words, you know yourself better than you know me because you are around yourself more often than you are around me. Just the opposite for me; I know myself and my mind better than anyone else's mind in virtue of constant proximity.

When I began studying self-knowledge, my project was negative. I felt the intuitive pull of the Cartesian doctrine, but, as a naturalistic materialist, was unsympathetic to philosophers that sought to preserve something mysterious or special about mind. In my view, our minds are ultimately the result of physical systems, and our understanding of their workings, even first-personally, must be contiguous with our

knowledge of the external world. However, despite the similarities between my project and Ryle's, the caricature I had in mind of his view was laughable. Ryle was wrong about self-knowledge, or so I thought, because he had an inescapable theoretical commitment to a hopeless metaphysical picture of the mental. As far as I was concerned, the mind is not simply behavior and, even if we use behavioral evidence as the sole evidential input in acquiring knowledge of other minds, we can know things about our own minds without the benefit of such inferences. As this dissertation developed, I found myself softening more and more toward Ryle's view. I now think that much of his view of self-knowledge can be resurrected without bringing back philosophical behaviorism.

In the first few chapters of this dissertation, I pursue the negative project. From certain assumptions about the mind and knowledge, I hope to show in those chapters that our knowledge of our own minds is not especially different from our knowledge of other minds or of the external world. I do this by rejecting accounts that seek to show something special about self-knowledge. It is my hope that these chapters can stand on their own regardless of the positive views I offer later on.

4. Stating Assumptions and Defining Terms

Self-knowledge is knowledge of one's own mind. In broadest terms, this would include knowledge of one's own intentional states, conscious sensations, emotions, and even character traits. However, the philosophical attention on self-knowledge in contemporary philosophy has been much more restricted. Primarily, studying self-knowledge now means studying our knowledge of our own intentional states. As mentioned before, a secondary thread in the literature which could be thought of in terms

of self-knowledge is concerned with our access to our conscious mental states. In general, this literature has moved away from framing discussions in terms of self-knowledge and knowledge of our intentional states is more commonly regarded as the subject of self-knowledge.

In this dissertation, I will have very narrow focus, discussing only our knowledge of our own beliefs.[26] While there is no way to be sure that a general account of self-knowledge covering our knowledge of all our various mental states can be found, it is my hope that at least some of the discussion of our knowledge of our own beliefs will generalize. For the purposes of this dissertation, I will use the term 'self-knowledge' to mean knowledge of one's own beliefs. On rare occasions when I do mean self-knowledge in a broader sense, it should be clear from context.

There are several terms I will make heavy use of throughout the dissertation that require definition. I have already stipulated that, for the purposes of this dissertation, I understand "self-knowledge" to refer to knowledge of one's own beliefs. This compels me to explain how I understand the term "belief". However, belief is too big a target for an introductory chapter. The nature of belief requires separate consideration for its central role in psychology and the philosophy of mind. Throughout this dissertation, I am concerned with trying to pin down exactly what belief is. In chapter six, I will address the nature of belief directly. Until then, unfortunately, I can only state a few assumptions that I make about belief.

[26] Limiting the discussion to belief simplifies things in many ways, but it also occasionally causes confusion. In discussion of knowledge of beliefs, discussion of second-order beliefs, also called meta-beliefs, is unavoidable. Meta-beliefs are beliefs that take other beliefs as content. I, as most people do, take it to be necessary to believe something to know it. Thus, in order to know that I believe P, I must have a belief that I believe P. It can sometimes be difficult to keep straight first or second-order (or higher) beliefs given the nesting structure of meta-beliefs and the way they are represented in sentences. It often helps to switch mental states if one loses track. Making the first order states desires or intentions usually makes it easier to keep track.

I assume belief to be a fixed or standing mental state. That is, belief is inherently dispositional, governing behavior by disposing us to act certain ways in certain conditions, depending in part on the rest of our mental states.[27] I do not exclude the possibility that we may have occurrent beliefs or that we may be conscious of a particular belief, but belief, as I understand it, is essentially a mental state underlying our dispositions to perform various actions and think various things.[28] How beliefs become conscious and the epistemology of our knowledge of conscious mental states are large and important issues that will be sidelined here.

I write about beliefs as if they are things that one either has or lacks. I talk about someone having the belief that P without any mention of believing P to a certain degree or having some amount of confidence in the belief that P. This is not because I want to be committed to thinking of belief as a kind of binary state. There are good reasons for thinking that belief might come in degrees, but I do not think degrees of belief will dissolve any of the problems I consider or generate any problems for the claims I make about belief.[29]

I also generally assume belief to be a propositional attitude. This means that I will often write things such as "the belief that P" where P is meant to be a proposition. I do not take any stand on how one should understand propositions. It is my hope that the great majority of the claims within would survive largely intact no matter what the correct theory of propositional content or even if beliefs had non-propositional contents. I have tried to make as few assumptions about the content of belief as possible. I represent

[27] An important starting point is Audi 1994.
[28] See Schwitzgebel 2010, especially entries 2.1 and 2.2.
[29] See Schwitzgebel 2010, entry 2.4.

belief the way I do for the sake of brevity and consistency in writing more than to take a side in any substantive metaphysical debate. I am assuming belief to be a propositional attitude, but I am considerably more interested in the attitude than I am in exactly what that attitude is directed toward.

This may be especially surprising to those familiar with some of the philosophical literature on self-knowledge. Perhaps the dominant debate about self-knowledge over the last thirty years or so has been about what implications content externalism has for self-knowledge. I take no stand on this debate because I see the issues there as largely orthogonal to the issues I discuss. That is, if there is a good solution to the problems raised by consideration of world switching between twin earths, it ought to work equally well for all the theorists I discuss. If there is no good solution, then the problem is shared by everyone. It seems to me that it is reasonable for the navigators to debate where to steer the ship while they wait for the engineers to declare once and for all whether or not they can keep it afloat.

Since this dissertation is on our knowledge of our own beliefs, I must make some minimal assumptions about knowledge. I will assume that for S to know that P, S must have a justified true belief that P. It is commonly thought, even by those that would agree that justified true beliefs are required for knowledge, that something more must be required to rule out Gettier style cases.[30] I ignore the complications this creates throughout and often write as if justified true belief was sufficient for knowledge. This dissertation contains no discussion of the nature of truth, but the nature of justification will periodically be at issue.

[30] See Gettier 1963.

Since this dissertation is on our knowledge of our own beliefs, I understand self-knowledge to require a person to have a justified true belief that he or she has some belief. Since I take belief to be propositional, believing that one's self believes something is equivalent to believing the proposition expressed by "I believe that P" where P is some proposition. For the sake of brevity, I will often refer to first-order beliefs and meta-beliefs rather than say explicitly what the content of those beliefs is supposed to be. Of course, there are many distinctions to be made among first-order beliefs depending on what proposition is believed. For my purposes, 'first-order belief' will always refer to a belief that P, where P is some proposition that ascribes no mental content. While 'meta-belief' could be used to refer to anyone's belief about any mental state, I will use it to mean exclusively a belief about a first-order belief; a second-order belief is a belief that x believes that P, where x is some person or other. I use 'second-order belief' and 'meta-belief' interchangeably. Unless I state otherwise, the reader should assume that by either term I mean a *self-directed* second-order belief or meta-belief. That is, for the duration of this dissertation, a meta-belief is a belief held by an individual with the content that that very individual has some first-order belief. To readers attuned to the nuances of multi-level beliefs, with nested mental states of various individuals, this way of speaking might be slightly confusing. However, I think that the net gain from simplifying to avoid discussions of belief in propositions such as the one expressed by, "I believe that Bob believes that I want Sally to think that Sarah likes me," is very high and it would add further confusion to an already muddled terminological background if I created new terms to pick out a subset of first-order beliefs or a subset of meta-belief.

I also need to be clear on what I mean by the terms 'naturalism' and 'naturalized epistemology'. These terms have been used to mean a rather wide range of things in the philosophical literature. I am committed to philosophical naturalism in at least three different senses. I endorse naturalism about the mind, a kind of epistemological naturalism, and a kind of methodological naturalism in philosophy generally. By committing myself to naturalism about the mind and to a naturalized epistemology, I mean to have somewhat minimal commitments and would not necessarily endorse many things that others have held under those labels. I view naturalism about the mind to be the view that the mind is in principle knowable by empirical methods. My view of the mind, then, is naturalized because I think that the mind is a real phenomenon in a physical world and that it can be understood in principle by science. I do not here want to deny the possibility of *a priori* knowledge, but I do want to maintain that empirical discovery is potentially relevant to almost any domain and that cognitive science and psychology are deeply relevant to philosophy of mind. I endorse a modest form of methodological naturalism which views scientific methods and results as relevant to philosophy. It is my view that empirical discovery might overturn even the most carefully thought out bit of *a priori* reasoning.

By 'naturalized epistemology', I mean the view that theories of epistemology must show how our actual beliefs and belief forming processes, as described by scientific psychological theories, count as knowledge. This is weaker than the claims many naturalized epistemologists would endorse.[31] I take it as a starting point that we possess some knowledge and I would not accept any skepticism about knowledge or

[31] See Kornblith 1994, pp. 1-14. and Feldman: http://plato.stanford.edu/entries/epistemology-naturalized/

eliminativism about belief motivated by traditional conceptual analysis.[32] Nor would I accept any epistemology or theory of mind that would require beliefs to have properties not possessed by our actual mental states have or to be formed in ways that do not match real psychological processes.

In the interest of full disclosure, I hold certain philosophical commitments, at least at the time of writing this, which I have mostly attempted to suspend for the sake of argument. Namely, I am a functionalist about the mind—or at least about belief—and a process reliabilist about justification.[33] If I make implicit assumptions that depend on those views being true, it is probably because I believe that they are true. However, it is my hope that not much of what I claim depends too much on either of these views.

5. An Outline of the Dissertation

In large part, this dissertation is an extended argument in favor of a naturalistic approach to self-knowledge. I think that arguing for a naturalistic epistemology of self-knowledge goes a long way toward making the case for a broader naturalism. This is because much of the philosophical literature on self-knowledge has remained somewhat anti-naturalistic even as naturalism has gained momentum in philosophy as a whole. While there is a growing body of naturalist philosophers working with and in response to cognitive science and scientific psychology, anti-naturalists have something of a valid complaint in that those in the naturalist camp have done little to show how a naturalized account of self-knowledge can account for phenomena that philosophers have traditionally considered both central to the mental and deeply mysterious. The primary

[32] In Kornblith's terminology, this means I accept ballpark psychologism.
[33] Goldman 1979.

value of this dissertation, I hope, will be that it bridges the gap between the certain traditional philosophical puzzles that have intrigued some analytic philosophers and the discussion of self-knowledge that has emerged from interdisciplinary discussions in philosophy and cognitive science in recent years. In general, I think that naturalists have been as misguided in ignoring these puzzles as anti-naturalists have been in thinking naturalism incapable of addressing them. If I do not solve these puzzles directly, I hope that a dialogue is established and that naturalists gain a foothold in debates they have traditionally ignored.

Chapters two through five, deal with apparent features of self-knowledge which are widely accepted in the philosophical literature that may be thought to challenge a naturalistic understanding of the mind. These features can be summarized with two words: 'authority' and 'directness'. Self-knowledge is typically thought to display a kind of directness distinct from other kinds of knowledge. Having this special form of directness creates the potential for self-knowledge to create a special problem for naturalistic accounts of mind and knowledge. Similarly, the authority that each person has with respect to his or her own mental states can be seen to create similar potential for conflict.

The first purported special feature of self-knowledge that I examine is the directness with which we are said to possess it. By this, it is meant that we do not need to infer what our beliefs are from our knowledge of other facts or anything more basic. Instead, some theorists propose that our knowledge of our own minds is special in that we can know what we believe groundlessly, directly, or non-evidentially. It is not difficult to see how this idea might threaten different theories of epistemology. Theories of

knowledge usually attempt to delineate the epistemic grounds for justified belief. It is something of a puzzle how any theory of knowledge can make room for the idea that there is knowledge without epistemic ground. Naturalism would have added difficulty making room for this idea, or so it would seem, because naturalized epistemologists tend to view epistemic grounds as coming from the formation of our beliefs. If meta-beliefs are groundless, the naturalist is forced to question where they come from; magic is not an acceptable explanation for meta-belief formation for a naturalist. In Chapter Two, I argue that theorists of self-knowledge should not view the apparent directness of self-knowledge as an *a priori* constraint on a theory of self-knowledge. In essence, I argue that self-knowledge seems direct in the same way that other domains of knowledge seem direct even though we do not typically take those other domains to be such.

In Chapter Three, I examine Sydney Shoemaker's views about self-knowledge with an eye toward whether self-knowledge comes out as direct on those views. I argue that it does not. Instead, Shoemaker recognizes classically indirect self-knowledge and, in my view, undermines the motivation for thinking of self-knowledge as direct. I claim that his attempted reductio against an indirect perceptual faculty of self-knowledge works too well. While he took himself to be showing that we clearly have self-knowledge without the kind of faculty described, I take him to show that we do not need self-knowledge for most uniquely human mental tasks. My take on Shoemaker's views is crucial to developing my own view of self-knowledge.

Another purported feature of self-knowledge is that we know our own minds *authoritatively*. In Chapter Four, I distinguish between epistemic and non-epistemic types of authority. I there argue that our self-knowledge is not possessed with any type of

epistemic authority which challenges naturalism. I show that the naturalist is in as good a position as anyone to explain the epistemic authority of self-knowledge, and, if there is a special authority for self-knowledge, it must be compatible with the idea that there could be a better way to know what we believe than we currently have.

Chapter Five continues the examination of authority by considering whether there are any non-epistemic kinds of authority possessed by self-knowledge which threaten naturalism. I pick up on a theme of Richard Moran's writing on self-knowledge that it represents authority via authorship. That is, self-knowledge is possessed with a special non-epistemic authority due to the fact that we create or control the mental states that our self-knowledge ranges over. Typically, naturalistic models are models in which we detect existing beliefs. I argue that naturalistic detection models are compatible with a variety of theories of doxastic control. However, detection models are compatible with doxastic control largely because the formation of a belief is irrelevant to its detection. There are possibilities to explore for alternative naturalistic models of self-knowledge that provide a central role for doxastic control. I sketch what such models might look like. To evaluate the impact that doxastic control has on our self-knowledge, we first need to know how much control we have over our own beliefs. I provide some reasons for thinking that our control over our own beliefs may be more limited than some thinkers assume. This is an open issue that is at least partially empirical. An important claim in the chapter is that the nature and amount of control we have should be independently decided and should constrain our theorizing about self-knowledge.

In Chapter Six, I seek to get to the heart of the diverging views about the nature of self-knowledge. I contend that different approaches to self-knowledge are encouraged by

different conceptions of belief. Some philosophers tend to emphasize verbal and ascriptive behaviors when discussing belief. Others tend to focus on things that might be thought of as lower-level such as non-verbal behaviors and representative capabilities. Most assume that these behaviors all cluster together and that all are important parts the role of belief in our mental lives. However, in studying self-knowledge, one is forced to look at situations in which people seemingly do not know what they believe. These situations show patterns of behavior that show a division in the commonsensical notion of 'belief', forcing a theoretical decision about what we consider to be the nature of belief. What counts as a good theory of self-knowledge will depend vitally on what decision is made. I consider the options and argue in favor of one.

Finally, I offer the outline of a positive model of self-knowledge. In order to do this, I turn my attention to meta-belief. To study self-knowledge, we need to have a full enough understanding of the conditions under which we succeed and fail to know our own beliefs. This involves understanding the role that our self-directed meta-beliefs play in our mental lives. My preliminary suggestion is that meta-belief is more connected to conscious reasoning processes than is first-order belief, suggesting a framework for future empirical and theoretical work to test this idea.

Chapter 2

Direct Self-Knowledge

1. Apparent Directness

Perhaps the most puzzling feature attributed to self-knowledge is its epistemological directness. Claims that self-knowledge is direct typically revolve around a discussion of how one might attempt to support an apparent statement of self-knowledge. While other facts that we know appear to be supported by reasons or believed on some basis or other, facts about our own minds seem to be known simply and directly. What, for example, might you say to someone who asked you how you know *that* you believe that Barack Obama is the President of the United States? This is importantly different from being asked *why* you believe Barack Obama is president—that question is easy to answer. Perhaps you saw a news report on the president's most recent foreign tour. You may remember the moment in which he was elected or the day he was sworn in. There are plenty of reasons why you might believe that Obama is president. What reason, however, would you offer for your believing that you believe that he is president?

There are true propositions you could cite to support the idea that you have such a belief. They would be much the same as ones you would cite to support a claim that another individual had such a belief. If you have a politically savvy friend, for instance, you could support your belief that your friend believes Obama is president by pointing out that Obama *is* president and that your friend reads the paper. One of the odd things about self-knowledge is that although you could say these same things about yourself—that you read the paper and that it routinely mentions the president by name—it seems patently unnecessary and perhaps disingenuous to offer similar support for believing that

you believe Obama is president. In offering those reasons to justify your claim that your friend has that belief, it can feel as though those may really be the reasons you ascribe a belief to your friend. However, *your* belief that you believe Obama is president does not subjectively seem to be based on knowing that you read the paper. It doesn't seem to be based on anything at all. Subjectively speaking, it seems that you know what you believe not by observation or inference—you just know it.

I will treat this case as a paradigm instance of self-knowledge. That is, I assume that before reading this chapter you both believed that Obama is president and believed that you believed Obama is president.[1] This means that I accept some form of dispositional view of belief. While this is not an uncontroversial assumption, I cannot fully defend this assumption here. Also, I will refer to beliefs such as your belief that you believe Obama is president as 'meta-beliefs'. Unless otherwise specified, when I use the term 'meta-belief', I mean a self-directed belief that one believes P, where P is some proposition. I do not view them as metaphysically different kinds of states. The difference is one of content; meta-beliefs are beliefs about one's own beliefs. In the case described above, the meta-belief is your belief that you believe Obama is president.

We take ourselves to have lot of mundane self-knowledge similar to the paradigm case described above. Yet, when pressed, we seemingly cannot defend it in any way more sophisticated than, "I just know". In other contexts, "I just know" can be considered the worst kind of support. It doesn't have the humility of "I'm not sure" nor does it hint at a better answer in the way that, "I must have read it somewhere" might. It cuts off discussion without offering any support at all. In domains other than self-knowledge,

[1] Perhaps the meta-belief was formed in response to the questioning. I do not think this would undermine my point. The idea is to have an example of a bit of self-knowledge that comes very easily, without any explicit inference or observation of behavior.

such a response might cause us to become skeptical about whether the person in question really did know. Unwillingness or inability to articulate a reason for your belief suggests shallow and dogmatic thinking. Yet, in the domain of self-knowledge, this response seems natural and does not prompt the same skeptical response. If anything, we regard those facets of your self-knowledge that you "just know" as especially secure. So self-knowledge appears to be genuine knowledge, yet we cannot say why or how our meta-beliefs are justified and, moreover, do not typically feel that we need to support our self-knowledge claims at all.

For Descartes, self-knowledge was more than especially secure. He thought that one could not fail to know what was going on in one's own mind. Descartes held what Shoemaker and others have called a transparency thesis: the view "that nothing can occur in a mind of which that mind is not conscious."[2] In his view, our knowledge of the contents of our minds is so secure that it can be used as the foundation upon which to build up our knowledge of everything else. If we were infallible about our own minds, it would be unsurprising that we feel no need to support our claims to self-knowledge. Perhaps it would remain an interesting question why nothing we say seems to capture our reasons for holding meta-beliefs, but any puzzle over whether we need to offer such support vanishes with infallibility. No one would bother asking God to support his beliefs.

However, the Cartesian notion of transparency is largely out of favor these days. Most concede that there are mental states and processes that we never become aware of and that introspection is sometimes wrong about what states we have and how mental

[2] See Shoemaker 1996, p. 50.

processes work. Once we admit the possibility of error in our self-ascriptions, it is unclear how much of the Cartesian picture should remain. Many philosophers continue to think that self-knowledge is direct—an idea somewhat Cartesian in spirit—even while admitting that we are fallible.

The recent return to the idea that self-knowledge is directly justified came partially in response to Gilbert Ryle's account of self-knowledge.[3] Ryle held that we know our own minds the same way we know the minds of others, by inference from observational evidence. The denial of Ryle's account often amounts to a statement of directness. Consider, for instance, Donald Davidson:

> Ryle was wrong. It is seldom the case that I need or appeal to evidence or observation in order to find out what I believe; normally I know what I think before I speak or act. Even when I have evidence, I seldom make use of it.[4]

Davidson is clearly in favor of some understanding of directness. However, his claim here conflates a variety of ideas that might amount to the claim that self-knowledge is direct. In saying we do not *need* evidence or observation, Davidson implies that we can *form* meta-beliefs directly. In saying that we do not *appeal* to evidence or observation, Davidson implies that our self-knowledge claims can be justified without *citing* observation or evidence. These are distinct notions that are each present in the self-knowledge literature—often conflated as Davidson seems to do here. Sydney Shoemaker also attempts to weaken the Cartesian theory of mind by holding a slightly different thesis. He accepts a version of a self-intimation thesis: "that it belongs to [intentional states'] very nature that having them leads to the belief, and knowledge, that one has

[3] See Ryle 1949.
[4] See Davidson 1987, p. 43.

them, or at any rate that it normally does so under certain circumstances."[5] However, this view that belief is self-intimating is consistent with meta-belief being indirect. Having beliefs might somehow provide good grounds to infer that beliefs exist.[6] The claim that self-knowledge is direct is distinct from and in certain ways stronger than the claim that belief is self-intimating.

In asserting that self-knowledge is formed directly, philosophers often cite phenomenology. For instance, Dorit Bar-On writes:

> Now consider the situation from the point of view of the person issuing the avowal… in the normal case, I do not consult my own behavior or memory; I do not reason, or draw some inference… for all appearances, my avowal is baseless.[7]

For Bar-On, an avowal is an utterance which ascribes a current mental state to the speaker. The apparent baselessness of this avowal—the fact that there is no subjectively conscious process leading to it—has been taken by some to show that our meta-beliefs are not formed in the same way as other beliefs. In contrasting self-knowledge and external world knowledge, Akeel Bilgrami puts the claim this way:

> Knowledge of the external world is in ordinary cases the result of such *cognitive achievement* (as philosophers call it), which if the agent does not do or have happen to her, then she will not form the belief that there is some fact or object in the world. In self-knowledge there is, in the ordinary case, no such cognitive activity nor any analogue to it, which is required to believe that one has some mental state… to put it as I have in various papers over the years, it comes, as it were, 'for free', without any cognition.[8]

But surely phenomenology is not enough to conclude that self-knowledge consists of meta-beliefs that do not result from cognition. I think there are two factors supporting the idea that self-knowledge is direct. The first and more widely recognized is the

---

[5] Shoemaker 1996, p. 51.

[6] Shoemaker's view denies any sort of inferential or evidential connection between belief and meta-belief; I merely want to point out the differences between self-intimation and directness and that someone could hold one without the other.

[7] Bar-On 2004, p. 2.

[8] Bilgrami 2004, p. 31.

phenomenology: that it seems direct. Some philosophers seem willing to theorize in line with this phenomenology and to grant that all apparently direct knowledge is genuinely direct. For instance, Richard Moran writes, "The claim that introspective awareness is not inferred from observational evidence is what is usually intended by the claim that it is 'immediate'." He continues to clarify his notion of immediacy, saying it "is to be understood as a wholly negative claim about the mode of first-person access, that is, awareness that is not inferred from anything more basic." However, he goes on to make the surprising claim that, "much of our ordinary perceptual awareness is also taken to be immediate in this sense…."[9] This claim is surprising because, in articulating a notion of immediacy that places self-knowledge on a par with perceptual knowledge of the external world, Moran seems to undermine the claims of many that agree with him that self-knowledge is direct.[10] For instance, John McDowell would presumably reply to Moran the way that he replied to Crispin Wright here:

> And the authority of observations is indeed *non-inferential*. But it is precisely *not* baseless. The question 'How can you tell?' is precisely *not* excluded as inappropriate… he accepts it and answers it when he says, perfectly properly, that he is in a position to see.[11]

Here McDowell is careful to make sure his support of the idea that self-knowledge is baseless is stronger than the claim that it is not inferred consciously. McDowell is right to point out that there is a sense in which perception and perceptual knowledge can be called direct because perception does not involve conscious inference. If this is all there is to the directness of self-knowledge, then there is no need to go to extraordinary lengths to come up with an account of it. Those that endorse perceptual models of self-

[9] Moran 2001, p. 11

[10] Moran, of course, has other problems with perceptual models of self-knowledge; he objects that they do not make self-knowledge inherently first-personal.

[11] McDowell 1998, p. 48

knowledge can claim with good reason that self-knowledge on their accounts can be considered at least as direct as knowledge acquired through other modes of perception.

But perceptual beliefs are taken to be indirect or grounded in another sense because the things we say to support them seem correct or reasonable. This is true for many beliefs that are not formed through conscious inference. For instance, if you read in a chemistry textbook that oxygen atoms have eight protons, you will very readily form the belief that oxygen atoms have eight protons. We all form such beliefs without explicitly or consciously reasoning that that book is asserting some fact and likely to be reliable. We can scarcely be said to be aware of the belief forming at all. Yet, if asked, we tend to produce as justificatory reasons for these beliefs facts such as that the book said there were eight and that it is probably correct. If this is taken to be a paradigm indirectly justified belief, then we cannot tell the difference between directly and indirectly justified beliefs simply by the way it subjectively feels to form them. Though meta-beliefs also typically form unconsciously, we regard offered justifications differently. Meta-beliefs seem justified in spite of the fact that any offered justification seems somehow inappropriate.

The point that it seems wrong to offer reasons supporting apparent claims to self-knowledge is often raised in connection to the claim that self-knowledge is held with special authority. First-person authority can be thought of in terms of it being inappropriate to *ask for* support for apparent claims to self-knowledge. Here, for instance, is Crispin Wright:

> The demand that somebody produce reasons or corroborating evidence for such a claim about themselves… is always inappropriate. There is nothing

> they might reasonably be expected to say. In that sense, there is nothing upon which such claims are based.[12]

Similar points are raised by McDowell and Bar-On and are suggested in claims like Davidson's above, that self-knowledge is direct because we do not need to appeal to our evidence.[13] It is unclear how the apparent inappropriateness of challenging first-person authority might bear on the perceived inappropriateness of offering support. However, it is certainly underappreciated how much the idea that self-knowledge is direct depends on this intuition that we cannot reasonably say anything to support our self-knowledge claims. As I pointed out above, the phenomenology, or lack thereof, of meta-belief formation does not separate self-knowledge from many other kinds of knowledge. What separates self-knowledge from other kinds of knowledge at the phenomenological level is that our offered justifications of our other beliefs seem appropriate or correct, whereas none seems appropriate or correct for claims of self-knowledge.

The subject of this chapter is the apparent directness of self-knowledge and the intuitions that motivate theorists to constrain their theories of self-knowledge. Let me reiterate that I am considering in this paper the idea that directness is a feature of self-*knowledge*. I will not, for instance, be considering here the possibility that apparent claims to self-knowledge appear direct because they are part of a language game or feature an unusual semantics while not representing real knowledge.[14] A story at the level of semantics does not pin directness on self-knowledge and is not obviously committed to apparent claims to self-knowledge representing genuine assertions of meta-beliefs.

---

[12] Wright 1998, p. 14. The quotation actually refers directly to phenomenal avowals, but he claims that attitudinal avowals have the very same groundlessness.

[13] See McDowell 1998, pp. 47-50. and Bar-On 2004, p. 3.

[14] See, for instance, Wright 1998, pp. 41-45 in which he attributes what he calls the 'Default View' to Wittgenstein. It is clear here that the default view represents the special features of avowals as parts of grammar and not representative of the epistemic relations between a person and his or her beliefs.

Though they are hard to pin down, some of the philosophers I discuss at times seem to endorse such a story.[15] I take it that most any epistemology of self-knowledge is likely to be compatible with most any account of the semantics of apparent self-knowledge claims. Because of this, whether and how we know what we believe can be studied independently from the way we talk about what we believe, even when our talk is suggestive.

In this paper, I will argue that we should not regard the apparent directness of self-knowledge as indicating deep truths that must be accounted for in a theory of self-knowledge. Some philosophers, including many mentioned above, seem to hold that we can know *a priori* that self-knowledge is direct and that any theory of self-knowledge incompatible with directness must be false. I argue that the most that is required of a theory of self-knowledge is that it explains the *appearance* of directness, although even explaining the appearance might be viewed as a separate task.[16] Once possible notions of directness are on the table, it becomes an empirical question whether self-knowledge is direct. A strong secondary goal for this paper is to begin laying the ground work for an empirical investigation.

---

[15] The project of Dorit Bar-On's book, *Speaking My Mind* (Bar-On 2004), is just this, to give a semantics level explanation of the special features of avowals that is separate from, though she hopes amenable to, any privileged epistemic access.

[16] If, for instance, the inappropriateness of offered support is not based on epistemic concerns, it is hard to see how it can serve as the basis for a constraint on a theory of the epistemology of self-knowledge. It may be that we may feel constrained by non-epistemic norms to avoid challenging each other about self-knowledge. If that were right, the story may be extended to say that it seems wrong to offer justifications for self-knowledge at least in part because of those same non-epistemic norms. If there is some implicit restriction against asking me to defend my self-knowledge, there won't be many circumstances in which it seems reasonable to defend myself. It often seems unreasonable to offer detailed defenses to unreasonable charges. Why should I, for instance, defend myself against a charge that I assassinated Lincoln in any way other than by saying, "I just didn't"? Offering a real defense, to some extent, dignifies the charge. If we have non-epistemic reasons to avoid challenging apparent claims to self-knowledge, then the need to offer good epistemic justifications for self-knowledge claims may appear lower on balance than it should be from a purely epistemic point of view. Though I like this story as an explanation of the intuitions cited by Wright and others, I will not defend it here. I believe that the attempt to use these intuitions to constrain theories of self-knowledge can be shown to be misguided without any story about their source.

2. Directness as a Constraint

In presenting a theory of self-knowledge, everyone can and should agree to the *apparent* directness of self-knowledge. We all take ourselves to know what we believe very readily even though we very rarely consciously reason about what we believe. It just seems to come to us. No active justification of this knowledge seems necessary or even appropriate. I wish to discuss whether we should take the appearance to be genuine and allow a commitment to directness to constrain our theorizing about self-knowledge. Let SKID stand for the thesis that self-knowledge is direct:

> Self-Knowledge Is Direct (SKID): In normal cases of self-knowledge, self-knowledge consists in the self-knower having a directly justified meta-belief.

The issue is not merely whether self-knowledge is direct, but whether it can be known to be direct *a priori* or prior to investigation into various psychological processes. I do not think epistemologists would be well served to view SKID as a constraint on theory. As far as I can tell, SKID is a substantive thesis in need of defense. Though I am somewhat suspicious of all claims to *a priori* knowledge, I will not rely on a general rejection of the *a priori* here. I think that the specific intuitions that encourage pre-theoretic acceptance of SKID are mistaken and can be shown to be so independently of a more general critique of the *a priori*.

My first task is to formulate a notion of directness that remains neutral between the different ideas of directness in the literature. Minimally, the claim that self-knowledge is direct is that meta-beliefs are justified in the absence of epistemic grounds. I believe that the confusion over the issue of directness is partly the result of a diversity of conceptions of justification and of epistemic grounds. Those notions will be explored

later in the chapter. The overarching idea of directness can be understood in general terms as follows:

> Directness: A person S is directly justified in believing P if and only if S has a justified belief that P and S has no epistemic ground for S's belief that P.

Understood in terms of the notion of directness above, SKID claims that our meta-beliefs are typically justified without our having epistemic grounds for them. As I have defined it, directness is neutral about two key notions: justification and epistemic grounds. It represents directness as a property of justified belief but exactly what the property it is depends on these notions. Different theories of justification and different accounts of epistemic grounds will yield different notions of directness. In order to be fair to SKID, we should take care to consider some different options. It is worth reiterating at the outset that no plausible theory of self-knowledge would require a conscious inference be made to form a meta-belief. In this sense, there is a notion of directness that can be had on the cheap but its compatibility with all the alternative views makes it thoroughly uninteresting. I have in mind something like this:

> Conscious Conception of Directness (CCD): S's belief that P is directly justified if and only if it is justified and S's belief that P was not formed on the basis of conscious inference.

It would be rather uninteresting to accept SKID in light of CCD because CCD makes out beliefs that we clearly consider indirect to be directly justified. Consider, for instance, something like a flash of insight. Suppose you know full well that Bob only brings his umbrella to work when rain is forecasted. If you see Bob carrying his umbrella or a friend causally mentions that Bob has his umbrella, you might instantly, with no conscious inferential thought process, come to believe that it will rain today. But surely, this belief depends on something like an inferential connection to other beliefs of yours for its formation and, ultimately, its justification. The fact that the inferential process was not

carried out consciously means very little in terms of the nature of the justification of the belief.

CCD might make SKID out to be a constraint on theorizing about self-knowledge, but it would do very little constraining of those theories. In particular, it would not constrain any theory that might seek to explain the justification of self-knowledge in the kinds of cases I am concerned with. My description of the case of knowing that you believe Obama is president was meant to be a case in which your meta-belief was seemingly justified before you evaluated it and seemingly remains justified despite your inability to say anything that seems reasonable in defense of it. In light of this, I will not be considering theories of justification in which beliefs are only justified if the believer undergoes a self-conscious evaluation of the belief. A theory of justification such that justification only attaches to actively considered beliefs might warrant some consideration, but the notion of directness there is likely to be somewhat different.[17]

Of course, in treating SKID as a constraint on theories of self-knowledge, most philosophers are careful to point out that we sometimes do form meta-beliefs in a more indirect or even consciously inferential way.[18] Those meta-beliefs might be justified on the basis of certain facts and we could reasonably be expected to cite those facts as justification. Such indirectly justified meta-beliefs stand in contrast to our normal meta-beliefs. The process of self-discovery undergone in therapy is taken as a model of indirectly justified meta-belief. Compare coming to believe that you hate your job on the

[17] Some have held that a certain class of meta-beliefs is self-verifying in something like the way that the *cogito* is often thought to be. For instance, see Burge 1988. However, I am unconvinced that even occurrent self-ascriptive thoughts are self-verifying, and even if there were some possibility of showing that occurrent meta-beliefs are self-verifying, it would not seem promising to extend this idea to cover unconsciously formed meta-beliefs.

[18] See Bilgrami 2006 p. 8., Moran 2001, pp. 10-11 and p. 85, and Wright 1998 pp. 15-16 for example.

basis of an hour long therapy session to just knowing, without any therapy, that you hate your job. The indirect way of knowing is thought to be more the exception, while the rule is that self-knowledge comes directly. The project of theorizing about self-knowledge is viewed as the project of explaining the normal or standard cases and SKID is meant to apply only to those cases.

Still, it is not clear what exactly directness is supposed to amount to. The idea is that nothing important stands between the believer and the belief. However, different epistemologists may have different ideas about which sorts of things that can stand between a believer and his or her beliefs are epistemologically important. For many epistemologists, directness may be a matter of not having the believer's other beliefs mediating. Clearly, if some belief B_2 plays a role in the justification of a belief B_1, then B_1's justification can be thought of as indirect, as something it has in virtue of its relationship to B_2. This idea can be expressed as follows:

> Belief Conception of Directness (BCD): S's belief that P is directly justified if and only if it is justified and the belief that P does not depend on any other justified belief of S for its justification.

While BCD focuses on S's beliefs, we might still wonder if there are any other mental states that can mediate between a believer and his or her beliefs in a way that is important for epistemology. The most likely way for another mental state to play an epistemically important role is for it to feature in the formation of a belief. While non-belief mental states may not be able to stand in the sorts of logical relations to beliefs that other beliefs can, they can certainly play a role in the formation of beliefs. Sensations, hallucinations, pains, and desires are all mental states that can cause us, or so some have claimed, to form beliefs. If some of these states can cause beliefs that are justified, we may think of

them as standing between the believer and his or her beliefs in an important way. This could lead to a separate notion of directness:

> Formation Conception of Directness (FCD): S's belief that P is directly justified if and only if it is justified and no mental state of S is involved in the formation of S's belief that P.

However, this will not quite do. While some theories of justification hold that the formation of belief is central, others will not attribute much significance to the way in which a belief is formed. Even those philosophers who do think belief formation can be crucial to justification may not view every bit of a belief forming process as epistemically relevant. In light of these concerns, I will consider MSCD instead of FCD:

> Mental State Conception of Directness (MSCD): S's belief that P is directly justified if and only if it is justified and no mental state of S is involved in an epistemically relevant way in S's belief that P being justified.

MSCD is stronger than BCD. Any belief that is direct on MSCD will also be direct on BCD since beliefs are mental states. The different notions of directness represent a disagreement among epistemologists over what sort of states can count as epistemically relevant mediators for a person's beliefs. Given the prevalence of discussing directly justified beliefs as groundless, I think we can usefully label these sorts of states as epistemic grounds. To be clear, my understanding of an epistemic ground is that they are mental states that play a mediating role in the creation or conferral of justification on a belief. BCD admits only a person's other beliefs as potential grounds while MSCD might admit other kinds of states as well. Some may want to cast the net even wider and let non-mental states of affairs or external objects count as grounds. While I can think of

some motivation for doing this—even in the context of a discussion of self-knowledge—I do not believe that considering such an option will be helpful here.[19]

The final piece of the puzzle for considering whether meta-beliefs are directly justified is to consider different theories of justification. I will consider three possibilities. These are not meant to be completely exhaustive, but are meant to carve up the majority of plausible theoretical space left by my starting assumptions about self-knowledge. Thus, I am considering only those views of justification which comport with a view of belief as a dispositional state that can be justified even when never actively evaluated.

Consider beliefs that non-skeptics will take to be justified, such as the beliefs one forms about chemistry while reading a chemistry textbook. There are two ways we can understand how offering support shows these beliefs to be justified. We may think that we are reconstructing the unconscious process of belief formation (or some part of it) in citing these facts. This idea is that the very facts offered in support of your belief mirror the way your belief was formed, and, as long as they constitute a good way to form a belief, your knowing how it was formed justifies the belief. Call this the *reconstructivist* view of justification. The other way to take offered supporting facts as justifiers is to think that, regardless of whether or not the process underlying the belief formation mirrors an inference from these facts, your ability to cite these facts on demand shows that you are clearly capable of performing this inference. On this view, being able to offer a good argument when called upon entitles you to the belief or, in other words, justifies your holding it. Call this the *discursivist* view of justification.

[19] My use of the term 'ground' is much in line with Alston 1988. Interestingly, though he remarks that he thinks an extremely wide conception of grounds—one that would admit external things like facts—is natural, he ultimately restricts grounds to mental states as well.

Both discursive and reconstructive theories of justification require justifications to be available to the believer. In light of this, both theories should count as varieties of internalism. There are genuine differences between reconstructive and discursive theories of justification, but we must also consider an externalist theory of justification. I will argue that on any of these three theories of justification, SKID is unmotivated as an *a priori* constraint on theorizing about self-knowledge.

3. Justification

In this section, I will consider the motivations for SKID under three different views of justification. Theories of justification are typically divided into externalist and internalist theories. I further subdivide internalist theories into reconstructive and discursive theories. I argue that SKID is not motivated as an a priori constraint in a theory of self-knowledge under any of these theories of justification.

3.1. Externalist Justification

The bulk of the argument will be over whether SKID can be considered an *a priori* constraint for reconstructivists or discursivists. Perhaps unsurprisingly, from the standpoint of the externalist, SKID seems something of a non-starter. In light of this, I will not elaborate the externalist conception of justification in much detail. If the reader would like to have a specific theory in mind for the externalist, consider something like Alvin Goldman's classic theory of process reliabilism: S justifiably believes that P if and

only if S's belief that P is the result of a reliable cognitive process.[20] I will discuss different possibilities for discursivist and reconstructivist theories in more detail.

The idea that SKID could be an *a priori* constraint on a theory of self-knowledge is, as I have said, a virtual non-starter with an externalist theory of justification. The reason is that the externalist views justification as something that a belief can have without the believer having access to various facts about how or why it is justified. SKID, considered as a constraint on a theory of self-knowledge, states that we can have justified meta-beliefs with no epistemic ground. The externalist can easily concede that we have justified meta-beliefs even though we have no idea what the grounds for them may be. To take on the further claim that these meta-beliefs are groundless or direct is to commit to something stronger. On what might the externalist base such a commitment?

The internalist, whether discursivist or reconstructivist, has some apparent motivation for accepting SKID as a constraint. If one trusts the intuition that meta-beliefs are justified, and one holds that we must typically have access to the justifiers of justified beliefs, then the fact that we do not apparently have access to justifiers in the case of self-knowledge suggest that these beliefs are justified though groundless. However, for the externalist, there is no requirement that one have access to the grounds of a belief for it to count as justified. Trusting the intuition that our meta-beliefs are justified—which some externalists may be independently reluctant to do—provides no reason to posit anything about the grounds of the belief. Since an internalist requires that we have access to justifiers, the internalist is required to trust intuitions not just about what beliefs are justified but must also trust intuitions about what actually justifies them. There is no need

[20] See Goldman 1979.

for the externalists to trust that second set of intuitions. The externalist could point out that for all we know there are a variety of beliefs at play in the formation of meta-beliefs that would count as epistemic grounds for the externalist. These intuitions will be addressed throughout the next several sections and I will argue that no one should rely on them for theorizing. It is worth noting here that it would be especially strange for an externalist to rely on them, but if one were tempted, the concerns raised in the next section will likely apply.

While it may turn out that SKID is true according to the externalist, we will have to make sure that our other beliefs do not unduly influence the formation of our meta-beliefs. It is somewhat harder to imagine that externalist would ever accept SKID in terms of MSCD. That would require that there are no mental inputs to the processes of meta-belief formation worthy of being called epistemic grounds. This is a strange possibility to consider, but it depends in part on a correct characterization of meta-belief forming processes to evaluate it.

Allow me to make one more note before moving on. The externalist can hold—perfectly consistently—that we are justified in our meta-beliefs and justified in thinking that our meta-beliefs are justified without trusting the intuitions that they are. Remember, if justification can be external, for instance if it is just about forming beliefs via a reliable process (and not about having access to the process), then I can form justified beliefs about the justificatory status of my beliefs via any old reliable process.[21] It might be nice, for all sorts of reasons, to do some empirical work to delineate the structure of our meta-

[21] Thanks to Hilary Kornblith, for pointing this out and for consistently noticing when I think like an internalist.

belief forming processes, but it is not necessary to do this work prior to believing that our meta-beliefs are justified.

3.2. Reconstructive Justification

Though the externalist and internalist may disagree about what should count as justification for a belief, the central disagreement can be seen as a disagreement about access to justifiers. The disagreement between reconstructivist and discursivist theorists of justification, by contrast, is about how to think of justifications. It is open to the internalist to think that justifications, which justified believers must have access to, must be involved in the formation of a belief or that mere access to them entitles one to hold the belief. Internalists are often vague about the nature of the relationship between a justified belief and the grounds of the belief. The reconstructivist position that I am outlining here requires this relation to be causal. Many externalists will be in agreement that there must be a causal relationship between ground and belief. The reconstructivist, as opposed to the externalist, is committed to the accessibility of these grounds.[22] The discursivist gives up this causal requirement in favor of a different kind of relation. I suspect that most internalists are reconstructivists, but it is not always easy to tell how epistemologists conceive of the relationship between grounds and belief.

Consider the case of a belief that is formed via a process of valid conscious inference. Such a belief will come out as justified on both reconstructivist and discursivist accounts of justification. The actual process by which such a belief is formed should

[22] Robert Audi is especially clear in articulating the options for such a view. He calls the view, "causalist internalism," and defends it in Audi 1993, pp. 332-352. William Alston has defended a reconstructivist view as well. See Alston 1988. He is careful to distinguish between the accessibility of the grounds (which he requires for justification) and the accessibility of the adequacy of the grounds (which he does not require), but all the worries I have will apply to him as well.

count as justifying for the reconstructivist, provided the details of that process are accessible to the believer. Furthermore, since the actual reasoning will presumably remain subjectively available as support, this will count as justified by the lights of the discursivist as well; in other words, you could offer that reasoning in a discussion. The interesting differences between the theories come out in response to beliefs formed via unconscious processes.

A reconstructivist would require that we have access to the way the belief was formed in order for it to count as justified. To avoid skepticism, a reconstructivist trusts intuitions about which of our beliefs are justified and presumably trusts that the things we offer as justifications for our beliefs are really things that led to the formation of the beliefs. Since a reconstructivist must concede that we are often unaware of the formation of our beliefs, the reconstructivist trusts that we can subjectively reconstruct how our beliefs are formed. If we could not, then beliefs formed via unconscious processes would be unjustified beliefs.

However, in the case of self-knowledge, the reconstructivist seems to be in the odd position of saying that meta-beliefs are justified despite not being able to subjectively reconstruct the process which forms them. I am granting, for the moment, along with those that think meta-belief is direct or groundless that there is nothing one can say that seems to be appropriate justification for a meta-belief. The reconstructivist makes justification contingent on having access to the reasons the belief was formed. Since we have no access to anything that seems like a reason the meta-belief was formed, the reconstructivist is pressured to say that meta-beliefs are unjustified. However, as I have previously discussed, meta-beliefs are typically taken to be among the most

authoritatively held beliefs. It would seem preposterous for any internalist to think that meta-beliefs are unjustified.

So reconstructivists are immediately faced with a dilemma. They may either accept that meta-beliefs are justified, which is seemingly at odds with their theory of justification, or think that they are not, which is at odds with their strongly held intuitions. Since I am interested in self-knowledge, it would do little good here to consider the possibility that all meta-beliefs are unjustified. Instead, we should consider how the reconstructivist would make an exception for the justification of meta-beliefs. One possibility, that meta-beliefs are justified despite a lack of access to the grounds on which they are formed, seems especially *ad hoc* for the reconstructivist. Better to trust all the intuitions at play and think that the meta-beliefs are both groundless and justified. Additionally, this creates the exception to the general reconstructivist picture in a potentially helpful way. That is, one might think for independent reasons that some beliefs must be justified without being grounded. If meta-beliefs are taken to be among them, then they are a species of foundational belief. I take no stand here on whether or not there are foundational beliefs or whether foundationalism is the correct theory of the structure of justification. Instead, I want to focus on the intuitions motivating the reconstructivist to include meta-belief in the foundations.

Before doing that, I should head off a worry about focusing on the processes that cause our beliefs. After all, our beliefs often depend on various maintenance processes to continue existing in our minds. The reconstructivist might place equal or even more emphasis on belief maintenance than on belief formation. However, to the extent that these belief maintenance processes are unconscious, I believe the worries raised below

apply equally well. If, however, one thinks of belief maintenance as conscious reflection and deliberation about one's beliefs, then one has gotten away from the sorts of cases I have taken to be paradigms of self-knowledge. Most would be willing to concede that SKID is true on CCD, even in cases of conscious deliberation.

As I have outlined above, one way to reach the conclusion that self-knowledge is direct is to trust that we could reconstruct the unconscious processes that form and maintain meta-beliefs if there were any such processes. Since we seem unable to reconstruct these processes in the cases of ordinary self-knowledge, SKID would have to be true. This constitutes an *a priori* argument for SKID that seems to me to be implicitly behind some assertions that self-knowledge is direct.

Unfortunately, the picture emerging from the empirical literature on unconscious mental processes does not look favorable for this argument. The problem is that our subjective access to cognitive processes seems to be poor, and the trend in the research has been to highlight more and more ways in which it is poor. If I am right, two things follow. First, it follows that reconstructivists face the threat of a very broad skepticism. Second, it follows that the motivation for accepting SKID is undermined. I will now run through a few examples in the empirical literature to attempt to show how poor our subjective access to cognitive processes can be.[23]

People are more likely to find a potential partner attractive if they are antecedently aroused due to other stimuli, which need not be sexual stimuli. One study used a bridge that felt dangerous to produce the effect.[24] While we are on the subject of attraction, there is a stimulus that studies have shown can have strong effects on

[23] A more thorough review can be found in Wilson and Dunn 2004.
[24] Wilson 2002, pp. 100-102.

interpersonal interactions which we are almost never consciously aware: the dilation of our pupils.[25] Our pupils dilate to let in more light when we look at something pleasing. We use the dilation of the pupils of others as an unconscious cue to make various judgments without realizing that we do. When photographs are manipulated to appear as if the photographed individual's pupils were more dilated, subjects judge the photographed people to be happier and more attractive than in otherwise identical photographs. However, no one listed the dilation of the pupils as relevant or could recall noticing the pupils at all.

Though no one ever feels as if he or she is making a selection based on the spatial positions of the choices, repeated studies have shown that there is a position effect on selection. In one classic study, psychology researchers posed as marketing researchers and asked subjects to choose the best among identical pairs of nylon stockings.[26] People readily made a choice and readily backed up their choices by appeal to differences that the stockings did not actually have. The most significant factor controlling their choice was the placement of the stockings on the table in front of them. The further right, the more likely a pair was to be chosen as best. This rightmost preference seems to be very real and yet no one ever feels, even after learning that it exists, as if it is influencing their choices.

The examples above are ones in which people fail to know what influences their behaviors. This shows at least some kind of failure to know the states and processes of their minds. The cognitive processes that led to their behaviors made use of information that they were consciously unaware of using and later felt to be irrelevant. I think that

[25] Hess 1975.
[26] Wilson 2002, pp. 100-102.

these kinds of results show that many cognitive processes work in ways inaccessible to us.

This conclusion has also been reached by leading researchers in psychology. Here is Daniel M. Wegner:

> The fact is that each of us acts in response to an unwieldy assortment of mental events, only a few of which may be easily brought to mind…we perform many unintended behaviors that then require some artful interpretation to fit them into our view of ourselves as conscious agent. Even when we didn't know what we were doing in advance, we may trust our theory that we consciously will our actions and so find ourselves forced to imagine or confabulate memories of 'prior' consistent thoughts.[27]

Timothy D. Wilson, who originally maintained (with Nisbett) that we have "little to no access to higher order cognitive process," has now backed down slightly to claim that "to the extent that people's responses are caused by the adaptive unconscious, they do not have privileged access to the causes and must infer them, just as Nisbett and I argued."[28]

Wegner thinks that the errors we make in judging whether or not particular actions were intended show that we have a tendency to attempt to fit a reconstruction to a preexisting theory. Wilson is prepared to admit some kind of privileged access to content and to conscious mental processes, yet still thinks that we are mostly confabulating when we try to reconstruct the structure of unconscious mental processes subjectively. The analogy he favors is, "introspection as personal narrative, whereby people construct stories about their lives."[29]

However, there is an obvious objection to the relevance of these examples. I introduced them to undermine the idea that we have access to the processes that form our beliefs. These processes, while apparently inaccessible, do not necessarily have anything

[27] Wegner 2002 pp. 145-146.
[28] Wilson 2002 pp. 104-106.
[29] Wilson 2002 p. 162.

to do with belief. Perhaps we cannot access the processes that lead to other kinds of mental states like emotions. The behaviors of the people in these studies may not have been influenced by their beliefs at all. If so, then we might dismiss all of these cases as irrelevant to epistemology, which concerns the justification of our belief.

This objection deserves a more full reply than I can give here. However, the central point of my response is that it is tricky enough to determine what beliefs a person has in normal circumstances. In the puzzling results of psychological experiments, our pre-theoretical abilities to ascribe belief to the people in these situations are strained to the breaking point. Take, for instance, the case of misplaced arousal. What should we say about a man that calls a woman for a date after meeting her on a treacherous bridge? It is clear that the arousal from the bridge is influencing behavior, but how? Does he believe that the woman is attractive in a way that he would not have had he met her elsewhere? Perhaps he simply feels attraction without believing anything unusual about the woman.

What is clear is that psychologists consistently find a disconnection between the psychological explanations offered by people to explain their actions and factors influencing the actions themselves. The disconnection is so severe that many psychologists claim that there are dual processes at work in generating behavior.[30] If they are correct, it becomes difficult to say how the behaviors presented relate to the beliefs of the subjects. One could try to hold that the subjects are getting their beliefs correct or even that the subjects are getting the sources of their beliefs correct. However, given the role that we—not to mention the subjects themselves—give to belief in behavioral explanations, it seems to me that the rational response to all the experimental data is a

[30] There is a wide literature on this. Anthologies that provide good starting places are Chaiken and Trope 1999 and Evans and Frankish 2009.

willingness to explore certain theoretical possibilities: that we sometimes do not know what we believe, that we sometimes do not know how we arrived at our beliefs, and that our beliefs do not always play quite the role in behavior that we tend to think they do. These options should all be on the table. Once on the table, they leave no room for blind trust in subjective reconstructions of our past beliefs, their formation, and their role in behavior. I am not suggesting that we give up on belief entirely. I merely suggest that we have not ruled out certain possibilities that, if they were to obtain, would seriously undermine the reconstructivist picture of justification. The reconstructivist is betting that we have access to the relevant states and processes forming and maintaining our beliefs. For all I have said here, the reconstructivist may still be right. If it turns out that we do not have the right access in some domain, those beliefs are not justified. In the case of self-knowledge, if the bet pays off, then meta-beliefs are direct. If it does not then they are unjustified. In either case, the bet will only be settled by empirical investigation. We cannot rule out *a priori* the possibility that we have beliefs or other mental states that play some epistemically significant role in the formation of our meta-beliefs.

3.3. Discursive Justification

Perhaps it does not matter if our justifications match any kind of cognitive process because what is typically important for our beliefs to be justified is that we be able to offer support when needed. Some internalists seem to have something like this in mind as a picture of justification.[31] The idea is that we must have *available* a justification for a

[31] For instance, Michael Williams seems to hold something like this view. See Williams 2001, pp. 146-150. Williams presumably gets much of his picture from Robert Brandom, but I worry that Brandom diverges from various assumptions I make throughout to such a degree that it is difficult to bring into discussion. Still, one can find interesting discussions of discursive practice in Brandom, such as his criticism of John

belief, not necessarily that the belief was actually formed on the basis of what is offered as justification. I am calling this notion of justification discursive because it is useful to think of an imaginary conversation between the believer and a challenger to test whether a particular belief is justified under this picture. The challenger can ask for support and the believer must offer support until the challenger is satisfied. This picture does describe some common discursive exchanges and, if we allow the believer and challenger to be the same person, some instances of self-conscious reflection. However, one should not think of discursive justification as something that requires an actual exchange. The core of the view is that justification consists in a disposition or ability: the ability to offer support for beliefs. While there are some internalists that do hold such a view, I have already pointed out that it conflicts with my assumption that meta-beliefs are justified even when not explicitly examined.

There are a variety of ways to flesh out a discursive notion of justification. One might hold that the available support must objectively support the belief it is meant to justify or that justification is relative to a context. Another variable is the notion of availability employed. One might define availability in terms of things one would be able to say in discussion, propositions that come to mind, or existing beliefs one has access to, among other options.

Discursive justification has certain features that some would view as advantages. It ties the notion of epistemic justification to the traditional intellectual practice of asking

---

McDowell's epistemology for not construing the "space of reasons" as social in Brandom, 1995. Some internalists closer to my assumptions surely also would not accept the necessity of causal dependence between justified belief and ground. Keith Lehrer, for instance, has at one point at least explicitly rejected this dependence. See Lehrer 1974, pp. 122-126. He uses 'evidence' in a way close enough to the way I am using 'ground'. This sort of picture may also be held by various pragmatists and contextualists, though I will not attempt to classify them here.

for and offering reasons for belief. A counter-intuitive feature of externalism is that it seems to imply that one can be justified in one's beliefs without being able to adequately defend them. In addition, by requiring that people be able to offer support for their beliefs in order for them to count as justified, it separates the justified beliefs of humans from the beliefs of animals and infants in a way that some philosophers seem keen to do. Moreover, it creates this separation at the level of justification and knowledge and not at belief itself, which seems a better place to posit a distinction.[32]

One serious complication in understanding the distinction between reconstructivist and discursivist justification is that existing beliefs are often checked or maintained through various processes. We must be careful to avoid characterizing justification as an active process of self-conscious belief-checking. While a version of a discursive justification theory could require you to be able to undergo conscious mental inferences that support your justified beliefs, a theory of justification based on having access to the actual maintenance processes that we undergo would essentially be a kind of reconstructivism. It would also be susceptible to the problem that we are bad at subjectively knowing how unconscious mental processes work. An internalist account of justification based on actual belief-updating either assumes the dubious claim that we can subjectively access belief maintenance processes or only counts as justified the small collection of meta-beliefs that we consciously check. One nice feature of a discursive account of justification is that, in virtue of making justification dispositional, it can allow beliefs to count as justified even if they have not been actively put through a process in which justifications have been offered for them. This is a convenient way for an

[32] I do think that non-human animals can have knowledge in addition to having belief, but I think it is more plausible to assert that they have belief and lack knowledge than that they lack both.

internalist to avoid skeptical conclusions about unconsciously formed beliefs. Now that the options have been laid out for the discursive justification theorist, we can consider whether self-knowledge must be counted as direct under a discursive notion of justification.

If one formulates a discursive theory in terms of having objectively good inferences available, then it seems obvious that we can have indirect justification for meta-belief. We very obviously have available the same sorts of inferences regarding our own meta-beliefs that we have for justifying ascription of beliefs to others. As I pointed out in setting up the discussion of directness, we *can* offer certain kinds of defense to support our meta-beliefs. We *can* say, "Well, I must believe P because it says P is true in the newspaper and I tend to believe propositions asserted in the paper." If we regard this sort of inference as good in a third-person case (and we had better on pain of solipsism), it is hard to see why the availability of the corresponding first-personal premises could not justify our meta-belief in the first-person case. Presumably such justifications will be available to us most of the time, perhaps even more often than they would be available to support our ascriptions of mental states to others.

Given the availability of these inferential justifications, it seems that the discursivist who favors treating SKID as a constraint should say that we have indirect justification that is typically trumped by our direct justification. However, it is not so clear to me that this will be the preferred strategy for all discursivists. Remember, one of the intuitions driving the idea that self-knowledge is direct was that these sorts of justifications were viewed as inappropriate. In our actual discursive practices, it seems to me that we rarely treat an inferential justification for a self-ascription of a mental state as

appropriate. If it was taken to be superfluous, common practice would dictate that we accept it when offered. For instance, if someone is asked to justify his or her believing that there is a tree outside the window, that person can provide more justification than necessary without the justification being deemed inappropriate. This suggests to me that common practice treats these justifications as something worse than superfluous. On the discursivist picture, if third personal justifications are deemed worse than superfluous, there is a mystery in our discursive practice. Why think that certain kinds of inferences are justificatory in third person cases but not in first person cases when they ought to do at least as well in terms of securing true belief? I am not sure that a discursivist could answer that question satisfactorily.

It would be better for the discursivist to hold that indirect justification was available but superfluous for meta-belief. Suppose that actual discursive practice does treat inferential justifications of meta-belief as superfluous rather than as non-justificatory. This would not show that we regard meta-beliefs as directly justified unless all there is to being directly justified is to be a belief for which we do not ask for justification. Surely the directness of self-knowledge is meant to be more significant than a claim about whether or not we typically ask for justifications. Presumably, it is a claim about whether or not we *should* sometimes ask for justifications. It would be dangerous to take facts about our discursive practice to demonstrate that no support is necessary to justify a belief. We cannot assume that our discursive practice reflects only epistemic norms. Perhaps we do not ask for support because it would be impolite. Even excluding the possibility of our discursive practice being influenced by non-epistemic factors, there might be alternative explanations for the existence of the practice. Previous generations

may have implicitly assumed that others had access to good justifications for their own meta-beliefs while never feeling compelled to ask them to produce reasons until this was codified into common practice.

If the discursivist does not take self-knowledge to be direct on the basis of the structure of discursive practice, it is unclear what he or she could base *a priori* acceptance of SKID on. It certainly can seem as if apparent claims to self-knowledge do not require justification, but this amounts to little more than a restatement of the facts of the practice. It seems that by asking whether or not apparent claims to self-knowledge *should* require justification, we are asking for an *evaluation* of discursive practice.

We clearly cannot rely on facts about actual discursive practice to conclude actual practice gets the structure of justification correct. While some discursivists might be tempted to do just this—to make epistemology nothing more than the generalizing of rules over what we do and do not accept as justificatory—such a view can become radically separated from truth. Even when practice tends to be truth-conducive, it is not safe to assume that it matches an ideal practice. Any sort of belief that people are overwhelmingly accurate about becomes relatively safe to refrain from asking for justification. This would hold even if there were substantive justifications available, and even if there were situations in which they ought to be regarded as holding the belief without justification.

Once again, there are reasons to doubt the intuitions that underlie *a priori* acceptance of SKID. For the reconstructivist, we saw that our intuitions about what justifies what were likely a poor guide to unconscious mental processes, which undermined the intuition that meta-beliefs form directly. I argued that this gave us reason

to distrust our intuitions as a guide to belief formation. For the discursivist, the problem comes from trusting our intuitions in guiding epistemic practice. In accepting SKID *a priori*, the discursivist effectively treats actual practice as a guide to ideal practice. We may not currently ask for justification for meta-beliefs, but a concern for the truth of our beliefs must lead us to question whether our discursive practice is a good one. Merely accepting our current practice is contrary to epistemic goals. It may lead us to assume that people are justified when they have formed a belief in a haphazard manner and have no available defense of it. The discursivist must be prepared to evaluate discursive practice and to discover potential flaws in it. Once we realize that the practice is open to evaluation, there is no reason to conclude that it currently gets things right with respect to apparent claims to self-knowledge. In light of this fact, the discursivist must regard SKID as a substantive thesis in need of support rather than an *a priori* truth that can be gleaned from current discursive practice.

4. Conclusion

I have argued that the idea that self-knowledge is direct is often motivated by two *a priori* considerations. First, we all lack subjectively felt processes of forming or maintaining meta-belief. Second, some philosophers have intuitions that the things we might say in support of our meta-beliefs do not seem to be genuine grounds for our meta-beliefs. Since the phenomenology of self-knowledge does not distinguish self-knowledge from other kinds of beliefs that we consider to be indirect, I conclude that the intuitions are doing more work than the phenomenology in establishing directness. I argued that, whatever one's theory of justification, these intuitions should not motivate *a priori*

acceptance of the directness of self-knowledge. If one takes the formation and maintenance of meta-belief to determine justification, one cannot trust our intuitions that self-knowledge is groundless to show that there are no significant mental states involved in the unconscious formation of meta-belief. If one thinks belief can be justified discursively regardless of the formation or maintenance of the belief, then one cannot reasonably infer from the current discursive practice of not questioning claims to self-knowledge that there are no important grounds for meta-belief. Ultimately, the directness or indirectness of self-knowledge must be determined empirically.

Chapter 3

In Search of Direct Self-Knowledge

1. Direct Knowledge

This chapter is an examination of the plausibility of the claim that self-knowledge is direct or groundless. The claim that self-knowledge is direct is often made as a pre-theoretical constraint on theorizing about self-knowledge. It is sometimes seen as one of the primary goals of constructing a theory of self-knowledge that the theory explain the direct character of self-knowledge and often seen as a roadblock for theories of self-knowledge that deny that it is direct. In the previous chapter, I argued that we should not allow intuitions that self-knowledge is direct to constrain theorizing about self-knowledge. However, I did not evaluate the plausibility of the idea that self-knowledge is direct. That is the goal of this chapter. I will primarily be examining Sydney Shoemaker's view of self-knowledge. While Shoemaker is not as committed an advocate of the view that self-knowledge is direct as some, I think that he provides the best model for an interesting view of direct self-knowledge. Before examining Shoemaker's theory, I will discuss the idea of directness and bring out several options for what the claim that self-knowledge is direct might mean.

Before turning to self-knowledge, it is useful to consider the role directness plays in another area of debate in epistemology. One important source of discussion of directness comes from its role in foundationalist theories of justification and knowledge. Foundationalism is a view of the structure of justification. It holds that some beliefs are justified directly as a means to avoid skeptical regress arguments. Looking at the kind of skeptical regress arguments foundationalism addresses will be helpful. I note here that I

am providing only a minimal sketch of a skeptical regress argument, not attempting to cover the intricacies of such arguments or add anything to the debate about them.

The skeptic begins by examining a belief that we would pre-theoretically judge to be justified. For example, we could begin with my belief that I am over five feet tall. Then, the skeptic raises questions about the justification of this belief. In normal conversation, I could respond by defending my belief. I might say that I was at the doctor's office recently and that they told me I was over five feet tall. But, of course, I offer these facts as justifications because I believe them. So, the standard thinking goes, those beliefs can provide justification for my belief that I am over five feet tall only if they are themselves justified. The skeptic can now ask for justification for the newly presented beliefs. If I respond with yet more beliefs, those beliefs have to be justified, and the skeptic asks still more questions. At this point, there seem to be a limited number of available strategies to stop the skeptic's questioning. It seems illegitimate to offer a circular series of beliefs and it seems difficult to maintain that a justified belief must be justified in virtue of a never ending chain of justified beliefs, though both options have been considered.[1] The most promising strategy, and the one favored in contemporary epistemology seems to be to find a belief which effectively cuts off the skeptic's line of questioning. In order to do that, a justified belief would have to be offered which did not depend for its justification on another belief.

Foundationalist accounts of justification all share the commitment to the possibility of justified beliefs that do not depend on other justified beliefs for their justification. We can call these beliefs basic beliefs, as they provide the base of the tree

[1] The former idea, that beliefs might be justified in something like a circular series would be a version of coherentism. Various philosophers have held such views, though most have attempted to avoid circularity. The later is most prominently held by Peter Klein. See Klein 1999.

(or pyramid, if you prefer that metaphor) of justification. Beliefs above them in the structure are to be justified, ultimately, in virtue of their connections to basic beliefs. Foundationalist accounts differ predominantly in their accounting of basic beliefs. Different accounts include different classes of beliefs among the basic beliefs and they can offer different accounts of the justification of these basic beliefs. It is to this facet of foundationalist accounts I now turn.

One class of foundationalist theories accounts for the justification of basic beliefs by appeal to factors external to the mind of the believer.[2] Such a view might allow for justified beliefs if the believer is reliable in a certain domain or if there is a proper causal connection between the believer and his or her beliefs. These external factors are justifying whether or not the believer can cite them. This effectively cuts off the force of the skeptic's questioning. The externalist does this by refusing to answer, but it is a principled refusal. Though the skeptic can still frustrate an externalist willing to engage with skeptical questioning, the externalist holds that there are can be justified beliefs which are justified despite his or her inability to provide justifications.

Another class of foundationalist theories seeks to explain the justification of certain basic beliefs in terms of their relations to mental states other than beliefs. For instance, perceptual sensations are sometimes thought to provide the basis for the justification of basic beliefs; these perceptual states may be thought to justify either beliefs about the existence and properties of external objects, or beliefs about the appearances of such objects and their properties. This second sort of view actually

[2] Bonjour provides a nice discussion of externalism as a foundationalist theory in his attempt to criticize those theories. See Bonjour 1980.

privileges a kind of self-knowledge, though not the kind that ultimately concerns me in this paper.

Finally, basic beliefs are sometimes seen as self-justifying. Certain easily grasped logical truths, for instance, may be such that anyone that entertains the question of their truth would be justified in believing them. While it is a promising suggestion that certain beliefs are self-justifying, there are obvious worries about the scope of such beliefs and what they can justify. Descartes, for example, may have been correct that one is always justified in believing in one's own existence. The problem is that not much else seems to follow from one's own existence. Certainly, most of our beliefs about the external world are not obviously justified in light of such a belief.

Above, I have outline three kinds of directness for justified belief. In each case, basic beliefs are presented as deriving their justification directly, in contrast with beliefs that are justified indirectly, via their relations to other justified beliefs.

2. Direct Self-knowledge

Now that we have seen the notions of directness at play elsewhere in epistemology, we can turn attention to the notions of directness at work in the claim that self-knowledge is direct. First, as I noted above, there is a brand of foundationalism which contains a central role for a kind of self-knowledge. Specifically, the view that beliefs about appearances are basic and justified, where appearances are understood to be mental, is a view of a certain kind of self-knowledge as direct. However, I am here not concerned with beliefs about appearances. I am concerned with our knowledge of our own beliefs.

As discussed in the previous chapter, recent discussion surrounding the directness of self-knowledge traces back to Gilbert Ryle's rejection of the idea. Ryle held that we know our own minds the same way we know the minds of others, by inference from observational evidence. Advocates of the directness of self-knowledge think that our beliefs about our own beliefs are (typically) justified, but they are not justified in virtue of evidence or observation. But just how or why they are justified is something of a mystery.

Perhaps we can understand the claim that self-knowledge is direct in terms of the discussion of directness in the previous section. If so, we might assimilate the idea that self-knowledge is direct onto one of the three models above. I will take each in turn. First, we can consider the externalist account of directness. On such a notion of directness, our beliefs about our own beliefs would be justified just in case we happened to satisfy some external condition, such as reliability with respect to our beliefs. While most are willing to concede that we are reliable in this way, this view of the directness of self-knowledge leaves open the mystery of why we are so reliable. Second, we have the possibility that there are certain non-belief intermediary mental states, which provide justification for our beliefs about our own beliefs. While some may accept this characterization of self-knowledge as introspection, it seems somewhat phenomenologically implausible, especially in the case of belief.[3] If we are to understand these intermediary mental states on the model of appearances and their role in justifying beliefs about the external world, then we would be committed to it appearing or seeming a certain way that we had a particular belief. Many have found this counter-intuitive or at

[3] Though I maintain its implausibility, it is worth noting that Alvin Goldman holds a view somewhat like this in Goldman 2004.

odds with subjective phenomenology. Discussion of the last possibility, that our beliefs about our own beliefs are self-justifying, is reserved for a later section.

Either of the first two views would constitute an interesting position on the epistemology of self-knowledge. Both stand in need of the filling in of significant details. However, they can mostly be discounted here since neither makes the directness of self-knowledge unique. To the extent that the directness of self-knowledge is a puzzle on these views, it is the same sort of puzzle that exists elsewhere in epistemology. Moreover, these sorts of views of the epistemology of self-knowledge are routinely criticized by theorists that discuss the directness of our knowledge of our own beliefs.

Note something about the conceptions of directness discussed this far. In each case, the directly justified belief was said to be justified in light of its relation to something other than another justified belief. So, while there was no justified belief that provided justification, each view considered so far posits something to do the work of justifying the directly justified belief. These beliefs then count as basic only insofar as they do not derive their justification from other beliefs; there is a sense in which these beliefs are still justified indirectly. To get a better handle on the notion of directness we are after, we should return to the notion of an epistemic ground. As before, my understanding of an epistemic ground is that they are mental states that play a mediating role in the creation or conferral of justification on a belief. We can understand any notion of directness, then, as a lack of epistemic grounding.

If our knowledge of our own beliefs is not grounded in other justified beliefs, there is a clear sense in which we can call it groundless and direct. However, theorists concerned about the directness of self-knowledge seem concerned to articulate something

more dramatic. After all, many different sorts of beliefs have been proposed as candidates for being direct in this sense and the need to have direct beliefs of this sort is felt by most epistemologists. If our beliefs about our own beliefs turn out to be among the sort of basic beliefs that we can use to halt the skeptical regress, so much the better for epistemology—though we should worry a bit that they may not be much help in building justification for many of the beliefs that do seem prey for the skeptic. In the next section, I will attempt to locate a more interesting notion of directness sometimes attributed to our knowledge of our own beliefs.

Different notions of directness can be understood in terms of different conceptions of epistemic grounding. In the previous chapter, I considered several different conceptions of directness. The most important two were as follows:

> Belief Conception of Directness (BCD): S's belief that P is directly justified if and only if it is justified and the belief that P does not depend on any other justified belief of S for its justification.
>
> Mental State Conception of Directness (MSCD): S's belief that P is directly justified if and only if it is justified and no mental state of S is involved in an epistemically relevant way in S's belief that P being justified.

The project of foundationalism is to find beliefs that are directly justified according to BCD in order to block the skeptical regress. While it would be a significant claim that self-knowledge is directly justified according to BCD, it would not necessarily threaten the naturalist. Directness according to MSCD is might be more difficult for the naturalist to handle, though some varieties of externalism might still safely countenance self-knowledge as direct under MSCD. Essentially, the move would be similar to the move from considering only beliefs as epistemic grounds to considering a wider variety of mental states. The externalist can be seen as allowing non-mental states of affairs to count as epistemic grounds. In the next section, I will consider the idea that theorists of

self-knowledge concerned with puzzles of directness are rejecting moves of this sort. That is, the reason the directness of self-knowledge is taken to be such a puzzle is that theorists have set themselves the goal of explaining self-knowledge as absolutely groundless, with nothing playing a mediating role in the creation of justification for self-knowledge.

3. Self-Knowledge as Absolutely Groundless

In the previous section, I reintroduced the notion of an epistemic ground and proposed an understanding of epistemic directness in terms of a lack of epistemic grounds. While there are clearly some epistemologists that would want to allow only other beliefs to count as epistemic grounds, I noted above that certain views of direct belief essentially shift the justifying work of an epistemic ground on to things other than other justified beliefs, creating a sense in which those beliefs are not genuinely direct; they receive justification from something which could plausibly be called an epistemic ground. In the next several subsections, I consider the possibility that what is really special about the proposed directness of our beliefs about our beliefs is that they are justified without resort to such a philosophical move. That is, our self-directed meta-beliefs are completely groundless, on any conception of epistemic grounding.

3.1. Claims of Groundlessness

It may be useful to review the sorts of claims made about the directness of self-knowledge. Different philosophers tend to employ their own idiosyncratic terms, but I think there is discernable thread throughout these comments which can be understood in

terms of directness and epistemic grounding, as I have laid out. Here, for instance, is Crispin Wright:

> The demand that somebody produce reasons or corroborating evidence for such a claim about themselves… is always inappropriate. There is nothing they might reasonably be expected to say. In that sense, there is nothing upon which such claims are based.[4]

Here, Wright is a discussing what he and others call first-person avowals, which are utterances of first-person, present tense, self-ascriptions of mental states. So, if I were to say, "I am experiencing some pain in my back," or "I believe there is some ice cream in the fridge," my utterance would count as an avowal. Specifically, in this passage, he is claiming that avowals ascribing phenomenal states like pain are baseless, though he later attributes the same property to attitudinal avowals, like those that self-ascribe beliefs. Similar claims are endorsed by McDowell and Bar-On.[5] These philosophers share a common influence in Wittgenstein. While it is far from clear how to interpret Wittgenstein on these matters, and it would drag us far afield to attempt such a monumental task, we can try to locate some unity among his followers.

They all seem to agree with Davidson, in the earlier quotation, that apparent claims to self-knowledge are not based on observation or inference. For instance, Dorit Bar-On writes:

> Now consider the situation from the point of view of the person issuing the avowal… in the normal case, I do not consult my own behavior or memory; I do not reason, or draw some inference… for all appearances, my avowal is baseless.[6]

[4] Wright 1998, pg. 14. The quotation actually refers directly to phenomenal avowals, but he claims that attitudinal avowals have the very same groundlessness.
[5] See McDowell 1998, pp. 47-50, and Bar-On 2004, p. 3.
[6] Bar-On 2004, p. 2.

That there is no self-conscious inference is a claim about phenomenology. Bar-on is certainly correct in asserting that we do not typically experience such a process before issuing avowals.

While there is space for thinking that some sort of inference or inferential process provides the justification for the belief despite the lack of experienced inference, we can here take this claim at face value. However, this claim does not distinguish self-knowledge from perceptual knowledge. For instance, Richard Moran writes, "The claim that introspective awareness is not inferred from observational evidence is what is usually intended by the claim that it is 'immediate'." He continues to clarify his notion of immediacy, saying it "is to be understood as a wholly negative claim about the mode of first-person access, that is, awareness that is not inferred from anything more basic." However, he goes on to make the somewhat surprising claim that, "much of our ordinary perceptual awareness is also taken to be immediate in this sense".[7] This claim is surprising because all four theorists mentioned are at odds to explain what they dislike about perceptual models of self-knowledge. It seems incumbent on them to find a notion of directness that cannot be matched by a perceptual model. John McDowell takes Crispin Wright to task for failing to do this well enough here:

> And the authority of observations is indeed *non-inferential*. But it is precisely *not* baseless. The question 'How can you tell?' is precisely *not* excluded as inappropriate… he accepts it and answers it when he says, perfectly properly, that he is in a position to see.[8]

It seems that McDowell is unwilling to accept anything as a ground for self-knowledge, even the statement of external conditions, such as being in the proper position to know.

[7] Moran 2001, p. 11.
[8] McDowell 1998, p. 48.

Ultimately, all these philosophers seem to want an account of self-knowledge that is of a wholly different character than perceptual models. This idea is given succinct expression by Akeel Bilgrami, another prominent theorist with views similar to Richard Moran. Bilgrami puts the claim this way:

> Knowledge of the external world is in ordinary cases the result of such *cognitive achievement* (as philosophers call it), which if the agent does not do or have happen to her, then she will not form the belief that there is some fact or object in the world. In self-knowledge there is, in the ordinary case, no such cognitive activity nor any analogue to it, which is required to believe that one has some mental state… to put it as I have in various papers over the years, it comes, as it were, 'for free', without any cognition.[9]

Though the theorists mentioned above obviously have diverse philosophical commitments and do not share a unified view of the epistemology of self-knowledge, it does seem clear that these philosophers share a common disregard for 'observational' theories of self-knowledge. Specifically, all these theorists seem to think that our knowledge of our own beliefs is not grounded in anything, even the sorts of external facts or non-rational processes that could be said to ground basic or foundational perceptual beliefs.

3.2. The Phenomenon of Transparency

One wrinkle in the self-knowledge literature is that, while self-knowledge is often thought to be direct, it is also often claimed to be transparent. It is not always clear exactly how transparency is meant to relate to directness. In order to fully evaluate whether or not self-knowledge is direct, we will have to consider the possibility that it is transparent. Here is how Gareth Evans frames a phenomenon that has come to be called transparency:

[9] Bilgrami 2004, p. 31.

> 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?'. I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p.[10]

The propositions expressed by "There will be a third world war" and "I believe that there will be a third world war" are quite different. They have different truth conditions. Why would a person direct his or her attention to the same facts when deciding whether such different propositions are true? More to the point, how could we be justified in doing this? [11] I claimed that the question "How do you know that you believe Barack Obama is the president?" was importantly different from the question, "Why do you believe Barack Obama is the president." The suggestion that Evans makes here is that the questions are not so importantly different.

In pointing out that the answer to a question about one's own mental states is sometimes addressed by consideration of external facts, people sometimes seem to think that the phenomenon of transparency explains or underlies the phenomenon of directness.[12] Their idea seems to be that once you come to have a belief about the third world war, you will unfailingly know that you have the belief without any further mental

---

[10] Evans 1982, p. 225.

[11] One thing that might be said about the Evans example is that most people do not have an existing first-order belief of any sort about a third world war before the question is asked. For those that favor thinking of self-knowledge as a kind of introspection or a perception of our own minds, it is tempting to say that we would have to form the belief before we can see it. Most of the time, when asked, such questions are not really meant to be exclusively about what a person believes. Typically, this sort of question is more a prompt for a discussion about the possibility of WWIII and for the interlocutors to offer reasons it may or may not occur. We recognize this implicitly and we rarely respond with a literal answer. Instead we consider the question and form a belief. Once formed, such a belief should be immediately available for introspection. This is a plausible story that I will not explore any farther at this time.

[12] For instance, Akeel Bilgrami's notion of transparency is as a presumption that meta-beliefs occur whenever first order beliefs occur. Bilgrami 2006, p. 31. Richard Moran also seems to think that this transparency phenomenon is key in explaining the difference between the direct self-knowledge and theoretical self-knowledge as he thinks transparency fails precisely when we form self-knowledge in the theoretical way. Moran 2001, pp. 60-69.

activity. This is the kind of transparency that Descartes's theory of mind included.[13] He thought your mind was perfectly transparent in that you would unfailingly know that you had any mental state just in virtue of having it. Though most view Descartes as wrong for thinking that our minds are infallibly transparent, it seems a relatively uninteresting consequence of his theory of mind that something like the Evans example would be true. Once a belief is formed about whether or not there will be a third world war, the believer will know that he or she has it. The more puzzling thing on the Cartesian conception of transparency would be how the kind of reasoning Evans suggested could ever fail to work. If our mind is typically or even can sometimes be transparent in this Cartesian way, what can obscure it during failures of self-knowledge?

However, it is not at all clear that this phenomenon deserves to be equated with Cartesian transparency. Once we have given up the Cartesian idea that the mind is perfectly transparent, it becomes a significant assumption to think that there is no epistemic intermediary between the first-order belief and the meta-belief. For instance if we have an internal scanner that can detect our beliefs, it might simply detect the newly formed belief as it is formed, registering it unconsciously. Another option would be that there is something like an unconscious inference drawn from the truth of the proposition considered to the separate proposition that you believe it. That is, once you take P to be true (once you believe P), you might have a process that takes P as an input and outputs a meta-belief. There are other possibilities, each of which involves some sort of epistemically relevant intermediary potentially worthy of being called an epistemic ground.

[13] Shoemaker 1996, p. 50.

If there is an extra unconscious step, then this phenomenon should be well understood already. It is not that unusual to get ourselves into a position to know one thing by learning another. Knowing the numbers showing on the scale puts me in a position to know how much an object weighs. Knowing 2+2=4 and 4+8=12 puts me in a position to know 2+2+8=12. We often get at the fact we want to know indirectly even when forming the belief seems direct. Take the scale for example. When I stand on the scale and look down at it, I seem to instantly come to know how much I weigh. But how could this be! The proposition expressed by, 'The scale reads 175lbs' has different truth conditions than the one expressed by, 'Jeremy weighs 175lbs'. I investigated the scale's display yet somehow came to believe that I had a certain weight. Of course, no one is surprised or perplexed by our ability to do this. All it means is that we have learned how scales work and automated a belief forming process. We may have a similar automated process governing our self-knowledge. That is, we might automatically register a meta-belief as we form our first-order beliefs. No matter how direct this feels, it would not count as absolutely groundless.

A slightly different response to the Evans example is the suggestion that the grounds for holding a first-order belief might be the very same grounds for holding a meta-belief.[14] That is, we can view transparency as a relation between beliefs in which one belief has identical grounds to another. On this notion of transparency, we do get a genuine epistemic puzzle. The puzzle arises because we take our meta-beliefs to be justified, but the grounds we accept as justifying our first-order beliefs seem poorly suited to justify meta-beliefs. As noted above, the truth conditions of a proposition and a self-

[14] See Fernandez 2003, pp. 360-361, Bar-On 2004, p. 113, Martin 1998, p. 114, among others.

directed ascription of belief in that proposition are quite different. It is hard to see how good grounds for believing one could be considered good grounds for believing the other. Conceived this way, transparency is worth further discussion. However, the purpose of the chapter is to understand the apparent directness of self-knowledge and this construal of transparency is a direct denial of groundlessness. The idea that we have the same grounds for our meta-beliefs that we have for our first-order beliefs, transparency entails that we must have *some* grounds for our meta-beliefs. Since the interesting notion of transparency is a denial of directness, further discussion of the phenomenon will have to wait, though it is worth noting that some have attempted to use a notion of transparency to generate a model of self-knowledge.[15]

3.3. Belief as Self-intimating

Above, I considered three models of directly justified belief. I have since ruled out two of them as useful in constructing an account of direct self-knowledge of the sort many theorists seem to want. Now, I want to turn to the neglected final option: belief as self-justifying.

Clearly, we do not want to regard all self-directed meta-beliefs as justified. I can be wildly mistaken in attributing a belief to myself. For instance, I may be confused by a bad argument, forcefully delivered by a therapist, to the effect that I believe I chose the wrong career. In such a case, mere possession of the meta-belief does not seem to justify it. To be clear, I am suggesting in this example that a therapist might convince me that I have the belief that I have chosen the wrong career as a way to explain, say, my vague

[15] See Fernandez 2003 and 2005 and Williams 2004 and 2006.

sense of discontent, not convince me that I in fact have chosen the wrong career. At a minimum, we ought to require that the believer has not arrived at the meta-belief in an unusual way, for the meta-belief to count as self-justified. But even this would seem to leave out something important. Free floating meta-beliefs should not count as self-justifying independent of some connection to the beliefs that they are about.

One way for a belief to be connected to a meta-belief about it is for the lower-order belief to signal its presence to the believer. Such a model is clearly quasi-perceptual. Something in the believer's mind would have to register the signal from the belief and form a meta-belief. Thus, even though this would provide some assurance of a connection, it loses the possibility of counting as direct. There is a causal process and a cognitive mechanism, which makes self-knowledge a cognitive achievement. Clearly, this is not what we are looking for. To understand meta-belief as self-justifying, we need to elaborate a connection between belief and meta-belief that explains why we should view the meta-belief as justified without providing something to do the work of an epistemic ground.

The problem is essentially caused by the separation of the belief from the meta-belief. As long as we conceive of them as separate and distinct states, we will be looking for assurance that they will occur together to regard the meta-beliefs as justified. The more detailed the supposed connection, the more likely we are to see some feature of the connection as an epistemic ground. This is, essentially, the problem that one will always run into when trying to give an account of directly justified belief. In other contexts, beliefs can count as justified because the features which are candidates for being considered an epistemic ground are different from the sort of epistemic grounds one is

trying to avoid. If one is looking for a directly justified belief to stop a skeptical regress, one only needs to avoid other beliefs as epistemic grounds. A foundationalist can concede that there is a sense in which foundational beliefs are indirect, as long as they are direct in the sense that halts skepticism.

Self-knowledge, however, may have a unique way to avoid any sort of epistemic grounding. Given the similar character of belief and meta-belief—unlike say, perception and perceptual belief—the self-knowledge theorist may deny that belief and meta-belief are wholly distinct states. This sort of view is called a constitutive account of self-knowledge. Essentially, the first-order belief constitutes knowledge of itself, in the right circumstances. While the previously mentioned Richard Moran and Akeel Bilgrami each have versions of constitutive accounts of self-knowledge, I want to turn my attention to an older and better known constitutive account: that of Sydney Shoemaker.

4. Shoemaker's Constitutive Account

In a series of papers collected in *The First Person Perspective,* Sydney Shoemaker outlines a theory of self-knowledge.[16] As he sees his view, he is pushing back against the tide of anti-Cartesian thought in contemporary philosophy. Shoemaker, like most contemporary philosophers, rejects large parts of the Cartesian conception of the mental, but he sides with Descartes in thinking that there is a special authority to self-knowledge resulting from its status as directly known. Shoemaker is always clear that he endorses a restricted kind of authority short of infallibility but in the spirit of Descartes.

[16] See Shoemaker 1996.

4.1. The Account

At times, Shoemaker can sound as if he endorses a quasi-perceptual view like the one outlined above. For instance, he accepts a version of a self-intimation thesis: "that it belongs to [intentional states'] very nature that having them leads to the belief, and knowledge, that one has them, or at any rate that it normally does so under certain circumstances."[17] However, as noted above, this claim is consistent with meta-belief being indirect. Having beliefs might somehow provide good grounds to infer, or enable some mechanism to detect, that those beliefs exist.[18]

Yet Shoemaker, like the philosophers mentioned above, is very concerned to differentiate his view from any sort of perceptual model. His view is not a view in which there is a tight causal connection between first-order beliefs and meta-beliefs. In the Royce Lectures, he asserts that, "the first-order belief and second-order belief have the same core realization," and denies that what is added to this core to produce second-order belief, "pushes the creature into a new state, distinct from any it was in before."[19] Indeed, nothing is added at all. So he does not think of self-intimation in terms of a causal process. His view amounts to the following claim: a rational person with an understanding of mental concepts like belief and desire will necessarily have self-knowledge of whatever first-order beliefs he or she has, provided that person reflects and barring a failure of rationality.

[17] Shoemaker 1996, p. 51.
[18] Shoemaker's view denies any sort of inferential or evidential connection between belief and meta-belief; I merely want to point out the differences between self-intimation and directness and that someone could hold one without the other.
[19] Shoemaker 1996, p. 244.

Some caveats are worth noting at the outset. First, Shoemaker's account makes a central role for the concept of rationality, which he is admittedly lax in describing.[20] Second, though Shoemaker permits the possibility that an unreflective person can be in error as to what he or she believes, he does not see reflection as a matter of inferring from behavioral evidence that one has or lacks a certain belief. That would be a kind of causal connection that he denies. Rather, his claim is that rational people act as if they know what they believe and that this is good enough to show that they do know.[21] Finally, he allows for individual failures of rationality to result in individual failures of self-knowledge.[22]

To reiterate, Shoemaker's view is that any rational, minimally reflective person in possession of mental concepts will automatically count as knowing what he or she believes. This kind of account makes self-knowledge automatic, not in the sense of being the result of a mechanism that automatically functions, but in the sense of something one counts as having in virtue of having other things. Perhaps a helpful analogy would be that one automatically counts as having citizenship in virtue of possession of various legal documents. The documents *make* one a citizen, but not in a causal sense of 'make'. So far, this account is providing a promising start to providing a notion of directness for self-knowledge in which it is absolutely groundless. The idea that self-knowledge is constituted by rationality and possession of mental concepts seems so far consistent with Bilgrami's idea that self-knowledge is not the result of a cognitive achievement.

---

[20] Shoemaker writes: "All of this, I realize, puts a rather heavy burden on the concept of rationality. Fortunately, that is a matter for another essay, which I haven't the slightest idea of how to write." Shoemaker 1996, p. 49

[21] Shoemaker 1996, p. 34

[22] Shoemaker 1996, p. 49

At this point, it would be helpful if Shoemaker's account could be spelled out in more direct detail. In particular, one might want to know more about the way in which self-knowledge is supposed to be constituted. The relation of constitution has, in other contexts, been thought to be philosophically problematic.[23] One might wonder if it is a part-whole relation—with the meta-belief being part of the belief, or both existing as part of some larger state. The fact that self-knowledge can fail, on Shoemaker's account, suggests that he thinks belief and meta-belief can occur separately. It is difficult to square this with a notion of constitution that can seem like identity at times, where the knowledge that one has a belief is said to be nothing over and above having a belief and being rational. Constitution can seem like a relation trying to find a middle ground between identity and distinct existence—and it is far from clear that there is such a middle ground to occupy. He does occasionally put his view in terms of more well established metaphysical relationships, such as supervenience. For instance:

> What I am inclined to say is that second-order belief, and the knowledge it typically embodies, is supervenient on first-order beliefs and desires—or rather, it is supervenient on these plus a certain degree of rationality, intelligence, and conceptual capacity.[24]

However, this still leaves questions about the extent to which self-knowledge is distinct from the qualities on which it supervenes.

Unfortunately, Shoemaker does not provide direct assistance in understanding this relation. His view is elaborated in contrast with views he rejects. Since Shoemaker is most concerned to reject perceptual models of self-knowledge, a large part of the argument in favor of his view is the argument against perceptual models. The strategy that Shoemaker employs to defend his constitutive view is to reduce to absurdity a

[23] See Zimmerman 2002 for a critique of constitution as a metaphysical relation.
[24] Shoemaker 1996, p. 34.

scenario that he thinks all non-constitutive theories are committed to. The scenario he dubs "self-blindness", and describes it as follows:

> A self-blind creature would be one which has the conception of the various mental states, and can entertain the thought that it has this or that belief, desire, intention, etc., but which is unable to become aware of the truth of such a thought except in a third person way.[25]

His choice of terms in calling this scenario 'self-blindness' is not accidental. He thinks that positing a mechanism by which a person monitors or senses his or her own beliefs should create the possibility for systematic failure of the mechanism. His claim is not that we should expect to see self-blind people if perceptual models of self-knowledge were correct. If there were a mechanism as the perceptual theorist suggests, then it may be physically impossible for it to fail without more widespread problems than self-blindness. But Shoemaker is adamant that non-constitutive models of self-knowledge entail at least the logical possibility of self-blindness.[26] It is this possibility which he seeks to reduce to absurdity.

His argument proceeds as follows. He first supposes that a particular person is self-blind. He calls this person George. He then attempts to reduce this possibility to absurdity by showing that George would be functionally identical to a normal person. It should be noted here that Shoemaker is a functionalist about the mind. This plays a fairly significant role in the framing of the argument, but I take it that most would accept that if George is really self-blind, he ought to be somehow different from a normal person. Shoemaker's goal is to reduce to absurdity the idea that self-blind George is any different from you and me. Interestingly, the direct conclusion from Shoemaker's argument is not that self-blindness is logically impossible. Rather, it is that that no one can be self-blind

[25] Shoemaker 1996, pp. 30-31.
[26] See Shoemaker 1996, p. 31.

or everyone actually is.[27] Though I ultimately find fault with Shoemaker's argument, I think it does have implications for the theories of self-knowledge. However, I think Shoemaker draws the wrong lesson from his argument. I think it should push us closer to the view that everyone is self-blind.

4.2. The Argument from Moore's Paradox

In order to reduce to absurdity the idea that George is self-blind, Shoemaker makes use of a peculiar kind of sentence. It is a sentence which asserts some proposition conjoined with a self-ascribed belief to the contrary. Thus, an example of such a sentence might be, "It is Friday and I believe it is Tuesday." Discussion of such sentences is usually traced back to G.E. Moore. The classic example is, "It is raining, but I don't believe it is raining". From this point forward, I will refer to all such sentences as Moore-paradoxical sentences (MPSs).

There is no inherent contradiction in an MPS. Both conjuncts can be true at the same time, as the truth conditions for a proposition about the world (it is raining) are very different from the truth conditions for a proposition about a person's psychological states. We are all easily able to understand, for example, how it might be true that it is raining when John does not believe it is raining. In fact, we are all perfectly capable of understanding how an MPS could be true even of oneself. The trick is that there is something funny about asserting one.

[27] He acknowledges this, but quickly brushes it aside without argument. See Shoemaker 1996. p. 36

Since there is no contradiction in an MPS, it is not clear what creates the oddity in attempting to assert one. Shoemaker's analysis of the oddity is that such an assertion would not communicate anything. In his words:

> Since, in asserting the first conjunct one would, if sincere, be expressing the belief which the second conjunct denies one has, one could not hope to get one's audience to accept both conjuncts on one's say so, and could have no hope of getting them to accept either.[28]

Thus, Shoemaker posits a connection between the expression of a belief and the assertion of its content. The first conjunct would be seen as an expression of one's belief that P and the second as a denial that one had that belief. A listener, then, would not be able to decide what it was the asserter of an MPS really believed, and would withhold judgment about both the contents of the assertors' beliefs and the proposition asserted.

Shoemaker assumes that George, being self-blind, would be more likely than most to get himself into circumstances in which he might truthfully attempt to assert an MPS. Since he has no first-personal way of knowing what his beliefs are, he is forced to rely on third-personal evidence. Since his evidence about his own beliefs is separate from his evidence about the world at large, faces the possibility that his evidence will point to rain while also pointing to his not believing that it is raining. Without any independent check on his beliefs, George should be in a position to assert a Moore-paradoxical sentence. Shoemaker's argument is, essentially, that George would still not be in a position to assert such a sentence. So, if self-blindness were possible, George could assert MPSs. He cannot. So he is not self-blind, contrary to the assumption that he was, showing the logical impossibility of the concept.

[28] Shoemaker 1996, p. 35.

I am inclined to grant that the self-blind person should be able to get into situations in which he or she might be expected to assert MPSs. Then again, I am inclined to grant that normal people can get into such situations. So what we have to see is why Shoemaker thinks George will not be in a position to make such assertions.

Shoemaker first points out that George would recognize that asserting an MPS involves some sort of impropriety. George, like us, need not have a worked out view of the nature of that impropriety. The important point is that George understands what beliefs are. He is, by stipulation, in possession of normal mental state concepts. Further, he is rational. As a rational man, he will avoid self-defeating assertions. As Shoemaker puts it, "Since we can assume him to be a rational man, we can assume that this recognition would lead him to avoid Moore-paradoxical utterances."[29] What leads him to avoid such utterances is, presumably, his intuitive grasp of the nature of assertion and its relation to belief.

Of course, George's avoidance of the assertion of MPSs might be thought, for everything said so far, to be merely a bizarre failure of the expressive power of language. On this picture, we might think of George being willing to assert an MPS, but refraining out of practical considerations, given that no one would believe him. George's situation so understood, would be a little like a man who saw a UFO, but refrained from telling his friends about it, knowing that they would judge him crazy and not believe him anyway.

But Shoemaker continues the argument. He claims, not merely that George refrains from asserting MPs, but that George can use the word 'belief' in a functionally identical way with everyone else, including in self-ascription. In other words, it is

[29]Shoemaker 1996, p. 35.

functionally as if George knew what he believed, just like the rest of us. Shoemaker envisions a situation in which George believes some P, but, being self-blind, is unaware of this belief. Shoemaker claims that, contrary to what one might expect, George would, in such a situation, be able to assert P or even that he believed that P. Here is Shoemaker:

> But it seems to me that George would have a motive for saying 'P,' or, 'I believe that P,' in this case. He could reason as follows. 'P is true. [This expresses his belief, but it of course doesn't say that he has it.] It is therefore in anyone's advantage, by and large, to act on the assumption that P is true… since this to applies to anyone, it applies to me,… In this instance, so acting would mean saying, "I believe that P," or just, "P."… Having given himself a good reason for saying, 'P,' or "I believe that P," he then says one of those things. It seems to me that he would always have such an argument available to him, in any case in which it would be rational for a self-aware speaker to avow his belief.[30]

There are several points that demand response. First, if Shoemaker's goal is to establish functional identity between George and a normal human being, he does not succeed here. At best, he establishes that George can function *outwardly* identically to a normal person. He says and does the same sorts of things. But if George ever has to go through anything like the convoluted reasoning above, his mind is clearly working differently than yours and mine. Functionalists like Shoemaker can make reference to the mental causes and mental effects of mental states in determining the makeup of the minds of others. Without this, after all, we would be forced to admit any computer program capable of producing linguistic strings like you, me, or George would have the same kind of mind as us, without knowing anything about the way in which it produced those words.

Second, since Shoemaker's move is to essentially allow George to go from his assertions or his willingness to assert, which are a function of the expression of his beliefs, to assertions of his beliefs, it is surprising that he takes George on such a

[30] Shoemaker 1996, p. 41.

roundabout path. Learning a simple rule, like, "When I am prepared to assert P, I should be prepared to assert that I believe that P," will do the trick more cleanly and more quickly. This would still, apparently, represent a functional difference between George and a normal person, but it would save a lot of mental resources.

Finally, given that Shoemaker links assertion to expression of belief, it is not clear what would prevent George, or any self-blind person, from learning to read their own minds by checking on what they would or would not assert. George, being intelligent and rational, ought to be just as capable as anyone of noticing that he makes assertions. Shoemaker already attributed to George at least an intuitive understanding of the connection between assertion and belief. So, George ought to be able to learn to form beliefs about his own beliefs via his assertions. Granted, it would be strange that he would have to check, in essence, to see what he would say before he could say what he thinks. We tend to think normal people do it in the opposite direction, but there is little to see what would stop George from going that route.

Notice that such a mechanism, one in which one's assertive powers were used to form beliefs about one's mind, would be a kind of perceptual mechanism for self-knowledge. One might think, then, that Shoemaker succeeded after all. George was supposed to be self-blind, but, contrary to this assumption, he turned out to be capable of developing a reliable mechanism for knowing his own mind. However, this mechanism would be clearly definable in causal terms. It represents separate states of belief and meta-belief, and meta-belief is arrived at via inference. So all we really know about George is that, given the capabilities Shoemaker has assigned him, he would be able to arrive at knowledge of his own mind in a very indirect way.

An assumption doing a lot of work throughout Shoemaker's argument is that a person who believes P will be able to assert P as an expression of his or her belief. Certainly we regard this as true in the ordinary case. However, would this always be true for the self-blind person? If so, then the reasoning above is sound. However, consider a case in which self-knowledge fails a normal person. Suppose one has a hidden unconscious belief that he is a failure. Shoemaker accepts such a possibility. Further, he grants that such a person, "is not someone who is prepared to assert 'I am a failure' but not 'I believe I am a failure.' Presumably, he will be no more disposed to assert the first than the second and in fact can be expected to deny both."[31] Here, Shoemaker leans heavily on the rationality of the man and its failure in this case. Shoemaker seems to think that man's hidden belief might influence his behavior in certain ways, but, overall, the behavior of the man will fail to rationalize with the rest of his beliefs and desires. I take it that, George, who is stipulated to be rational, could not similarly find his behaviors in conflict with his beliefs, even if he did not know what they were.

This gets into tricky ground about the function of belief and meta-belief. Shoemaker's assumption that George would be able to express his belief that P without knowing that he believed P, attributes assertion to the functional profile of belief. But, perhaps, we could make more sense out of actual failures of self-knowledge if we understood meta-belief to at least sometimes have a role in assertion. Moving between assertions of 'P' and 'I believe that P' can work in either direction. So, sometimes, an individual's assertion of P might be a function of that individual believing that he or she believes P, rather than a function of the belief in P. This is certainly a possibility worth

[31] Shoemaker 1996, p. 49.

considering. It would also create possibilities for much more severe differences between George and a normal person.

I do not think that Shoemaker's reductio successfully establishes a problem for non-constitutive views of self-knowledge. Even if they do entail the metaphysical possibility of self-blindness, it seems to me that we can, with the proper understanding of the role of belief and meta-belief in our mental lives, make sense of this idea. Shoemaker's discussion has helped focus our attention on these issues going forward. For now, I return to consideration of his view as a way for self-knowledge to be absolutely groundless.

5. Lessons for Directness

I began discussing Shoemaker's view with an eye toward articulating a notion of directness which could make self-knowledge absolutely groundless. That is, I was looking for an account of how our beliefs about our beliefs are justified, despite having no epistemic ground. Shoemaker's view is that possession of beliefs constitutes knowledge of those beliefs, given rationality and possession of mental concepts. Above, I argued that Shoemaker fails to make the case that self-knowledge is so constituted. However, his failures should not be interpreted as failures of the possibilities of any constitutive account providing the crucial notion of directness.

I believe the successes of Shoemaker's account are here more illuminating than his failures. In attempting to articulate a constitutive account of self-knowledge, Shoemaker is effectively showing us how much we can do without separately arrived at representations of our own mental states. Remember, his goal is to show that any person

with beliefs, rationality, and the right concepts, will behave for all the world as if they know what they believe. Suppose he had succeeded completely. As noted above, the proper response to a successful argument like Shoemaker's is that no one is self blind or we all are. But what, besides stubborn conviction to pre-theoretical intuition, should push us one way rather than the other?

Here, I see a flaw in the structure of constitutive accounts of self-knowledge. In order for us to be legitimately said to know our own beliefs, that knowledge must manifest itself somehow. If everything human beings do, say, and think, were explainable in terms of their first-level beliefs and desires, plus rationality and mental concepts, it would not be as if we all had knowledge of our own mental states. Rather, it would be as if we didn't have any self-knowledge, because that knowledge would be a useless extravagance.

At points in the argument, Shoemaker does make genuine progress in showing how much is capable with beliefs and desire, rationality, and mental concept. For instance, he effectively argues that this would be enough to use language assertively.[32] However, in doing this, he effectively undermines anyone that might have thought self-knowledge was required to make assertions.[33] This, I take it, is the correct position to take with regard to any facet of an argument for a constitutive view of self-knowledge. Once the theorist succeeds in showing that some task, once thought to require self-knowledge, can really be accomplished with some smaller base of abilities and capacities, we should not conclude that self-knowledge is constituted by those abilities and capacities, but that self-knowledge is not required for that task.

[32] Shoemaker 1996, pp. 39-40.
[33] See Mellor 1977.

Shoemaker is, to his credit, one of the few philosophers who have written on self-knowledge who has devoted some consideration to the function and value of self-knowledge. He thinks that self-knowledge is important in communicating one's beliefs and desires to others for the purposes of cooperative action and in the self-critical enterprise of deliberation about what to do or what to believe. I will only discuss deliberation here. Here is Shoemaker on the role of self-knowledge in deliberation:

> Suppose a man confronts evidence which contradicts things he believes, and he makes suitable modifications in his beliefs—he modifies them in the way rationality dictates, or at any rate in one of the ways rationality allows, instead of sticking with an inconsistent or incoherent set of beliefs. Does this show self-knowledge?... I think it plainly does if the modification of belief is the effect of deliberation or reflection… But I wish to argue that even if no explicit rationalization or deliberation went on, such a readjustment of one's beliefs requires self-knowledge, or *at least something very much like it*. The man, *or something in him*, must be sensitive to the fact that there is the inconsistency, and must know what changes in his body of belief would remove it—and this requires knowledge (*or something like it*) of what belief he has.[34]

Since the first time I read this passage, I have been fascinated that Shoemaker adds the qualifications that I have italicized without more seriously considering the possibility that our beliefs can be updated and revised in light of new evidence without any genuine self-knowledge. In discussing what actions or reasoning beliefs prompt (or "rationalize" as he sometimes says), he claims that, "if the beliefs and desires are all first-order states… then one thing they do not rationalize is changes in themselves."[35] His picture of how belief is updated in light of evidence is one in which our, "desires to promote consistency and coherence in the system of beliefs and desires, and beliefs about what changes in the beliefs and desires would be needed in order to satisfy [the desires to promote

[34] Shoemaker 1996, p. 31. Italics added
[35] Shoemaker 1996, p. 33.

consistency and coherence]," require, "beliefs about what the current beliefs and desires are," in order to enact revision[36].

Consider a case of belief revision. Suppose I somehow come to believe that P, that Q, and that P entails not-Q. There is a contradiction among my beliefs, but no immediate prescription for which belief to get rid of. Shoemaker suggests that to resolve the contradiction, we would need a lot of complex mentality. As pointed out above, he says that we would have to want the contradiction to be resolved, have beliefs about which changes would best resolve it, and recognize which beliefs we have to change. Perhaps, though even this is not completely clear to me, this story is plausible for conscious deliberation and belief revision.[37] However, it seems wildly implausible for cases of unconscious belief revision. I see little reason to think that unconscious processes could not be set up to modify the content of beliefs (thus changing the first-order beliefs) without representing those contents as beliefs. Indeed, some of Shoemaker's own comments suggest that he is committed to this as well. I have in mind his brief comments about non-human animals in the same paper in which he sets out his reductio.

He writes, "I am prepared to ascribe beliefs, desires, and intentions to such lower animals as chimpanzees and dogs," but denies that they have self-knowledge on the grounds that they lack the conceptual capacities to represent their own minds.[38] Dogs, according to Shoemaker, can have beliefs but do not have the concept of belief. This is a fairly widely held position. Without the capacity for self-knowledge, however,

---

[36] Shoemaker 1996, p. 33.

[37] To some extent, I think we can get by without meta-belief even in conscious reasoning. I might say to myself, "Q seems true, but P is definitely true and so is P entails not Q. So, Q is false after all." The only word I worry about is 'seems'. Perhaps something like "Q was indicated" would work better.

[38] Shoemaker 1996, p. 31.

Shoemaker's view implies that chimpanzees and dogs should lack the ability to modify their beliefs in light of evidence. But it is preposterous once one has conceded that non-human animals have beliefs to think that their beliefs are not updated in light of evidence. Perhaps there are adequate grounds for denying that non-human animals have mental concepts, but dogs and chimpanzees are, at least as animals go, pretty good problem solvers. Their abilities show not only that they modify their beliefs but that they can, in some sense, reason from their beliefs. If they can do this without self-knowledge, so can we. To summarize, the problem for Shoemaker is that he allows that some non-human animals have first-order beliefs and desires without having self-knowledge, yet fails to consider whether these animals can rationally update their beliefs, an ability which he attributes only to self-knowers.

As mentioned before, the concept of rationality is doing a lot of work for Shoemaker. Perhaps he would concede that non-human animal beliefs are updated as if they were rational, but deny that such updating is actually rational. If dogs and chimpanzees could self-consciously deliberate over their beliefs, making use of self-knowledge, then they could count as rational. This move, aside from being largely a semantic move, glosses over the possibilities for human belief updating. It certainly seems to us that our self-conscious deliberations are involved in the maintenance of our first-order beliefs, but it is not clear how necessary such deliberation is or how much effect it has when it does occur.

We tend to take our ability to do things like revise our beliefs as support for thinking that we have knowledge of what we believe. The previous discussion showed that we should think, even if Shoemaker did not, that belief can be updated and revised

rationally without being represented to the believer as beliefs.[39] I take it that the lesson here is that self-knowledge is not required for belief revision. This represents another task that we might have supposed required self-knowledge being separated from self-knowledge and another step toward thinking that we don't have self-knowledge at all.

Above, I suggested that a completely successful argument for a constitutive account could just as readily be seen as an argument that we lack self-knowledge altogether. Removing the causal efficacy of something theorized to exist always threatens commitments to the existence of that thing. A completely successful argument in the style of Shoemaker's, one that showed a certain set of capacities enabled humans to act exactly as they do now would show us that we were mistaken in thinking that self-knowledge played such a dramatic role in human affairs. It would show us that we lacked it entirely.

However, I think it apparent that no such argument will be given. If in no other way, third personal reasoning about one's own mind demonstrates a human capacity for self-knowledge. Suppose I have, upon observing my behavior in certain situations, come to realize that I have a belief that a certain person is deceitful. If this realization prompts me to avoid that person for fear of a social backlash, my knowledge of my own mental states is prompting me to act. In this case, at least, self-knowledge has something to do.

Of course, this sort of third person reasoning about one's own mental states is supposed to represent an unusual connection to them. It is supposed that the ordinary case is one of more direct connection. However, if we show that what is often taken to be the ordinary case does not require one's knowledge of one's mental states, then we should give up our ideas about how self-knowledge functions in the ordinary case. In assessing

[39] He seemingly rejects this picture, "because it leaves out entirely the role of the person as an agent in deliberation" (Shoemaker 1996, p. 28). The notion of agency is more than can be brought in here. A later chapter specifically addresses the connection between agency and self-knowledge.

the epistemology of self-knowledge, we should look for cases in which our knowledge of our own mental states makes a genuine difference in our behaviors. When we find such a difference, we can then examine the justification of the meta-beliefs involved. This is, methodologically, to rule out the possibility of a constitutive view of self-knowledge. It is, also, to rule out the possibility that self-knowledge is absolutely groundless. With separate meta-beliefs impacting our mental lives and our behaviors, we will invariably look for explanations of the justification of such states. If we found no cognitive process or connection to the lower-order states they represent, we should conclude that such states are unjustified.[40]

6. Conclusion

I argued that notions of directness associated with foundationalist epistemologies are insufficient to display the sort of unique properties attributed to the directness of self-knowledge. Foundationalists attempt to present basic beliefs as justified by positing a relation between the basic beliefs and something else. Whether this other item in the relation is another sort of mental state, such as a perception or sensation, or some external state of affairs, it effectively functions as the epistemic ground of the basic belief. Thus, basic beliefs are not justified indirectly by way of other beliefs, but they are justified indirectly, in a sense. Theorists of self-knowledge seem to want to maintain that there is nothing which could possibly serve as the epistemic ground for our knowledge of our own beliefs.

[40] This raises an interesting question as to whether various anti-naturalist philosophers would be willing to accept that our meta-beliefs are, in fact, unjustified. Some of them flirt dangerously close to this. However, as this possibility is a sharp divergence from my starting assumptions, consideration of this idea will have to wait.

Of the views of self-knowledge in the literature, it seemed that constitutive accounts of self-knowledge stood the best chance of articulating a view in which self-knowledge was absolutely groundless. I examined Sydney Shoemaker's account as a representative. I argued that he failed to make the case that self-blindness, which he claimed was entailed by all non-constitutive accounts of self-knowledge, was impossible. Further, I argued that there is a fundamental flaw in the idea of a constitutive account. Essentially, such an account eliminates the need to think of the phenomenon supposedly accounted for as anything over and above the elements supposedly constituting it. Thus, it seems to me that the directness of self-knowledge is unlikely to be absolute.

Chapter 4

The Epistemic Authority of Self-Knowledge

1. The Challenge of Authority

One reason that self-knowledge is often taken to be special is that there is an apparent contrast in the authority with which we possess it to our authority with respect to other kinds of knowledge. We can make mistakes in domains such as art history or geography, chemistry or mathematics, which seem wholly unlike the mistakes we can make about our own minds. These other subjects, detached and distant, require skillful inquiry and careful scrutiny for genuine knowledge. This detachment means that there is always the possibility of error and that our pronouncements in these domains, even when well grounded, are defeasible. Our self-knowledge, however, can seem indefeasible. From my perspective, what I believe does not seem like some detached and distant fact; it is something I *know* in a way that does not leave the matter open for discussion.

Descartes made his knowledge of the existence of his own mind the foundation of his epistemology because he felt it was so firm. However, the Cartesian view of mind goes well beyond the claim that we cannot doubt our own mind's existence. In the Cartesian view, everything that happens in our minds is transparently conscious and we are infallible with respect to its contents. If this were correct, you would know exactly what you are thinking whenever you were thinking it. Cartesian infallibility depicts us as perfectly aware of every aspect of our mental lives.

However, in contemporary philosophy and psychology, virtually no one thinks that we have infallible access to our minds. It is largely acknowledged that there is a huge amount of mental activity that goes on in each of us of which we are completely unaware. Many philosophers now take intentionality to be the mark of the mental rather than conscious experience. Also, our knowledge of those parts of the mind of which we are typically aware still displays a susceptibility to error.[1] Our emotions, our desires, even our core beliefs can form and change without us realizing it, leaving us confused about how we feel, conflicted in our desires, and wrong about our beliefs. Yet the idea that we have a strong authority over our own minds is a persistent one. Tyler Burge, for instance, asserts, "Descartes held that we know some of our propositional mental events in a direct, authoritative, and not merely empirical manner. I believe that this view is correct."[2] However, Burge and others that continue to hold on to the idea of special authority do not share all of Descartes's outdated view of the mind. The idea that there is a special character to first-person authority has survived the shift away from Cartesian infallibility. Consider, for instance, this claim by Donald Davidson:

> Because we know what we believe (and desire and doubt and intend) without needing or using evidence (even when it is available), our sincere avowals concerning our present states of mind are not subject to the failings of conclusions based on evidence. Thus sincere first-person present-tense claims about thoughts, while neither infallible nor incorrigible, have an authority no second- or third-person claim, or first-person other tense claim, can have.[3]

Davidson takes on some of the commitments of the modern understanding of mind by accepting that we can make mistakes, but holds to the idea that our authority over self-knowledge claims is special.

---

[1] For an overview, see: Wilson and Dunn, 2004

[2] Burge 1988, reprinted in Cassam 1994, p. 65.

[3] Davidson 1987, reprinted in Cassam 1994, p. 44.

As evident from Davidson, this first-person authority is often taken to be tied into some sort of privileged access granted by our first-person perspective on our own mental lives. The apparent special character of this access, its *directness* or lack of inference from evidence, was discussed in the previous two chapters. Davidson represents first-person authority as somehow dependent on this directness. As should be clear from the previous chapters, I do not put much stock in the directness of self-knowledge. However, I will not argue against the special authority of self-knowledge by denying directness. In the next two chapters, I want to approach the apparently special nature of first-person authority from a more general point of view, without presupposing a connection to directness.

I want to tackle the issue of whether it makes any sense to say that we have special authority over self-knowledge claims given the denial of Cartesian infallibility. At issue in this chapter is whether this idea of first-person authority is a vestigial concept holding us back from a full understanding of the mind or an important and unexplained mystery which lies at the center of the questions we must take seriously if we are to fully account for human mentality.

Many thinkers take special first-person authority as a starting point: as something to be preserved and explained in an account of self-knowledge. That is, some philosophers view it as a constraint on a theory of self-knowledge that it account for, or at least remain compatible with, some version of special authority. It can be quite difficult to pin down what is meant by claiming that the authority of self-knowledge is special. This is not surprising, since philosophers discussing this special authority are typically attempting to describe what is special about it. The starting point is supposed to be that

we intuitively see that self-knowledge is especially authoritative in a way that needs to be explained. The project is to explain its special character.

Special authority is sometimes framed in terms of a presumption of truth: that a claim to self-knowledge is presumed to or should be presumed to reflect genuine self-knowledge. Consider, for instance, the formulation of an authority constraint by Akeel Bilgrami:

> It would be natural to call the presumption we've been discussing the presumption of 'authority'. That we have epistemological authority over our states of mind is a good way to describe the idea that it is presumed that when or if we believe that we have a state of mind, say, a desire or belief, we are right. No other person has such epistemological authority over our states of mind… "It is a presumption that, if S believes that she desires (believes) that p, then she desires (believes) that p.[4]

Bilgrami's claim might not suggest the full extent to which he and others consider this authority to be special. For instance, Shoemaker agrees with the idea that authority is about a presumption of truth, but he writes:

> Another claim is that a person has 'special authority' about what such states he or she has… at a minimum it is the claim that it is in some sense necessary that our beliefs about our own mental states of these kinds be for the most part correct, and that a person's belief that she has such a state creates a presumption that she has it, in a sense in which it is not true that someone's having a belief that some *other* person has such a state creates a presumption that the other person does indeed have the state.[5]

Shoemaker is careful to claim that the presumption of truth is different than any other way in which we presume claims to be true. For instance, our ability to know things on the basis of testimony may require a presumption that what others say is true. It is widely held by those that discuss first-person authority that the authority of self-knowledge is different in character. For instance, Crispin Wright states:

[4] Bilgrami 2006, p. 30.
[5] Shoemaker1996, p. 51.

> It is striking that attitudinal avowals would appear to exhibit a form of weak authority nevertheless: that is, they provide criteria—empirically assumptionless—justification for the corresponding third-person claims… It might be suggested that it is nothing other than the presumptive acceptability of testimony generally… Actually, however, I think the suggestion is wrong. What distinguishes the presumptive acceptability of attitudinal avowals from anything characteristic of testimony generally is that the authority which attaches to them is, in a certain sense, *inalienable*.[6]

Wright means to say that, while there may be a presumption of truth assigned to a claim anyone makes about any subject, most such claims only go as far as those, "subject-matters which an informant is deemed competent to know about".[7] We know exactly how and when it is appropriate to challenge most third-person testimony. Not so for apparent self-knowledge claims. Underlying the idea that there is something special about this presumption that self-knowledge claims are true is the thought that they are inappropriate subject matter for challenge.

To help get a foothold on the idea of this apparent special authority, imagine that you were to stumble upon a friend, apparently lost in thought, in some public place. You offer this friend the proverbial penny to find out what he or she is thinking. Suppose your friend says, "I am thinking about whether or not human beings have free will." Depending on what kind of friends you have, this may be very much in or out of character. In this situation, you have a range of responses that seem reasonable. You might leave your friend alone to think or sit down to debate the issues of free will together. However, one thing it seems inappropriate to do is to argue about whether this really is his or her topic of thought. Imagine saying to your friend, "No you weren't! You were thinking about the Spanish Inquisition!" Absent strong evidence that your friend is

[6] Wright 1998, p. 17.
[7] Wright 1998, p. 17.

lying, it seems inappropriate to challenge your friend on this point.[8] When anyone says, "I am thinking about *X*", we just simply accept that *X* is what they were thinking about. The default position we take others to have toward knowledge of their own minds is one of authority. This becomes clearer when you reverse the roles in the thought experiment. Imagine your friend challenging you about the contents of your thoughts. Why should you care what opinions an outsider has toward the insides of your mind? If I were so challenged, I might say to the challenger, "Who are *you* to say what's going on inside *my* head?"[9]

This point about the oddity or inappropriateness of challenging self-knowledge claims is made often in the self-knowledge literature. For instance, Dorit Bar-On writes:

> …it would be highly irregular to request of someone issuing an avowal to offer reasons for her pronouncement, or to criticize her for lacking an adequate basis or justification. And, except under very special circumstances, we do not expect the sincere author of an avowal to stand corrected.[10]

And Wright asserts:

[8]There are contexts in which we can imagine evidence that a friend is lying in a case like this. Perhaps you know that your friend has been researching the Spanish Inquisition rather than performing some other duties and is embarrassed to have been caught by you in the library whilst surrounded by history books. The important cases are ones in which your friend seems sincere and there is no preponderance of evidence against his or her claim. In those cases, his or her proclamation might seem authoritative.

[9] I should note, especially to those who might be sympathetic to my ultimate stance, that I have become increasingly less convinced that this is a good question. At one time, it seemed to me a very good question, but the more I think about everyday life, the more I think challenges to people's claims to believe this or that are relatively commonplace and often appropriate. I might say to a friend, "You don't really believe that," because the friend has acted in the past in a way that is incompatible with the belief he or she has just self-ascribed, because the friend is usually reasonable and self-ascribed belief seems unreasonable, or because he or she has told me differently some other time. These sorts of cases, at least sometimes, seem genuine and appropriate challenges to a person's knowledge of his or her own beliefs rather than challenges to the truth of those beliefs. That is, sometimes when I say to a person, "You don't really believe the moon landing was faked," I am actually challenging the idea that he or she believes the moon landing was faked and not merely arguing we did land on the moon. However, I am willing to grant the presumption of authority in order to examine it.

[10] Bar-On 2004, p. 3.

> The demand that somebody produce reasons or corroborating evidence for such a claim about themselves… is always inappropriate. There is nothing they might reasonably be expected to say.[11]

Elizabeth Fricker agrees that, "Self-Ascriptions are 'authoritative' in the sense that they are ordinarily *treated* as such," and contrasts self-ascriptions with third-person ascriptions which are "treated as neither authoritative nor basic." She claims of the third-person case that it is reasonable to ask questions, "and we may be skeptical about [the] claim while not doubting [the maker of the claim's] sincerity."[12] However, Fricker explicitly frames this authoritative treatment in terms of a language-game. Without here delving into the exegesis of Wittgenstein that this phrase demands, I take her point to be that granting that we talk as if there is a presumption of authority does not immediately require granting any particular view about our actual knowledge or authority. She writes:

> This idea of Special Access as a real phenomenon of reliable tracking underlying the language-game surrounding Self-Ascriptions (i.e. everyday epistemic practice in response to them), far from being a datum, is a controversial theory of what underlies the 'grammatical' phenomena.[13]

This is an important point. There is logical space between inappropriateness of challenge and a metaphysical presumption of truth. In granting that there is an apparently special first-person authority, I take the phenomenon to be that apparent claims to self-knowledge seem inappropriate things to challenge, where apparent claims to self knowledge are understood as utterances of sentences in the first-person that apparently ascribe mental states to the person making the utterance. This weaker claim should still be granted by anyone who wants to frame authority in terms of truth. By framing authority in this way, though, we are able to consider the possibilities that apparent claims to self-knowledge do not reflect meta-beliefs, are not interpreted as assertions, or

[11] Wright 1998, p. 14.
[12] Fricker 1998, pp. 157-158.
[13] Fricker 1998, pp. 158-159.

are not appropriately challenged due to non-epistemic norms. I am not here denying that the presumption of truth is real or that it is guiding the feeling that challenging apparent self-knowledge claims is inappropriate. I merely want to keep as many options on the table as possible.

I understand authority of any kind to be a normative notion. In particular, I consider person A to have authority over (or to be an authority with respect to) person B when it is inappropriate, according to some norm or other, that B challenges A. If it is inappropriate for Charlie to challenge his mother about when he must go to bed, Charlie's mother has some kind of authority over Charlie (or is some kind of authority with respect to Charlie). Parental authority is, I think, not particularly relevant to epistemology. No doubt there are some kinds of authority that will be more interesting than others to philosophers. At issue here is the appropriateness of challenging people's claims to knowledge of a very particular subject matter: their own minds. This sort of authority, if it exists, ought to be of some philosophical interest.

If the proponents of a special first-person authority are right and we are constrained by a notion of special authority in theorizing about self-knowledge, the naturalized epistemologist has a *prima facie* problem. For the naturalist, it is seems to be an open question whether or not we are even very good at knowing what is going on inside our heads. On the naturalist picture, a presumption that meta-beliefs are true would seem at best unwarranted (until the empirical evidence is in) and at worst at odds with the very idea that the methods of science could uncover failures of self-knowledge.[14] The weaker version of the authority presumption that I am using presents a similar problem.

[14] I am lumping together, to a certain degree, thinkers that simply reject perceptual models of self-knowledge (Shoemaker, Bar-On) with thinkers that are more openly anti-naturalist (Bilgrami, Moran). This is difficult not to do as perceptual models of self-knowledge are often equated with naturalized models.

Why should we think claims to self-knowledge are above challenge before we know how good people are at knowing their minds? For that matter, why should we think such claims are inappropriate to challenge given the empirical possibility that people are not even very good at knowing their minds?

The dialectic here is messy. Naturalists do not typically think that philosophical issues about the mind are to be settled purely on conceptual or *a priori* grounds. Just how good we are at knowing our own minds should be something we can discover empirically; any method of knowing anything should be subject to external empirical scrutiny. The naturalist is unlikely to accept *a priori* or conceptual arguments that rule out what he or she takes to be empirical possibilities. However, it remains an interesting issue what kind of authority a naturalistic model of self-knowledge might countenance without accepting a conceptual constraint. There is, after all, a persistent idea that there is something special about the first-person perspective. It would be nice to see how special this perspective can be without creating problems for naturalism.

Of course, there are different ways to be an authority, even with respect to claims of knowledge. Different sets of norms create different senses in which a challenge can be considered inappropriate. One way to be an authority is to be in a better position to know something than another person, perhaps because of expertise or evidence. Call this *epistemic authority*. The point is that, if someone knows better than you do, it is, in some sense, inappropriate for you to challenge that person. There are, of course, other ways to be an authority. For example, one might count as an authority over another due to an established relationship in which it is conventionally—or morally—inappropriate for the second to challenge the first. I will simply call all kinds of authority other than epistemic

authority, *non-epistemic authority*. I recognize that this implies unity among kinds of authority that will be importantly different, but, for this chapter, all the matters is that epistemic authority have some degree of unity. Specifically, all kinds of epistemic authority must involve an inappropriateness to challenge of a knowledge claim due to some epistemic norms.

The notion of non-epistemic authority evokes the sorts of relationships that exist between parents and their children, children and their elders, students and their teachers, and so on. Non-epistemic authority can exist in any facet of interaction between two parties or groups. Authority based on convention can even cover inappropriateness of challenges to claims of knowledge. We often have non-epistemic reasons to disagree or not to disagree with someone. I will return to this idea later.

In this chapter, I consider only epistemic authority. I argue that there is no reason to think that there is a special epistemic authority attached to self-knowledge. By 'special', I mean a kind of authority that would somehow conflict with naturalism about the mind or a naturalized epistemology. Marking this distinction between epistemic and non-epistemic kinds of authority allows for several possibilities to become clear. For instance, one might think that there is a special authority associated with the first-person perspective that is non-epistemic in nature and challenges naturalism about the mind. For example, if facts about our mental states were somehow determined by conventions in which we take each other to be authorities about our own minds, then naturalistic theories of the mind would be in trouble. The idea that minds could change with changing conventions is surely at odd with the naturalist's picture of the mind, but it does not immediately create a special notion of epistemic authority. Perhaps self-knowledge under

such a view would be simply a person's application of conventional wisdom to himself or herself. Another possibility might be that our way of self-knowing itself is unnatural. For instance, if introspection was the result of an inner sense which was sensitive to divine illumination of our souls, naturalism about the mind would be out the window. However, there would still be a further step required to show that this divine introspection created any special epistemic authority. Perhaps divine introspection could be fallible and subject to critique from external evidence.

One way for first-person authority to conflict with naturalism is for it to stem from some non-natural feature of the mind. I will return to this possibility in the next chapter. In this chapter, I will consider a different possibility for the conflict between authority and naturalism. In the comments above, it sometimes appeared as though the claim was that the naturalist's model of first-person authority was fundamentally misguided because it did not fully articulate the epistemic authority we have over our own minds. That is, one potential criticism of the naturalist's view of first-person authority is that it fails to appreciate the conflict between naturalism and the epistemic authority of the first-person perspective. What the naturalist cannot tolerate is a kind of epistemic authority that places self-knowledge outside the realm of correction and critique from external sources.

If we take the claim of special authority for self-knowledge to be an epistemic authority claim, there is a sense in which naturalistic accounts make our epistemic authority out to be rather mundane. My contention is that once fallibility is admitted, any account of the epistemic authority of self-knowledge will be mundane in just this sort of way. Essentially, by admitting the fallibility of self-knowledge, contemporary theorists

place self-knowledge claims in the same realm as other kinds of knowledge. In order to establish this, I will first look at the mundane authority associated with a naturalistic model of self-knowledge.

2. Perceptual Models of Self-Knowledge

Before we can understand why the authority of self-knowledge is seen by some philosophers as so important, we must see the way in which some accounts of self-knowledge are seen as minimizing or ignoring it. There are certain common features to the majority of models of self-knowledge considered by psychologists, cognitive scientists, and naturalist leaning philosophers. These general features enable us to group somewhat disparate models together as *perceptual* models of self-knowledge. My use of the term 'perceptual model' follows Richard Moran.[15] It is suggestive of inner sense models of self-knowledge. While anyone defending an inner sense model should be considered a perceptual model theorist, the term is actually meant to be a bit broader. The unifying idea of perceptual models is that some kind of detection process is directed toward our own minds. While one way to envision a detection process is in terms of perception, not all of these models make heavy use of the perception metaphor. What is common to all of them is a certain causal picture of how we come to believe that we have certain beliefs. The causal picture, at its most general, is that via some mechanism or other, our first-order beliefs cause us to have second-order beliefs about those first-order beliefs.

[15] Moran uses the term throughout his book. See Moran 2001.

This categorization allows us to group together almost all naturalistic models of self-knowledge. Even though some views involve a mechanism that relies on theoretical information, it still depicts self-knowledge as detection of existing states.[16] As long as a view tells a causal story about the formation of meta-beliefs in response to the existence of first-order states, it counts as a perceptual model as I understand them. This means even a view which posited that we individually rely on observation of our own behavior—such as Gilbert Ryle's view—could be counted[17]. Granted, using the term 'perceptual' is a little strained for some naturalistic accounts, but you might think of these sorts of views as being *outer* sense theories of self-knowledge, since they are based on external data or observation (plus theory). Indeed, given the prevalence of perceptual talk with regard to other minds, this sort of view should probably be given more consideration than it usually is. Ultimately, I will defend a view that has more in common with outer sense views than with inner-sense views. The larger point for the moment is that, however real the debates between perceptual model theorists, they can be seen as engaged in something of a family dispute. Disagreements between theorists proposing perceptual models are disagreements about the causal processes involved in self-knowledge. Some think that information from outside is involved. Some think there is some amount of theoretical information employed. Some think that mental states are detected via an internal process. What all of these accounts agree on is a picture of

[16] See Gopnick 1993.

[17] It is not clear that Moran would go this far, but I think he should given the way he typically uses the term. I think you could only exclude such inferential accounts by stipulating what features of the existing states the "perceptual" process can take as inputs. Behaviors might be unreliable inputs, but they do seem to me to be a way of detecting the existence of the mental states. It is something like crediting the inhabitants of Plato's Cave with seeing the world on the basis of seeing the shadows.

existing lower-order mental states and a mental process generating higher-order mental states in relation to the lower order states.

It is in conceiving of self-knowledge as a process of detection that some philosophers take issue with naturalistic accounts of self-knowledge. This is apparent in regards to the topic currently under discussion. The idea that we are special authorities over our own minds is seen as incompatible with naturalism mostly because it is thought to be incompatible with perceptual models of self-knowledge. I do not mean to suggest that there cannot be a naturalistic model of self-knowledge that is not a perceptual model, but virtually all models of self-knowledge put forward by scientists and scientifically minded philosophers share as a common element a view of self-knowledge as a process of detection of existing states.

Perceptual model theories have come under fire in recent years for their supposed inability to explain special authority. Richard Moran, for instance, claims that a perceptual model, "presents an essentially superficial view of the differences between my relation to myself and my possible relations to others… as a faculty that happens to be aimed in one direction rather than another."[18] This idea is put forth very strongly by Crispin Wright:

> …there can be no scaled-down observational model of self-knowledge which preserves the advantages of the Cartesian account while avoiding its unaffordable costs. The problem is that the kind of authority I have over the avowable aspects of my mental life is not transferable to others: there is no contingency… whose suspension would put other ordinary people in a position to avow away on my behalf, as it were.[19]

The idea is that, in presenting self-knowledge as just a different kind of perception-like mechanism, one puts it on a par with the other mental mechanisms that impact our ability

[18] Moran 2001, p. 91.
[19] Wright 1998, p. 24.

to know things. A perception like mechanism is simply one way of knowing among many. It could be subject to error and, in principle, corrected by external evidence. But, if this is true of self-knowledge, then it is not special; it is just like every other kind of knowledge.

The perceptual model theorist posits a mechanism for acquiring self-knowledge and any mechanism in an imperfect being is likely to be imperfect. Those endorsing perceptual models will likely view it as an empirical question just how good the mechanism is. However, the controversy is not over the fact that perceptual models posit *imperfect* mechanisms. Nearly every theorist has given up the idea that we have perfect access to our own minds. Rather, the controversy is over positing a mechanism for self-knowledge at all. The threat to authority is over the idea that there could be a way to know our own minds more reliably or accurately than we do now. In other words, positing a mechanism suggests that there might be a better mechanism, perhaps a mind-reading device straight out of science fiction. A further worry for some thinkers is that a better mechanism might be trusted over internal mechanisms. If I said that I believed one thing and the machine said I believed another, it would seem perfectly appropriate to challenge my assertion. The possibility of a mind-reading machine does seem to have implications for special authority.

Let us consider the possibility that a scientist might one day be able to build a machine that can read minds. Perhaps that day is not so far off.[20] If such a machine were merely pretty accurate, its existence would not jeopardize the idea that we are individually authoritative about what we believe. To really get a conflict with the practice

[20] See Borel 2009.

of treating ourselves as authorities, we would need a machine that is more reliable than the individual it scans. In conceiving of self-knowledge as an imperfect detection process, perceptual models strongly suggest the possibility of such a machine, whether or not we can ever make it an actual reality. Call this the Mind Reader Hypothesis.

> MRH (Mind Reader Hypothesis): It is in principle possible, metaphysically if not physically, to have a device that can detect a person's mental states in a way that is more accurate than that person's own meta-beliefs.

Let us grant that the perceptual model theorist is committed to MRH. It would seem that, given this commitment, we can have authority with respect to our own minds only insofar as the internal mechanism that we have for detecting our mental states is the best one available. If there were a super accurate mind-reading machine scanning your mind then it would seem reasonable for you to defer to the more accurate machine whenever there was initial disagreement. It would suddenly be appropriate for other people to say to you, in some circumstances, "You aren't thinking about that at all!" and to similarly challenge all sorts of first-person mental reports. If we are more reliable and accurate about the contents of our own minds than anyone or anything else, it is, under MRH, a contingent matter because we just happen to have the best available way of knowing. Accepting MRH, one might think, reduces "special" authority to something we only have right now because no one has built a machine like this.

With respect to the idea of special authority in particular, here is an argument against perceptual models that may be lurking in the background.

1) Perceptual models are committed to MRH.
2) If MRH is true, we do not have any special authority over our own minds (our intuitions of special authority are unjustified).
3) We do have a special authority.
4) Therefore, perceptual models cannot describe self-knowledge.

Perhaps no one would accept the argument in this form. The central idea however, that a commitment to MRH undermines our claim to a special authority, is a very important one. What the mind-reader would do is provide an independent check on our mental self-ascriptions. With such an independent check, any presumption that self-ascriptions are true begins to seem unwarranted. That the presumption could be defeated would show that our current presumptions of authority are based on something contingent rather than on something necessary to the first-person perspective. An anti-naturalist might argue on the basis of our strong intuition that there is something fundamentally special about first-person authority and that we ought to reject the MRH scenario as impossible.

Perhaps we have reached an impasse. Perceptual model theorists will not be able to accept any version of authority that rules out conceptually what they take to be the empirical possibility of a more accurate and reliable mechanism. Those that favor a special authority constraint on an account of self-knowledge have an objection to perceptual model theories of self-knowledge, relying on conceptual intuition, which naturalists are not inclined to accept when they run counter to otherwise good theories.

As I see it, this apparent impasse can be avoided because this anti-naturalist argument is fundamentally flawed right at the point of its central idea. The possibility of a more accurate mind reader device is not uniquely connected to perceptual models. Rather, I will argue, commitment to MRH depends only on assumptions that (virtually) all theorists of self-knowledge share. In light of this, MRH is relatively innocuous and should be accepted by nearly everyone. Anyone wishing to provide an account of our special authority had better be careful regarding ruling out MRH in an *a priori* manner. I think that there are quite a few ways one might grant MRH and yet hold on to the idea

that there is something special about first-person authority, but a view that seeks to rule out MRH is fundamentally misguided.

The interesting thing about perceptual models is not that they are committed to the possibility of a mind-reading device. Rather, it is that they posit the actual existence of a mind-reading device and claim it exists in each of our brains. The internal scanning mechanism posited by inner-sense theories, however it is supposed to work, is essentially an internal mind-reading device. The idea that there might be a better device follows naturally from accepting that there is an existing and fallible device; it is always possible to build a better mousetrap. However, MRH does not depend on thinking of self-knowledge as detection, or even on thinking of self-knowledge as a process. As long as one concedes that we are not infallible with respect to self-knowledge, it is open in principle that there could be a mechanism that was more accurate than we are about our own beliefs. This point, once made, might seem fairly obvious, but it is worth dwelling on for the sake of clarity and to help us understand where the confusion over this point lies.

All that is really needed is that there be a fact of the matter what we believe and that our beliefs about our own beliefs are sometimes incorrect. It does depend on thinking of self-knowledge as a matter of having meta-beliefs, but it does not matter how or why those meta-beliefs are present. If they exist and sometimes misrepresent the facts, there exists a possibility for making fewer errors, both in number and in kind. One need not even imagine an infallible detector. All that MRH entails is a detector that represents our minds more accurately than our meta-beliefs typically do. Think of your meta-beliefs and the scanner as providing two different maps of your beliefs. You can compare two maps

for accuracy without knowing anything at all about how they were made. Everyone concedes that our internal maps are fallible, thus creating the possibility of there being a better map.

There are some worries about the possibility of making such a device. One way to be worried is to conceive of the mind or mental states in a non-naturalist way. If beliefs do not have features or properties that are physically detectable, then it is less than likely that we could build a physical detector. As will become clear later in the dissertation, I think some have problems with naturalistic models of self-knowledge for precisely this reason. Of course, if beliefs are real and have physical effects, then they must be at least indirectly detectable. So even on non-naturalistic views of belief, there is some reason to think that a mind reading device is physically possible. Of course, all that is required is that the mind reader be metaphysically possible, and this is difficult to threaten, even with extreme anti-naturalist positions.

Another worry is that we seem to lack an independent test of any mind reading device. The idea is that you cannot have a device that is more accurate at reading an individual's mind than that individual because you would have to rely on his or her reports to know whether or not the device was right. Otherwise, we would need to know the facts perfectly or at least more accurately than either of the two processes and we have no such access to the facts. This is confused. It may be, and I accept this point as a practical concern, that it would be difficult to calibrate a scanner and test it against an individual. However, all this shows is that, if we had a scanner and an individual's reports, we would not be in a good enough position to know which was right more often without additional information. That the mind reader *might* be more accurate is

guaranteed by the fact, accepted by everyone in the debate, that we are sometimes wrong. Whether or not we can know how good the reader is remains an open question.[21]

In light of this reasoning, it seems that everyone had better learn to live with at least the metaphysical possibility of a very good mind reading device. This does not win the day for perceptual models, but does help make them prima facie more plausible to those that believe we have a special authority. If everyone must accept MRH, then any special authority must be compatible with it and may be compatible with perceptual models as well.

I think there are several questions worth attempting to answer. First, what kind of authority might satisfy our intuitions without ruling out MRH? Is there a notion of epistemic authority that can do the trick? This will be the guiding question for the remainder of this chapter. However, we must also consider whether there is a non-epistemic notion of authority that can capture at least some of our intuitions about the first-person perspective. In the next chapter, I will consider various notions of authority explored by philosophers in recent times. I will argue that each offers a notion of authority that is non-epistemic. However, there will remain a worry about whether or not these non-epistemic notions of authority might themselves cause problems for naturalism about the mind, even if they do not cause problems for a naturalized epistemology.

3. Epistemic Authority

As we have seen, many philosophers think that there is something inappropriate about asking for support for mental self-ascriptions. However, our fallibility in the

[21] And one that seems more promising to me than it might to some. We seem to accept already that there can be, in at least some cases, adequate evidence contrary to people's reports for concluding that their reports are wrong. Why, then, not think we could use such additional information to help build and test a mind reader?

domain of self-knowledge raises genuine epistemic questions. For instance, how accurate are our self-directed meta-beliefs? How often are we wrong and are there patterns to the errors? How are these beliefs formed and what external factors influence their formation? Even if there are times that it is inappropriate to ask these questions, the epistemologist cannot ignore them for fear of violating conventions of normal discussion. The philosopher must sometimes ask the questions no one else wants to ask.

To be clear, I do not want to dismiss critics of naturalism as putting etiquette ahead of epistemology. Rather, what I am suggesting is that no notion of authority grounded in non-epistemic norms can be taken as decisive between differing theories of the epistemology of self-knowledge. To the extent that the epistemologist considers one of these notions of authority as a constraint on an epistemic theory of self-knowledge, he or she is putting something, even if not etiquette, ahead of epistemology. Some philosophers, no doubt, think that the inappropriateness to challenge is epistemic. If questioning a person's mental self-ascriptions were nonsensical, because that person could never be wrong or because we could never know that person to be wrong, it would be easy to see how there might be epistemic norms behind the inappropriateness. However, since the shared position of almost all theorists is not only that people can sometimes be wrong, but that we can sometimes know that they are wrong, it becomes difficult to see how any epistemic norm can underlie the inappropriateness in question.

Interestingly, I think there is much potential for naturalized epistemologists to make use of non-epistemic accounts of first-person authority. Roughly, the strategy would be to claim that our self-knowledge is on par with all other kinds of beliefs—defeasible, in principle knowable through a variety of processes, open to the normal

range of epistemic errors, etc—but admit that treating it as such violates some interesting normative practice. The intuitive authority of self-knowledge, then, can be accommodated without an epistemic notion of authority. I will sketch this sort of strategy in slightly more detail in the next chapter.

For now, I want to restrain the topic of discussion to epistemic norms. That is, let us restrict our discussion of authority to epistemic authority. The idea of epistemic authority is that it is possessed only when epistemic norms make it inappropriate to issue a challenge. If we take the challenge to be over the truth of some fact, then epistemic authority must be authority grounded in the fact that one person is in a better position to know than another. Unless the proponent of a special authority of self-knowledge is articulating a kind of epistemic authority, the naturalized epistemologist has nothing to worry about. Naturalistic models of self-knowledge should be compatible with non-epistemic conventions, good or bad, against challenging self-ascriptions.

The fact that some philosophers view first-person authority as a presumption of the truth of meta-beliefs suggests that they do think that epistemic norms create the inappropriateness of challenge. However, it cannot simply be the presumption of truth that creates a special authority. Without conflicting evidence, we tend to presume anything that anyone says is true. Yet, most claims are perfectly fine, or so we think, to challenge when we have good reason. A naturalized epistemology, then, is compatible with a presumption of truth in favor of testimony, for instance, precisely because this presumption can be over turned with the right sort of evidence. If there is a special epistemic authority for self-knowledge, it must be explained why challenge is

inappropriate even when the challenger has apparently good reasons or, alternatively, why there are no genuinely good reasons.

In the remainder of this section, I will consider whether or not a version of epistemic authority might be found that could serve as a basis for rejecting a naturalized epistemology of self-knowledge. Providing any account of epistemic authority will involve filling in the idea of being in a "better position to know". As a naturalized epistemologist, the most natural way for me to understand the idea of being in a better position to know is in terms of reliability, either of the knower or the method of belief formation. I have argued that, no matter what your account of the way that we possess self-knowledge, accepting fallibility means accepting the possibility of a really good mindreading device. It is thus extremely difficult to entertain the notion that our self-knowledge represents any kind of special epistemic authority if "knowing better" is cashed out in terms of reliability. Perceptual accounts of self-knowledge are open to the empirical possibility that we are or are not even very reliable in our meta-beliefs. Of course, even if it is not incredibly reliable, our internal detection processes may be the most reliable currently available way of knowing. It would be comparable to a situation in which an eye witness to a crime is debating someone that heard the details from a friend. Even though we can have reasonable doubts as to the accuracy of the eye witness, he or she seems to know better than most, most of the time. Certainly an eye witness is typically in a better position to know than someone who heard from an eye witness.

Perhaps a better parallel would be to our knowledge of germs gained from an electron microscope; it is the best we can do now, and might be, as a matter of fact, the best we can ever do, but there will always remain a clear sense in which we are not in the

best possible position to know what germs are like. Having the exclusive access to the best available way of knowing might grant us a very real kind of epistemic authority, but it is hardly unique and it is not at all incompatible with perceptual models. What remains for anyone that thinks a naturalized account of self-knowledge is getting something fundamentally wrong about our authority over our own minds is to articulate an account of epistemic authority that does not rely on reliability and demonstrates that our special relationship to our self-knowledge grants us, in some sense, a better way of knowing.

I can see two ways that someone might attempt to cash out a notion of epistemic authority that does not make reference to reliability. I do not claim that these two ways are exhaustive. Rather, they strike me as directions that philosophers hostile to naturalized accounts of self-knowledge would be most likely to go. If there are other ways to formulate accounts of epistemic authority that I have not considered, I can only hope that this discussion forces the thinkers that articulate them to be clear about how they differ from the notions that I do consider while also attempting to show that they constitute what might genuinely be a special epistemic authority.

Consider the following example to help understand two alternative ways in which we might understand epistemic authority. Suppose John is an eager young college student and Clara is a tenured professor of chemistry. One day, Clara says in John's presence, "Hydrogen is the first element on the periodic table". It would be inappropriate, perhaps in several different ways, for John to challenge this claim. Absent any other details to the story, it seems obvious that Clara is in a better position than John to know the layout of the periodic table. In terms of reliability, for example, Clara would likely be much more reliable about the periodic table. Since we know that that will not do for a notion of

epistemic authority, we should consider other reasons for thinking that Clara knows better than John.

One tempting line of thought is that, whatever John's reasons for disagreeing with Clara, Clara will have better reasons for her belief. Clara, as a tenured professor of Chemistry, has a rich network of well justified beliefs and experiences grounding her claim. Perhaps she has even been working with the periodic table for longer than John has been alive. She has conducted hundreds of experiments, all of which depend in some way on the periodic table being accurate and many of which support her various beliefs about it. No matter what John has seen or done to make him think otherwise, his reasons cannot compare to Clara's. Perhaps we can understand the epistemic authority of self-knowledge in terms of having, as a matter of our relationship to the subject matter, inherently better reasons for our meta-beliefs than anyone else might have.

This is a tempting, and in some ways promising, line for the anti-naturalist to take. It fits well with the idea, for instance, that we would have no way to show a mind reading device to be more accurate than our own judgments. That is, if your internal reasons for forming a meta-belief are inherently better than the evidence the scanner is sensitive to, you might have a reason to trust yourself over the scanner even if it is in fact more reliable. However, dialectically, this line of thought is completely cut off from most anti-naturalists. In addition to self-knowledge being thought to be special because of our special authority, it is typically claimed by anti-naturalists that self-knowledge is held groundlessly or directly. That is, rather than claiming our self-knowledge is based on especially good reasons, anti-naturalists tend to think that self-knowledge is not based on reasons at all. If self-knowledge really is something possessed directly, in a way that is

incompatible with a naturalized epistemology, perhaps the anti-naturalist will have won the day. However, articulating an account of direct self-knowledge undermines the prospects for thinking we have a special authority even further. If this groundless self-knowledge is also especially authoritative in an epistemic sense, a separate account of this authority must still be provided and it cannot be in terms of having good reasons.

If, instead, the anti-naturalist would give up the idea that self-knowledge is direct and attempt to articulate the reasons we form our meta-beliefs, there would be an interesting question of whether those reasons might be in some way better than other kinds of reasons. However, it seems to me such a move would be ultimately unsuccessful. The problem is that, however we form our meta-beliefs, we already think that others can sometimes have reason enough to think our self-directed meta-beliefs are wrong. This indicates to me that, whatever our reasons for our self-directed meta-beliefs, on the assumption that we have reasons, they must be the sorts of reasons that fit into the normal epistemic scheme of reasons, capable of being outweighed by enough contrary reasons. Even if these internal reasons are especially good, they are not of a fundamentally different kind. They are certainly not good enough to disregard a mindreading device if they are not good enough to regard a friend's observation. A perceptual model theorist, while perhaps not talking of reasons, can say much the same: that our self-knowledge is pretty accurate, but there might be more accurate ways to know or considerations outweighing our normally accurate method from time to time.

The other possibility I see for an alternative version of epistemic authority is in terms of expertise. Going back to the classroom example, Clara is an expert on chemistry. Any student of chemistry should, on some level, defer to her. Clara, after all, is teaching

the class. If it is appropriate for the students to doubt and challenge everything she says, one might think, the students would have no reason to listen to her. You have to have some trust in and limited deference toward experts. Perhaps if we think of each person as an expert on his or her own mind, we will see that no one other than you can truly become an expert on your mind. Any other person trying to learn what you are thinking is at best a student and should defer to you, the teacher.

Ironically, this idea shares much with the ultra-naturalistic view of self-knowledge proposed by Gilbert Ryle.[22] He held that a person knows better than any other what he or she is thinking by exposure. To put it blithely, we are around ourselves very often and we have a lot of opportunity to gather data. We can scarcely help but become experts on our own minds. This view has been attacked from all sides for not taking seriously the differences between the first and third person perspectives. These days, even naturalists tend to think that what goes on inside the mind is relevant to self-knowledge. Perhaps, though, Ryle was right to think of us as *de facto* experts, even if our expertise is due in part to our privileged point of observation and the internal things observed rather than just due to making more of the same kinds of observations.

To understand the epistemic authority of experts, we must separate it from their non-epistemic authority. Complicated relationships can exist between sets of norms. Imagine a world in which certain conventions about who can challenge the assertions of whom are written into law. Suppose in this world, there was a king and that it was illegal for anyone to challenge the word of the king. Such a king would have a great deal of authority, but clearly the king's authority would be non-epistemic authority. Even when

[22] See Ryle 1949.

the king's proclamations are about facts or are claims to knowledge, the inappropriateness of challenge is due to political and legal conventions.

Kings are not necessarily in a better position to know certain facts. If the king says X is true, people may have, quite obviously, reasons against challenging X. There might even be certain kind of reasons to believe X merely given the king's assertion, but such reasons would be non-epistemic in character. There would be, for instance, pragmatic reasons; if your beliefs are in line with the kings, you are much less likely to run afoul of the king. The important thing to note is that nothing about the fact that the king, at least *qua* king, thinks that something is true makes it more or less likely to be true. The king is not necessarily in a better position to know anything than anyone else simply in virtue of being king.

This is not to say that epistemic and non-epistemic authority cannot overlap. Add one detail to the story above and the overlap between non-epistemic and epistemic authority becomes more complex. Suppose that the way one becomes king is to get the highest score on a very comprehensive test of knowledge in a variety of subjects. Knowing this, the king can be thought of as a very reliable knower, perhaps even an expert in many subjects. There would be genuine epistemic reasons not to challenge at least some of the things he says. Of course, the laws against challenging him make these epistemic reasons somewhat redundant in the eyes of the subjects, but good sense can be made of the distinction between when it is appropriate to challenge the king in a legal sense (never) and when it is appropriate in an epistemic sense to challenge the king.

A similar multifaceted dimension of authority may exist between experts and non-experts in our society. In addition to the epistemic authority of experts, it is sometimes

also, out of deference or respect, inappropriate to challenge an expert. However, if such conventions are set aside, it becomes immediately clear when it is epistemically appropriate to challenge an expert. All you really need is a good reason for your contrary belief. Experts, being experts, often have better reasons for their beliefs than non-experts, but the risk in challenging them is only in being embarrassed for having presumed to know better than someone who knows very well.

If I am right about there being implicit conventions against challenging experts, we can even accept that such conventions have some epistemic merit. All other things being equal, people will probably have more true beliefs if they defer to the experts. However, we know, at the same time, that expertise is defeasible. Whatever the value of attaching some non-epistemic authority to expertise, we cannot epistemically afford to take experts as indefeasible. It is precisely when we have good reasons for our beliefs that it becomes epistemically appropriate to challenge the experts. And, since we can sometimes have good evidence for the beliefs of others, it should be epistemically appropriate to challenge people about their own beliefs with good conflicting evidence. Since all of this would be readily allowed by a perceptual model of self-knowledge, expertise is an unlikely basis for a special epistemic authority of self-knowledge.

4. Non-Epistemic Authority

Most philosophers writing on first-person authority have presumed that if we have special authority, it must be epistemic. However, there is logical space to think about first-person authority as non-epistemic. That is, first-person authority might be grounded in some set of norms outside of epistemology. Some writers seem to be moving in this

direction. In the next chapter, I will discuss the work of Richard Moran and Akeel Bilgrami. While neither explicitly gives up an epistemic component to authority, both make explicit reference to non-epistemic norms surrounding self-knowledge claims. I argued above that the prospects for a special epistemic authority of self-knowledge are not good. Of course, my suggestions for how to understand epistemic authority were not exhaustive. It remains an open possibility that someone will articulate a sense in which the first-person perspective is associated with a kind of epistemic authority incompatible with naturalism. However, in his section, I want to consider the possibility that we might do better in accounting for the apparent special features of the authority of the first-person by positing a kind of non-epistemic authority.

As a first pass at understanding non-epistemic authority, think of the myriad of social conventions that govern our daily interactions with each other. Without much conscious thought, most people do or avoid doing all sorts of things in accordance with what they deem appropriate. For example a friend of mine once refused to take off his shoes when entering a party held by an acquaintance. While certainly not universally done, many party hosts require guests to take off their shoes. My friend's refusal made me suddenly aware of some kind of authority which I had never thought too much about. I, and every other party guest, took it for granted that the host, inviting us into his or her home, had the authority to ask us to remove our shoes. When my friend refused, his challenge to that authority seemed inappropriate. The awkward moments of silence following this exchange indicated to me that I was not the only one suddenly aware of this hidden authority dynamic.

Similar commonplace examples of authority might also count as non-epistemic authority. For instance, the legal system is explicitly concerned with authority and it is often at issue what aspects of life the court or law enforcement has authority over. According to legal conventions, police have the authority to do certain things and the courts have authority to settle certain disputes. What the government can or cannot do and what citizens of the government can and cannot do is all (mostly) defined by the legal system. Moral issues are also sometimes framed as issues about authority.[23] Does a doctor have the authority to deny a patient risky treatment? Does a parent have authority to hit a child as punishment? Imagine a scene in which you see a parent being publicly forceful with a misbehaving child. When I see something like that, I typically experience some internal conflict that centers on questions of authority. Does a parent have some amount of authority to inflict corporal punishment? Certainly there are lines that I know no parent should cross, but I might wonder what amount of force a parent can exert that does not qualify as abuse. Even if I suspect the parent might be crossing some line, I might wonder if I have the authority to challenge their parenting.

While my discussion of non-epistemic authority is conflating a variety of norms that are importantly different in some respects, I think it is useful in the context of the discussion of first-person authority. While calling that authority 'non-epistemic' does little to inform about what first-person authority is, it does make it clear what it is not. The primary puzzles for first-person authority result from thinking of that authority as epistemic. If we let go of that idea, then naturalism about the mind and naturalized

[23] I do not want to claim here that morality is a matter of convention. There are, no doubt, many ways morality is importantly different from etiquette, for instance. All I wish to highlight here is that there are a variety of considerations about what is or is not appropriate that govern our actions.

epistemology become less hostile to the possibility of some sort of interesting notion of first-person authority.

There are a myriad of ways one might try to frame a non-epistemic first-person authority. Each is based around the idea that we might find a set of non-epistemic norms that explain the way we talk and act with regards to apparent claims to self-knowledge. Since apparent claims to self-knowledge are apparent claims to knowledge, one might think that only epistemic norms would ever be appropriate. However, very few discussions are carried out solely under the constraints of epistemic norms (even among epistemologists); often we have non-epistemic reasons not to present the strongest argument or to prematurely cease discussion. For instance, for a wide variety of reasons, I would not personally want to argue in front of a meeting of the NRA that the sale of handguns leads to more accidental deaths than it prevents murders. Epistemically, I might have a duty to argue for the truth. However, no matter how certain one is that something is true or how many good arguments one can think up, even epistemologists need to pick their battles. Even two epistemologists disagreeing at a conference are unlikely to obey only epistemic norms. The fact that non-epistemic authority sometimes creates situations in which it is inappropriate to challenge claims to knowledge is widespread, not unique to self-knowledge.

What is special about self-knowledge claims is that debating facts about self-knowledge can scarcely help but involve norms other than simple epistemic norms. Two people can debate even sensitive political issues in an abstract way, directing their attention outward or toward the issues. Yet, when one of them is challenged because of a claim to have a certain belief, the debate becomes, for lack of a better word, personal. It

is a much different thing to abstractly discuss an idea than it is to challenge, in any way, a person's belief in or adherence to that idea. This is equally true whether the challenge is to the person's actually having a belief or if the challenger grants that his or her opponent has the belief and attacks that opponent for having it. When I issue a claim to self-knowledge, the claim is not just about my mind. It is, on some level, about me.

Interpersonal discussions and interactions invariably bring a host of norms to bear on what is appropriate other than simply a concern for the truth. The crucial thing is not to lose sight of the epistemic questions when they become entangled with other norms. Unless a theorist can show that the presumption against challenging claims to self-knowledge is rooted in the fact that we know our minds better than others it cannot begin to act as a constraint on a theory of self-knowledge. The most we can say as epistemologists about any account of non-epistemic authority is whether or not that non-epistemic authority is a reasonable practice from an epistemic point of view. That is, the epistemologist can criticize non-epistemic authority as misleading or not reflecting the truth, but the fact that such authority exists cannot tell us anything about how well we know.

In the case of self-knowledge, everyone should agree that the amount of evidence required and the difficulty collecting it make actual justifiable challenges to self-knowledge relatively rare, at least for the moment. That is, our accuracy and our access to our own minds give us some kind of epistemic authority independently of any other kind of authority. Admittedly, epistemic authority of this sort is not particularly special. It is defeasible. However, if there are other normative reasons outside of the epistemic that rule out challenges to apparent self-knowledge claims, then our natural epistemic

authority, perhaps already fairly robust, might seem exaggerated. The different norms under which the challenges are inappropriate could be conflated, suggesting one overarching sense in which it is inappropriate to challenge claims to self-knowledge. It would then seem as if there must be a special kind of epistemic authority at work. I think that this is roughly what has happened, and it is, in a sense, special. Our beliefs are unlike other topics of discussion in that our mental states are a part of us. We cannot discuss them with each other without involving the norms of personal interaction, causing us to be especially cautious in critique, treating each other as if we know our own minds best even when we have evidence otherwise.

There are as many possible accounts of the non-epistemic authority of self-knowledge as there are norms governing normal discourse. I will not attempt to be exhaustive or to spell out any proposal in detail, but I will canvass a few options. There are norms that we tend to think of as somewhat trivial, such as etiquette and linguistic convention. It would be interesting to examine if other cultures and speakers of other languages display the same apparent presumptions toward first-person authority. If first-person authority turned out to be nothing more than English convention, then epistemologists would rightly be more skeptical of it.

There are other possibilities. Sydney Shoemaker has claimed that self-knowledge is essential to rationality. He sees self-knowledge as central to human mental life. He argues that it is necessarily connected to our conceptual capacities and our rationality. Here, again, is Shoemaker:

> Suppose a man confronts evidence which contradicts things he believes, and he makes suitable modifications in his beliefs—he modifies them in the way rationality dictates, or at any rate in one of the ways that rationality allows, instead of sticking with an inconsistent or incoherent set of beliefs. Does this

> show self-knowledge? … I think it plainly does if the modification of belief is the effect of deliberation or reflection—we have already seen the central role self-knowledge plays here.[24]

His idea is that the way in which human beings form their beliefs and respond to evidence requires self-knowledge. This is a conception of belief formation and revision as a highly cognitive activity. As we take in information, we have to notice how the new information affects our previously held beliefs. To make rational decisions about which beliefs to revise or replace, we have to know which beliefs we have. For Shoemaker, rationality and self-knowledge go hand in hand. Self-knowledge is, while fallible, more or less guaranteed in anyone that is rational and understands mental concepts.

For Shoemaker, having second-order beliefs follows directly from our having first-order states, conceptual capacities, and rationality.[25] To presume that someone lacks self-knowledge is to presume that he or she lacks one of these things. When we discuss failures of self-knowledge in ordinary people, we can safely assume that they have first-order states and mental concepts. Thus, in his view, any failure of self-knowledge is a failure of rationality.

Shoemaker's idea of a connection between rationality and self-knowledge might be used to explain the authority of self-knowledge. I think that many people would be disturbed to learn of failures of self-knowledge in part because they take it to be centrally connected to their rationality. I think of myself as a rational agent and I would be disturbed to learn of systematic (or even singular) failures of self-knowledge that resulted in my acting or thinking in ways that do not comport with the way I think of myself. If

[24] Shoemaker 1996, p. 31.

[25] His actual view is that there is some sort of constitution relation between the first-order states, the concepts, and the rationality, on the one hand, and the second order states in the other. See the section on Constitutive views in the introduction for more detail. For purposes here, discussing constitution adds a great deal of complexity to the outlining of the view but changes what I want to say about his view very little.

there is a connection between rationality and self-knowledge, then a challenge to a claim of self-knowledge could be seen to imply, in a way that Shoemaker's account would entail, the irrationality of the person making the claim. It is one thing to challenge people's beliefs (or their reasons for their beliefs); it is quite another to accuse them of being irrational.

In everyday discourse, a charge of irrationality can be problematic. For one thing, a charge of irrationality is often a rhetorically unfair attempt to pressure the accused to conform to the thoughts of the accuser. I shall return to this point in discussing agency in the next section. There is another problem special to a charge of irrationality. In order to have a meaningful discussion with someone, we must treat them as rational. That is, to accomplish the normal goals of discourse, in which reasons are discussed in order for each side to test and possibly to revise their beliefs, all parties must assume the others to be minimally rational. If anything like the Shoemaker connection were true, in order to treat each other as rational in the way that we must for discussion, we would need to presume that people had accurate self-knowledge. To avoid derailing discussions, we would be forced to refrain from challenges to self-ascriptions.

While resting on a host of questionable premises, this scheme does create a sense in which it would be inappropriate to challenge someone's self-knowledge. To be clear, this framework is not endorsed by Shoemaker. Rather, it is one based on his views which might provide, in a way that his views do not, a special authority for self-knowledge claims. The central idea is that, you cannot question the rationality of an interlocutor without undermining the purpose of a dialogue. This idea has merit. If you truly believed someone to be irrational, there would be little point in any discussion with that person.

An irrational person is unlikely to make a compelling case for you to change your beliefs. And, if you find the case he or she makes compelling, you'd have at least one good reason to doubt it. Also, irrational people, by definition, do not revise their own beliefs properly in light of argument or evidence. Unless both sides are rational, there is little point to attempting rational argument. You might as well draw straws.

Can the proponent of naturalized accounts of self-knowledge accept this? I do not see why not. There are no obvious constraints on what counts as fair game for discourse imposed by accepting a naturalistic account of self-knowledge. The sense in which it is inappropriate to raise the issue of another person's beliefs, in this picture, is not because one interlocutor is incapable of thinking or knowing that the other is wrong about what he or she believes. Rather, the issue is that talking about *that* won't get you anywhere. If you want to keep debating someone, it is inappropriate to tell them that they do not know what they think.

Of course the story would have to be more complicated than this. The naturalist is unlikely to accept the tight connection between rationality and self-knowledge that Shoemaker proposes. It seems to me that since we do, at least occasionally, make what appear to be justifiable challenges to claims of self-knowledge, we must, *pace* Shoemaker, think that people who make self-knowledge errors can be otherwise rational. The naturalist is likely to think that an otherwise rational person might make some errors of self-ascription, be able to accept that he or she has made errors in light of good evidence, and potentially revise his or her first-order or second-order beliefs accordingly. Perhaps there is a default presumption that we should not challenge self-ascriptions for

fear of casting aspersions of irrationality, but it is not clear that perceptual models are incompatible with such a presumption.

This idea that norms of discourse based on treating each other as rational create presumptions toward authority is similar to proposals more closely examined in the next chapter. One might think that there are discursive norms in place to make sure we respect each other's autonomy. Debates about the truth of a proposition are fine, so long as each debater is free to decide the matter individually. However, if the subject of a debate is the mind of one of the debaters, than the debate might create a challenge to that debater's autonomy. I believe that this idea is a thread running through the discussions of Richard Moran and Akeel Bilgrami on self-knowledge. I return to it in the next chapter.

Importantly, these proposals of the non-epistemic authority of self-knowledge are compatible with our self-knowledge being systematically flawed. None of them rules out the possibility that science can find patterns of error or domains in which our self-knowledge breaks down. We should be open to that empirical possibility. I think that this is the lesson that we should learn from our considerations of MRH. Once we give up the idea of infallibility we accept the possibility of error, and where the possibility of error exists, epistemologists need to try to learn if and when errors actually occur. Also interesting about my account of first-person authority is that if scientists did discover major flaws in our self-knowledge, it would not immediately follow that the conventions underlying the non-epistemic authority of our apparent self-knowledge claims should be abandoned or revised. The conventions might remain perfectly reasonable, at least in non-epistemic contexts.

The remaining issue is whether any kind of non-epistemic authority can create a problem for a perceptual account of self-knowledge. As long as the mental states of individuals are seen as real states having physical properties or producing physical effects, there will be questions about how well we know those states in ourselves. Perceptual model theorists propose processes that make predictions and can be tested. Those tests may or may not violate whatever norms ground a non-epistemic first-person authority. If they do, epistemologists may or may not feel constrained by those norms. If the correct account of first-person authority locates it in etiquette, for example, then scientists and philosophers should show no more restraint in studying self-knowledge than they do studying digestion. However, if there is a moral dimension to first-person authority, more care may be required after all. Suffice it to say that none of these concerns are currently decisive against the naturalized epistemologist or the perceptual model theorist.

5. Conclusion

From the arguments above, it would be slightly hasty to conclude that there is no special epistemic authority of self-knowledge. What I have shown, is that some notions of epistemic authority are perfectly compatible with perceptual models of self-knowledge. It is not a problem for the perceptual model theorist to admit that we are very reliable about our own minds as long as we are not perfectly reliable. It is not incompatible with perceptual models to treat people as having better reasons for their own meta-beliefs than others might in ascribing beliefs to them. As long as these better reasons are themselves capable of being overturned, they do not create special epistemic

challenges. It is not incompatible with perceptual models of self-knowledge to treat people as experts with respect to their own minds. Expertise is a phenomenon than can be handled by any sophisticated epistemology, so long as the experts can in principle be wrong.

If there is any epistemic authority attached to self-knowledge that is incompatible with perceptual models of self-knowledge, it cannot be cashed out in terms of reliability, especially good reasons, or expertise. Perhaps there are other ways for one person to be in a better position than another to know something. The above arguments should place the burden of proof on the theorist claiming that there is some such epistemic authority. Additionally, a theorist making claims about epistemic authority must be careful to show that the authority they articulate is epistemic authority. Perceptual models of self-knowledge should be compatible with a wide range of non-epistemic kinds of authority. If there is some tension between a non-epistemic authority that implicitly underlies our claims to self-knowledge and perceptual models, it is far from obvious that the problem lies with perceptual models or naturalized epistemology.

Chapter 5

Authority as Authorship

1. Moran's Account of First Person Authority

This chapter is an extended consideration of the account of authority presented by Richard Moran in *Authority and Estrangement*.[1] There are many ideas touched on in *Authority and Estrangement*, and I make no claim to an exhaustive response here. Primarily, what I want to focus on is one thread through the book. It involves the idea that our self-knowledge is knowledge by decision and not by discovery. This is a fascinating idea. The positive account Moran presents is somewhat vague and incomplete. One of the primary goals of this chapter will be explore some of the options for filling that account out. Moran also has an explicitly negative goal. He seeks to discredit a certain kind of naturalistic model of self-knowledge. His target, what he refers to as perceptual models, encompass virtually all the theories of self-knowledge proposed by psychologists and naturalistic philosophers. The second major goal of this chapter is to see if any of Moran's ideas cause large problems for perceptual models. I claim they do not.

The principal contrast in *Authority and Estrangement* is between naturalistic or perceptual models of self-knowledge, which view self-knowledge as a process of discovery, and his model, which views self-knowledge in terms of decision. Since perceptual model theorists typically have to explain the authority of self-knowledge in epistemic terms, they face a challenge in accommodating the intuitions that there is

[1]See Moran 2001.

something special about first-person authority. As I argued in the previous chapter, perceptual model theorists, like all theorists about self-knowledge, seem forced to view whatever epistemic advantages the first-person perspective provides about the mind as contingent and they have to recognize some level of possibility that there could be a better way of knowing the contents of one's own mind.

Moran's contrasting model is supposed to be based on decision rather than discovery. I think that this is most clearly shown in his casting authority as authorship. As Moran puts it:

> …there is also a sense of 'authority' as describing the relation of the person to the report itself, indicating the person as responsible for the report, as its author. And it is not just the *report* that the person is the author of, but also, in a central range of cases, the person can be seen as the author of the state of mind itself, in the sense of being the person who originates it and is responsible for it.[2]

The idea, very roughly, is that the reason it is not appropriate to challenge me about what I think is that it is up to me what I think. Perceptual models suggest a kind of internal mind reading, which, according to Moran, fundamentally miscasts the relationship we have toward our minds. For Moran self-knowledge is more "writing" than "reading". Not only am I able to decide what to believe, I am entitled to do so and held responsible for what I decide. He believes that an account of self-knowledge should make central the person as epistemic agent.

Unlike other things one might know about, the subject matter for self-knowledge, the things the knower must have beliefs about, are a part of his or her mind. External facts can vary due to external factors. The weather outside can change suddenly and unexpectedly and one can be wrong about it if it has changed since the last time one

[2]Moran 2001. p. 113.

looked. One might think that self-knowledge is unlike external world knowledge because one's own mind is not entirely at the mercy of external factors. Even if one can change one's mind without realizing it, perhaps the fact that an agent can control, to some extent, what he or she believes represents an important difference between self-knowledge and other kinds of knowledge. At the very least, the amount of control we have over our minds is worth studying and it is worth thinking about the implications doxastic control might have for the epistemology of self-knowledge.

In discussing what control we have over our mental states, limiting discussion to belief, as I have done throughout this dissertation, might seem an especially questionable move. Emotions, for instance, are often thought of as outside of a person's control. Describing an angry person as tempestuous suggests just the sort of parallel with the weather that was suggested above. Our emotions seemingly change due to factors beyond our control and noticing those changes, at least sometimes, seem to be beyond us as well. Belief, by contrast, is sometimes thought to be something we can regulate, with important caveats to be addressed soon. Other mental states, intention for instance, might be even more strongly thought to be the kind of thing we have control over. Hopefully, grappling with issues about our control over our own beliefs will be relevant and helpful to those interested in control over other mental states, but I will not directly explore those issues here.

To see why our control over our beliefs might be important, it is useful to contrast first-person and third-person mental ascription. The process of coming to have knowledge of the minds of others involves forming a belief which ascribes a mental state to another person. This process has been called mindreading in the literature. The content

of a belief in an instance of self-knowledge is very similar to the content of belief from an instance of mindreading. In both self-knowledge and mindreading, the belief formed as an instance of knowledge is one that has a mental state in its content. In terms of our relation to the thing known, mindreading shares certain features with learning about the weather. In both cases the subject matter is external to and outside the control of the agent investigating. The mind of another person, much like the weather, can change when you aren't looking. Success, either in knowing what another person is thinking or in knowing what the weather is like, is a matter of successfully detecting external features. In many of the accounts of self-knowledge offered by naturalists, self-knowledge is also a matter of detection, but detection of internal features of your mind.

Thus far, I have been focused on the relationship between naturalistic models of self-knowledge and the idea that we have a special first-person authority. Richard Moran, one of the leading philosophers interested in the relation between agency and self-knowledge, thinks that the picture of self-knowledge as detection of our own internal states leaves out, "the fact that I not only have special access to someone's mental life, but that it is mine, expressive of my relation to the world, subject to my evaluation, correction, doubts, and tensions."[3] Moran views the project of characterizing self-knowledge, then, as showing how the way in which we know our minds is inherently first-personal. His contention is that a perceptual model is incapable of accomplishing this goal. He writes:

> …[the naturalistic] picture of privacy presents an essentially superficial view of the differences between my relation to myself and my possible relations to others. For in essence what we have here is a picture of self-knowledge as a

[3] Moran 2001, p. 37.

> kind of mind-reading as applied to oneself, a faculty that happens to be aimed in one direction rather than another.[4]

His primary idea for locating the special first-person nature of self-knowledge is to examine its connection to agency. Moran thinks that in discussion of self-knowledge, "inadequate attention is given to the person as epistemic agent, and hence to the mutual interaction between mental life and first-person awareness of it."[5] This is an idea he gets, at least in part, from Sydney Shoemaker. As seen in previous chapters, Shoemaker holds that self-knowledge is an essential component in the ability to revise your beliefs in light of tension or contradiction. He claims that, "if the beliefs and desires are all first-order states… then one thing they do not rationalize is changes in themselves."[6] He asks us to imagine such changes, "as a battle between contending beliefs or inclinations to believe" to attempt to show the role self-knowledge must play. He argues:

> If this were right, it would seem unnecessary that the deliberator should have knowledge of the contending beliefs and desires; he would merely be the subject of them, and the battleground on which the struggle between them takes place. But this model seems hopelessly unrealistic, in part because it leaves out entirely the role of the person as an agent in deliberation; it represents deliberation as something that happens in a person rather than as an intentional activity on the part of the person.[7]

Shoemaker seems here committed to people, as opposed to other kinds of systems, as being able to decide via deliberation what to believe. He thinks that this ability depends on our having self-knowledge. Moran agrees but argues that our abilities as agents actually underlie the special nature of our first-person access. He says, "there would be nothing that counted as agency or deliberation at all if a person could not generally claim the conclusion of his reasoning as making it the case that, as a matter of psychological

---

[4] Moran 2001, p. 91.
[5] Moran 2001, p. 27.
[6] Shoemaker 1996, p. 33.
[7] Shoemaker 1996, p. 28.

fact, *this* is his belief about the matter."[8] He notes approvingly that, "with respect to knowledge of one's own intentions, philosophers sometimes invoke a distinction between certainty that is based on evidence or discovery, and certainty that is based on a decision made by the person."[9] And he suggests that self-knowledge is to be understood in terms of an analogous distinction; he thinks that self-knowledge is based on a decision rather than a discovery.

There is no denying that this idea—that we can decide what we believe—has received little attention in the philosophy or psychology literature on self-knowledge. However, what remains unclear is the extent to which the proposed connections between epistemic agency and self-knowledge impact a theory of the epistemology of self-knowledge. Even if agency, in some sense, requires self-knowledge, it does not immediately follow that our self-knowledge depends in any interesting way on our epistemic agency. It is not even clear precisely what dependence in this direction could amount to; it is not obvious how to regard *decision* as a model for *knowledge*. There is a legitimate worry about whether such a model could even be constructed in a psychologically plausible manner. The problem is that, in certain important ways, our beliefs do not seem to be up to us. Bernard Williams, for instance, has pointed out that beliefs, unlike actions, seem to be constrained in various ways.[10] Moran is sensitive to the idea that, even when actively deliberating or engaging with the question, "what should I believe?" rational agents believe what seems to them most likely to be true. No matter how hard we try, we cannot simply make ourselves believe P while P seems to us to be

[8] Moran 2001, p. 120.
[9] Moran 2001, p. 55.
[10] See Williams 1973.

false. Not even if we would like P to be true. Much of the maneuvering in *Authority and Estrangement* is designed to describe something like a process of choosing what to believe while avoiding the obvious objection that we cannot choose, willy-nilly, what to believe. Since Moran's account of deciding to believe something involves sensitivity to evidence, we could make a distinction between a standard decision and a doxastic decision. The idea would be that doxastic decisions, unlike the standard kind, are constrained by rationality and evidence in a certain kind of way. This move leaves Moran the difficulty of describing doxastic decisions, but does seem to allow us to bracket the above obvious problem.

Moran often seems to think that the most important feature of his account is that it makes self-knowledge *inherently* first-personal. He routinely rebukes perceptual models for failing to do the same. I have in mind passages like the following:

> [on a perceptual] model, then, there would seem to be no deep reason why one couldn't bear this quasi-perceptual relation to the mental life of another person as well as oneself.[11]
>
> A person can make reliable psychological ascriptions to himself immediately, without needing to observe what he says and does. And this capacity lies in the nature of the first-person position itself; it is not a kind of access he may have to the mind of another person.[12]
>
> …[the perceptual] picture of privacy presents an essentially superficial view of the differences between my relation to myself and my possible relations to others. For in essence what we have here is a picture of self-knowledge as a kind of mind-reading as applied to oneself, a faculty that happens to be aimed in one direction rather than another.[13]

In passages like these it seems that Moran wants to make it central to an account of self-knowledge that it takes into account the differences between our relations to our own minds and our relations to other minds. Moreover, it seems he would require that an

[11] Moran 2001, p. 33.
[12] Moran 2001, p. 12.
[13] Moran 2001, p. 91.

account explain why it would be impossible to know other minds in the special way that we can know our own. Moran seems committed to a very strong claim about the characteristic self-knowledge relation: that an epistemic agent simply could not bear this relation to another. The idea is supposed to be that fully accounting for ourselves as epistemic agents would show that self-knowledge is inherently first-personal. I think that if our epistemic agency could reveal how our self-knowledge is inherently first-personal, it would go a long way toward showing that self-knowledge was special.

Any perceptual model will treat the subject matter of self-knowledge, to put it in Moran's terms, as things to be discovered. That is, existing beliefs are thought to be detected or somehow gotten at by a mechanism or process that is sensitive to their existence. There is no room in such a model for discussion of how those beliefs are formed because the model assumes that beliefs must already be formed in order to be known about. For the theorist defending a perceptual model of self-knowledge, our control over our own beliefs is potentially interesting but somewhat unrelated subject. Moran stresses this disconnection; he claims that epistemic agency is an integral part of the study of self-knowledge.

It is also not difficult to see that perceptual models do not allow for some deep connection between self-knowledge and the first-person. Indeed, Armstrong is explicit in denying that his view is inherently first-personal, saying that we "could conceive of a power of acquiring non-verbal non-inferential knowledge of current states of the minds of others."[14] Telepathy is an old idea despite being confined mostly to science fiction today. It would be off topic to discuss the possibility of something like telepathy here. I bring it

[14] Armstrong 1968, p. 326.

up to highlight what bothers Moran about perceptual models. Essentially, perceptual models invoke a kind of internal telepathy: an ability to *read* one's own mind. Moran thinks that our abilities as epistemic agents, you might say the ability to *write* our own minds rather than read them, underlies a relation that is much different in character than the kind presupposed by perceptual models. That is, he thinks it is conceptually impossible that we be aware of another's mind in the same way we are aware of our own. Supposedly, this becomes clear when you fully account for our ability to make up our own mind.

While there is a normative dimension to every epistemological issue, the issues of responsibility for our beliefs seem somewhat unavoidable in the context of a discussion about our control over our beliefs. Moran touches on this idea at times. He wants a view of self-knowledge that captures, "that I not only have special *access* to someone's mental life, but that it is *mine*, expressive of my relation to the world, subject to my evaluation, correction, doubts and tensions."[15] Moran clearly advocates a connection between the authority of self-knowledge and an agent's causal responsibility for the existence of a belief. But what, exactly, is the connection?

There are several issues worth considering. Firstly, is the authority proposed here epistemic or non-epistemic? Ultimately, I think Moran would concede that it is non-epistemic, though it sometimes does seem like he might want his account of authority to deliver some sort of special epistemic authority as well. Below, I explore both possibilities. Another important concern is how the proposed authority is supposed to be incompatible with naturalistic or perceptual models of self-knowledge. As I consider

[15] Moran 2001, p. 37.

different options for agent centered accounts based on Moran's ideas, their compatibility or lack thereof with perceptual models will be a primary concern.

In the next section, I explore the extent to which we do have control over our beliefs. If Moran's account is to have much merit, then it must actually be up to us what beliefs we have. However, there are good reasons to doubt that we have a high degree of control over our beliefs. In sections three and four, I will explore the possibilities for an agent centered account to produce a special kind of authority. First, in section three, I will examine the possibility that such an account provides a special epistemic authority. I conclude that it does not, but note that there is some reason to include an agent centered account of the epistemology of self-knowledge as a competitor with perceptual accounts. Second, in section four, I examine the possibility that Moran's idea of authority as authorship creates a special kind of non-epistemic authority. I think that there is a potentially very interesting idea of non-epistemic authority that can come from Moran's account. I think that ultimately, the notion of authority he presents in the book is best viewed as a kind of non-epistemic authority. However, I do not think that this sort of authority will create any problems for perceptual models of self-knowledge. Indeed, I think most perceptual models can easily accommodate the idea that active introspection can play a causal role in first-order belief formation and can thereby recognize a non-epistemic authority for self-knowledge much in the spirit of Moran's account.

Finally, in section five, I examine another account of self-knowledge that has much in common with Richard Moran's. As I understand Akeel Bilgrami's account in *Self-Knowledge and Resentment*, it succeeds in framing a kind of first-person authority that meets Moran's goals of being inherently first-personal and incompatible with

perceptual models of self-knowledge. However, it accomplishes these goals by fully giving up the naturalist conception of mental states and the causal relationship between believer and belief that Moran seems committed to in much of his book. Bilgrami's account nicely highlights how far away from naturalism an account of self-knowledge has to go in order to be incompatible with perceptual models.

2. Deliberation in Action and Inquiry

Before examining the role of voluntarily control in epistemic matters, some remarks should be made about how to understand freedom and volition. While it is well beyond the scope of this chapter to take a stand on issues of free will, there is a clear need to have some basic options for understanding free will on the table in order evaluate theories of the role free will plays in self-knowledge.

The first point that should be made about free will is that the philosophical motivation for providing an account of free will is often taken to be the feeling of freedom that accompanies our actions. That is, when we act, we always feel as if we could have done otherwise; we feel as if our action was performed because we willed that that action be performed rather than any of our alternatives. Throughout this dissertation, I have been skeptical of allowing phenomenology or intuition to constrain theorizing. I remain so now. Still, it is interesting to compare the feeling of deciding what to do with deliberating about what to believe.

In both cases, there seem to be norms governing the process. In the case of action, these norms are things like morality and self-interest; in the case of belief, justification and truth. During the process of either kind of deliberation, there is a feeling of

uncertainty, often an uneasy feeling. And, I think, in both cases, during the process of deliberation, there is a feeling that the outcome is, as Moran puts it, up to you. However the end result of a deliberation about what to do is an intention to act. Subjectively speaking, this intention does not seem to guarantee that any action is performed. For lack of a better way to characterize it, it seems to me that, even after reaching a decision about what to do, an *act of will* is required before I actually do it. Suppose I think about asking a question at a lecture and it ultimately seems to me to be a good idea. Ignore for now whether there is any difference between it seeming like a good idea or the right idea and deciding that you will do it. Even after I have decided, I must actually raise my hand and I must make the words come out of my mouth. There are sometimes disconnects here, between forming an intention and carrying out an action, that seemingly have no analogue in deliberation about what to believe.

Indeed, as I consider evidence for and against the truth of some proposition, the scales typically begin to feel as if they tip one way or the other.[16] Once this happens, my belief seems to be constrained by the way the scales tipped. If P seemed true, we take ourselves to believe it. If P seemed false, we take ourselves to believe not P. If the evidence was not overwhelming one way or the other, we think that we withhold belief. The belief, at least as we typically understand it, forms in us the way that an intention does. Though it seems, especially while deliberating, to be up to us what to believe, there is some feeling of constraint as well. I cannot, or so it seems to me, get myself to believe P when P seems to be false in light of the evidence.

[16] Ignoring, for the moment that many beliefs are formed unconsciously.

Free actions often feel totally unconstrained. For instance, in a morally ambiguous situation, even when I do what feels like the morally correct thing, I feel as if I could have done something else instead. It seems as though my judgment that one action was correct does not in itself determine what I actually do. Sometimes this is taken to be a disconnection like the one above, in which an intention is formed but does not lead to action. However, deliberation about what to do also typically seems unconstrained in sense that the norms involved do not appear to constrain the formation of an intention. Judging what is right or in one's self-interest does not, at least subjectively, lead immediately to the intention to do it. This is a feeling experienced by almost everyone that has ever judged it to be in his or her interest to clean the dishes or finish writing a paper. Sometimes it seems as if the intention forms and the action is performed. Sometimes it seems as if the intention forms and the action is not performed. Finally, sometimes the intention is seemingly not formed at all. From the inside, it seems possible to regard an action as right or beneficial while explicitly deciding not to perform it. By comparison, it is not possible, subjectively, to believe contrary to what seems true or warranted by the evidence. So we have reason, purely from the internal perspective, to be wary of any notion of decision which entails that beliefs are amenable to acts of will above and beyond the judgment of what is true or justified. If our self-knowledge is connected to agency, it is through our judgments about the truth, not via some ability to directly determine our own beliefs.

However, in being suspicious of this subjective story, we should note the role that self-knowledge plays. In particular, we take ourselves to know when we have formed an intention and to know when we have formed a belief. In theorizing about action, one's

knowledge of one's intentions is often taken for granted. Subjectively, this seems more than merely plausible; it seems obvious. However, objectively, it appears a substantive assumption and perhaps even an unjustified one. How, for instance, can we be sure that we really do form the intention to ask a question at a lecture? From the external perspective, when a person does not ask a question, it is possible that that person formed no intention or that the person did intend to ask a question but failed to for some other reason. Though you take yourself to have intended to ask a question, how is it that you know you really did form it in cases in which you do not ask a question? Likewise, when we reach the end of a deliberation about what to believe we take ourselves to have formed a belief. How is it that we know the belief is there? In both cases the answer seems to be that we trust the phenomenology of the first person and that we rely on a presumption that we know our minds.

The questions here require an explanation of how the meta-belief, or belief about one's intentions in the case of action, is formed. One answer would be that I know I have the relevant belief or intention because it seems to me that I have made a decision. Another would be that I have some sort of internal detection mechanism capable of detecting the intention or belief once formed. Yet another is that, upon reaching an end to deliberation, we immediately form a meta-belief that is consistent with the result of deliberation. That I am considering these alternatives suggests that I am not taking seriously Moran's idea that considerations of epistemic agency should reveal a problem with the perceptual model. I am here viewing it as one of the possible ways you can know that you have a particular belief after deliberation has run its course. This is, no doubt, partly because I am concerned with the formation of the meta-belief. In the next section, I

will consider whether there are genuine alternative views of meta-belief formation that have so far been over looked.

In the remainder of this section, I will examine the issue of trusting that certain mental states have formed after our deliberations. I propose to do this by analogy. Daniel Wegner has extensively studied the phenomenon of the subjective feeling of will in action. As I have noted, there is a similar feeling at work in deliberation about what to believe. Wegner argues that the feeling of will is separable from performed action in two ways. Under the right conditions we can be made to feel as if we are not doing something when we are (automatism) or as if we are doing something when we are not (illusion of control). He argues that the way in which these phenomena are separable indicates that conscious intention and voluntary action are produced by distinct processes and that the feeling of conscious will is the result of an inferred connection between the conscious intention and the action.[17]

I find Wegner's case to be largely compelling, but I worry about certain assumptions he makes. However, I will not evaluate Wegner's account at length. Instead, I will pursue a speculative theory of the phenomenology of deliberation inspired by Wegner's account of conscious will. I will start with a weakened version of Wegner's account.

> **WEG**: The feeling of will associated with normal voluntary action is the result of the mind's association of a felt intention with a perceived action.

Unlike Wegner's account, WEG is neutral with respect to whether a subject has an intention even when that subject self-attributes one. This is an important difference. It allows for a possibility that Wegner does not consider; that deliberating might produce a

[17] See Wegner 2002, especially chapters 1 and 3.

felt intention without producing an actual intention. Wegner believes that action is often caused as the result of unconscious states and processes that result from a separate system than the conscious thoughts one has about one's actions. Given the systems Wegner describes, there seem to me to be four options for what an intention is. Intentions might be identified with the unconscious states resulting from the unconscious cognition leading up to action, the conscious states resulting from conscious deliberations, some sort of combination of the two, or with nothing in the system at all. This division in logical space mirrors one that I suggest for belief in the next chapter. There are low road (unconscious) views, high road (conscious) views, hybrid views, and eliminativist views.

Wegner seems drawn to a high road view. He often seems to identify the result of the conscious system (a felt intention) with actual intention, leaving the causal work of performing the action to be done by whatever states result from the relevant unconscious processes, whatever those are. However, if there are unconscious states that produce actions, it is not clear to me that these states are not intentions. To assume the high road view is to trust a kind of self-knowledge. In particular, Wegner seems content to equate feeling as if one intends to act with having an intention to act. If we had perfect knowledge of our intentions, this would be an innocent equivocation. However, in the context of a discussion of self-knowledge it does not appear so innocent. Given that self-knowledge theorists often assume we have imperfect knowledge of our intentions—and that there can exist unconscious intentions—we need to consider the option that the mistakes made in the cases Wegner describes are not failures of intentions to cause actions, but failures of agents to know their own intentions.

To some extent, this position and Wegner's position may be in agreement. Both views describe systems in which conscious and unconscious processes happen in parallel and which produce unusual effects in rare cases in which they arrive at mismatched states. However, the differences threaten to be more dramatic than just the way in which the systems are described. What Wegner describes as failures of mental causation might better be thought of as failures of self-knowledge.

Notice that WEG preserves Wegner's controversial claim that the feeling of will involved in normal conscious action is an illusion, even when combined with this alternative view of intention. Further work must be done to determine the character of this illusion. Is the illusory connection one of efficacy (between intention and action) or one of epistemology (between intention and self-attributed intention)?

In addition to having potentially serious consequences for theories of free will and action, the alternative view I discuss has the potential to shed some light on deliberation in the case of belief. If we view the illusion in normal action as being that self-attributed intentions do not always match up with actual intentions, we might wonder if there is a parallel situation for belief. That is, we might wonder if, just like when we deliberate about what to do deliberating about what to believe involves two systems, one conscious and one unconscious which results in two separate states conflated by the mind. I return to this idea in the closing chapters of the dissertation.

3. Agent-Causal Models

Having noted the worries about taking apparent control over belief as actual control, let us now bracket them and assume for the sake of argument that we have some

measure of control over our beliefs. Return to the suggestion that deliberating about what to believe was supposed to be critically important in a full understanding of self-knowledge. How are we supposed to incorporate our control over our beliefs into a theory of self-knowledge?

Consider a game of capture the flag. We can think of the difference between two ways of knowing that there is a flag tied to a branch in a tree. One way to know such a thing would be to see the flag tied to the tree. You might come to know where the enemy flag is in this way. You might know where your own flag is in a different way: by having climbed up in a tree and tied the flag to it. Of course, you would likely see the flag in the tree during and after the process of tying it, but perhaps there is something different about being (knowingly) responsible for the flag's being in the tree that has weight. One might say that it cements the belief. Later, when I am no longer looking at the tree, I can be confident that it has not fallen because I tied a good knot. Knowing the reason it was put there, I can know something about conditions under which it might be removed. It is still there unless my opponents have found it and removed it. Perceptual models of self-knowledge hold, very roughly, that you "see" your own beliefs like you might see a flag hanging in a tree. Above, we saw that Richard Moran and others want to claim that this model fails to take into account the active role we have in determining our beliefs. Thus, if the analogy here holds, we should look for a model which takes into account your role in tying the flag in the tree.

As I noted earlier, it will have to make use of something like a doxastic decision. That is, a model of self-knowledge that maintains a central role for epistemic agency must spell out the details of doxastic deliberations in such a way that the role of an

individual's evidence is understood as a part of the process yet does not undermine the status of the resultant belief as having been *decided*. Assuming that task can be accomplished, there are two issues that jump to the forefront. One is a question about the scope of the model. The other is about its tenability.

It becomes plain rather quickly once one starts thinking about instances of self-knowledge, that the feeling of deliberation that accompanies introspection is not ubiquitous.[18] It is *prima facie* implausible to think that each and every meta-belief a person has is the result of a doxastic decision. People are just not *that* reflective; it seems they sometimes know what they believe without going through much cognitive work. For instance, once a doxastic decision has been made, how does the agent later know that he or she has the belief decided upon? It had better not be by anything like a perceptual mechanism. If it were, the two models would not be in competition and the most that could be said for epistemic agency is that it too is part of a full understanding of the mind. Some other faculty, like memory, might be appealed to but the worry about the scope persists. It seems that we sometimes know what we believe without any conscious deliberation. What is clear is that deliberation about what to believe is much less common than belief formation and any idea of privileged access should be broader than cases of conscious deliberation.

There is also a worry about the tenability of a model that connects knowing what one believes with deciding what one believes. Suppose I decide to believe that it is raining outside when I hear sounds much like raindrops hitting the roof. If I am also to come to know that I believe it is raining outside at this time, is it by knowing what

[18] Jane Heal makes this point in the PPR symposium on Moran's book. See Heal 2004, pp. 428-429.

decision I have made? How do I know what decisions I have made? Here it would be even more damaging to the prospects of creating an interesting model of self-knowledge to inject a quasi-perceptual element. If you have to "see" what decisions you have made, you might as well just "see" what beliefs you have.[19] It's not clear what the motivation for positing the extra element would be or what difference it could make for establishing self-knowledge as something special.

The terrain, then, is rough for a theorist that wants to claim self-knowledge is special knowledge because of epistemic agency. That theorist must articulate a notion of doxastic decisions that is coherent and tenable in light of the considerations raised above. That theorist must also provide a comprehensive account of self-knowledge that competes with perceptual models by covering even commonplace cases of self-knowledge. I think that the most natural way to attempt this would be to make doxastic decisions responsible for beliefs and meta-beliefs contemporaneously. Call any model like this an *agent-causal model*. Rather than a causal mechanism "perceiving" an existing first-order belief, a doxastic decision creates both the first-order belief and our knowledge of its existence. The most significant difference between agent-causal models and perceptual models is a difference in direction of causation. According to agent-causal models, agents cause beliefs in themselves as they make doxastic decisions, and, during that process, come to know what they believe. The perceptual models have the direction of causation going the other way. According to perceptual models, beliefs, however they are caused, somehow cause you to recognize that you have them once they are set.

[19] This point was made to me by Hilary Kornblith in conversation, although in a slightly different context.

Proceeding with this rough idea and ignoring the potential difficulties in filling it out, we can return to the goal of introducing agent-causal models. Perceptual models, remember, treat self-knowledge as much the same as other kinds of knowledge. The discussion of epistemic agency was supposed to reveal to us something about the nature of self-knowledge that perceptual models could not account for. To paraphrase Moran, perceptual models supposedly failed to account for the inherently first-personal nature of self-knowledge by articulating a faculty that merely happened to be pointed in one way rather than another. On an agent-causal model, would self-knowledge be arrived at via a mechanism that is inherently first-personal?

No. Remember the flag analogy. If perceptual models hold that you "see" your beliefs as you might see a flag in a tree, one might argue that it is mere accident on such a model that you can see only your own tree. On the agent-causal model side of the analogy, you come to know the flag is there when you tie it. But why think it is impossible, rather than just impolite, to tie flags in other people's trees? Whatever causal work doxastic decisions are supposed to do, it seems they could in principle do to establish someone else's belief. If I somehow come to know that I believe P by undergoing a deliberation and deciding to believe P, I could, in principle undergo a deliberation and decide for someone else to believe P, thereby coming to know that he or she believes P. It could only happen in a bizarre possible world much different from our own, but it does not seem incoherent. To help imagine it, restrict the ability to doxastically decide for another what he or she believes to a certain domain. Perhaps parents in such a world would be able to doxastically decide certain things for their

children with respect to sex and drugs, or scientists for politicians with respect to the climate and the environment.

Useful as an ability like this might be, it is not one that we happen to have. Yet, that is no reason to think it impossible. Perceptual models can make a similar claim. They claim that my mental states cause me to know that I have them in a particular way. The mental states of others do not, not in the same way at least, cause me to know about them, although they could in principle. Both accounts describe relations that we surely do not bear to the mental states of others in the actual world, but which it is at least logically possible that we could. The lesson to learn here is that, in order to establish an account of self-knowledge which makes it impossible, even in principle, to know the mind of another in the same way that we know our own minds we must abandon strictly causal accounts of the formation of our second-order beliefs. Any proposed causal story will amount to nothing more than a "faculty that happens to be aimed in one direction rather than another".[20] No matter how restricted a certain causal process is in the actual world, a casual process mirroring it without those restrictions could exist in a stranger possible world.

The epistemic authority generated by an agent causal model will also be on a par with the epistemic authority that can be countenanced by a perceptual model. The story for a perceptual model theorist is that our epistemic privilege comes via no one else being able to employ the same introspective processes that we can apply to our own minds. It is a practically important privilege, but one that could be defeated. Granting some connection between epistemic agency and self knowledge produces the same sort of

[20] Moran 2001, p. 91.

contingent authority. It is a noteworthy fact that once I have made up my mind, no one has any actual ability to change it for me. This privileges my epistemic position with respect to my own mind to some extent. However, it is in principle possible, if not in practice, for someone to change it. Moreover, since both the perceptual model and the agent causal model recognize the possibility of error, both should admit the same sorts of external evidence as potential defeaters of self-knowledge, whether or not discursive practice actually does.

The upshot here is that epistemic agency might well have something to do with self-knowledge, but positing connections with agency does not immediately reveal self-knowledge to be different from other kinds of knowledge. Focusing on and attempting to elaborate this idea of doxastic decision might be fruitful. As epistemic agents, our beliefs can be subject to change. It is not at all unreasonable to think that our coming to know which beliefs we have might occur, at least in part, along with the formation and revision of our beliefs. The issue, which will require further empirical work to settle, is to what extent there are separate processes involved with the formation of second-order beliefs and the formation and revision of first-order beliefs.[21]

4. Authorship as Non-Epistemic Authority

Recognizing a causal role for an epistemic agent to make up his or her own mind in a theory of self-knowledge does not seem to generate any more interesting epistemic authority than perceptual theorists can already accommodate. However, there is some reason to think that it generates a kind of non-epistemic authority. Here, I grant for the

[21] Given the earlier worries about the scope of a theory of self-knowledge, it should be noted here that, even if there are times when the process of forming a belief is relevant to the forming of a meta-belief about it, there might still be many other instances in which meta-beliefs are formed via a perceptual mechanism.

sake of argument an implausibly strong kind of doxastic agency to consider what consequences such agency would have for the idea that we have a special authority of self-knowledge. Presumably, if an agent's role as the source of his or her beliefs generates some non-epistemic authority, the extreme case in which an agent is in full control of his or her beliefs will generate no less than a more plausibly restricted case.

Suppose that a person had complete and total control over what mental states he or she had. Specifically, suppose that, contrary to what seems to be the case, if you will yourself to believe something, you come to believe it.[22] Call this view Doxastic Voluntarism (DV) The first thing to note is that DV does not guarantee infallibility. I might have a belief that P, while thinking that I have a belief that not P and simply never realize this predicament.[23] Perhaps, if I realized it, I would will myself to believe not-P. However, while unrealized, it seems to me the belief would likely remain unchanged, creating the possibility for error.[24] In other words, DV posits the ability to correct detected errors, not the inability to make errors.

Since DV does not rule out the possibility of error, it is compatible with MRH. I take it that, even if the mind is dynamic enough to change beliefs from minute to minute as a person wills, there will always be a fact about the beliefs for the mind reader to register at an instant. If DV is true, a person would be in a superior position to the mindreading device in one sense. Anyone learning of a belief that he or she did not want to hold could change his or her mind. Presumably the scanner could register a change, but

[22] Moran himself concedes that this is too strong and that when "deciding" what to believe, evidence constrains or influences in some way. See Moran 2001, especially chapter 2.
[23] See Williams 1973.
[24] I'm making this out to be an act of conscious effort and intention. Perhaps voluntarism should be as simple as believing whatever you want to believe. While this removes the possibility I describe, it creates similar ones. Perhaps I want to believe P and I do, but I do not realize that I want to believe P. In that case, I could still believe that I believe not P.

not enact one. As important as this difference between detecting a change versus making one might be, it is not a difference in knowledge. Even holding DV to be true, one could not hope to know one's own mind better than all possible mind reading devices.

Accepting DV would provide an account of how we come, at least some of the time, to have first-order beliefs. It does not provide, on its own, an account of how we know about these beliefs. An additional account is necessary. To see this, consider that DV is compatible with a perceptual model of self-knowledge. The idea would be that you could use your internal mental scanner to check your beliefs and help maintain your beliefs in the way that you want. Perhaps, before moving on, it is worth thinking one more time about MRH. Suppose, still assuming DV is true, that a mind reading device has given you a surprising result. In what sense could the result of the device be said to impinge on your right to make up your own mind? Clearly the device has no stake in whether you believe one thing or another. If you do not like what the device has told you, you should simply will yourself to believe differently. If anything, it seems to me that a device like the mind scanner would help you to make sure you had the beliefs you would like to have. Without an independent check on what you believe, your ability to make mistakes about your beliefs threatens your ability to make up your own mind. This seems true regardless of how much control we have over what we believe.

One way to extend the idea that challenges to claims of self-knowledge are inappropriate because they impinge on our authorship of our minds is to think of the challenge as a kind of interference: an attempt to influence the one making the self-ascription in an inappropriate way. After all, as Moran pointed out in the quotation above, it is not just that we think a person capable of deciding what to think. We actually hold

people responsible for the beliefs that they have. Going hand-in-hand with the idea that people are responsible for their beliefs is the idea that they are entitled to whatever beliefs they want to have. Thus, when someone claims that they believe something, maybe it can be seen not merely as a report of fact but as a statement of what they have decided to endorse. To debate this claim might be more like interfering in their affairs than disagreeing about facts. As with the exploration of views similar to Shoemaker, I do not want to attribute this view to Moran. I am offering it as a possible way to get special authority from some of his starting points.

Here's the most plausible kind of case I can imagine. Suppose you report to a friend that you believe that men make better firefighters than women. If your friend says, "I know you don't really believe that because you aren't a sexist." You might see your friend as passing judgment on your having a particular belief and exerting pressure on you to ensure you have the proper first-order belief: the belief that women can be as good at firefighting as men. With a charge like this in a debate, your friend may be applying tactics of shame and ridicule to influence your beliefs. If, by some set of norms or other, we view people as having the right to determine for themselves what to believe, what your friend is doing may be inappropriate under the normal rules of discourse.

There are several important points to make here. The first is that the norms under which this sort of challenge can be seen as inappropriate are not epistemic. If the inappropriateness is a result of wielding undue influence, there is no presumption made that you know better than your friend what is in your head. Rather, the presumption is that your friend ought to let you make up your own mind. The second thing to note is that this sort of inappropriateness is not especially tied to the domain of *self*-knowledge. If

people are entitled to form their own beliefs in the absence of certain kinds of external pressure, then it would be just as inappropriate and for the same reasons for your friend to say "I hate people that think women can't be good firefighters," or "only an idiot would think that," or even, "change your mind or you can't come to my party," all of which implicitly accept the self-ascription and apply pressure to change the belief. This undue influence might be exerted even without anyone mentioning a mental state at all. For instance, one person might assert some proposition P, prompting the other to say, "It's really obvious that P is false." As long as the implication is made that one would be foolish, irrational, or immoral to accept a proposition, people can exert influence on each other's beliefs by discussing the truth of matters directly.

This point raises questions about exactly what one is allowed to say in disagreement with another.[25] Presumably, we do not want people to refrain from saying everything that might unduly influence a person's beliefs. I should think that even a theorist that held DV to be true would think that discussion of current beliefs, including the presentation of arguments and evidence, is not just an important part of epistemology, but also our daily lives. Respect for a person's right to make up his or her own mind cannot go so far as to rule out interpersonal discourse. It seems to me, as hinted above, that an inappropriate challenge to a person's right to make up his or her mind has much more to do with the use of shame or rhetorical force to coerce than it does with challenges to self-knowledge. If we do have a presumption against challenging claims to self-knowledge because we are, in some sense, worried about interfering with the domain of others, then we would be refraining from a certain kind of challenge on non-epistemic

[25] This point was originally made to me by Hilary Kornblith in conversation.

grounds. In this picture it is not our place to make such challenges, even when we have perfectly good epistemic reasons to do so. As interesting as such a restriction on discourse might be, it is not an account of a special authority that undermines perceptual accounts or naturalized epistemology.

If there are implicit non-epistemic norms governing discursive practice surrounding apparent self-knowledge claims, we can potentially criticize these norms in several different ways. First, we could look at these norms from an epistemic standpoint. Unsurprisingly, such norms would get in the way of uncovering the details of an epistemology of self-knowledge and would be seen unfavorably from the standpoint of epistemology. However, we recognize all sorts of non-epistemic norms as restricting epistemic practice. The easiest way to get to the truth may sometimes be by asking rude questions, running immoral experiments, or operating without government permits. Yet all of those norms serve important purposes. To advance a naturalistic view of self-knowledge, it is important to try to understand any non-epistemic norms surrounding it in order to arrive at the truth in a responsible way. From the epistemologist's perspective, I suggest that this is the message to take from Moran's work.

5. Inherently First-personal Self-knowledge

So far, I have been focusing on the causal role that an agent might have in generating his or her beliefs. However, there is another dimension to the suggestion that authority is authorship. Moran notes from time to time that we are subject to evaluation for our mental states. Perhaps the idea that we are responsible for our mental states can be extended beyond the idea that we cause them. While Moran does seem to think that

causal responsibility is part of what perceptual models miss, it is not clear whether Moran thinks that causal responsibility underlies all other sorts of responsibility that we have with respect to our mental states.

Another recent account of self-knowledge more explicitly examines our responsibility with respect to our mental states. It takes more than three quarters of Bilgrami's book, *Self-Knowledge and Resentment,* for him to spell out and defend what he takes to be the connections between agency, value, thought, intentionality, and normativity. My summary below of the steps of his argument cannot hope to do justice to the details of the view or to defend the steps of his argument from various objections. Indeed, there are many steps of the argument that I would take issue with, even in spite of the skillful defenses he mounts as he proceeds through the book. My reason for bringing Bilgrami in here is that I think his account comes closer to meeting certain goals that Moran set out for himself: the goals of being incompatible with perceptual models and inherently first-personal. However, once it becomes clear how Bilgrami manages this, it becomes clear that it is not the first-person perspective that creates trouble for the naturalist. Rather it is the thoroughly anti-naturalist view of the mind adopted at the outset of Bilgrami's project. While some will no doubt be attracted to his view, I think it serves to illustrate exactly how far away one must move from certain presuppositions many theorists of mind hold in order to meet Moran's ideals.

In order to understand the view, one must understand the concept of intentionality that Bilgrami is working with. As I mentioned above, for Bilgrami, intentionality is deeply interconnected with several other complex subjects, such as agency and normativity. Genuine intentional states, for Bilgrami, are states that are justifiably subject

to certain kinds of evaluative reactions. It is helpful to continue to focus on belief here. Roughly, the idea is that genuine beliefs are the sorts of things that an agent can be held accountable for: both for having or not having and for acting or not acting in accordance with them. More specifically, Bilgrami classifies beliefs as a kind of commitment. To have a belief is to be committed to the truth of some proposition as well as to holding certain other beliefs, ones entailed or at least consistent with it, and to actions that are consistent with that belief (in conjunction with your other beliefs and desires, all construed as commitments).

Suppose I believe that it is raining. On this picture, I would be committed to believing other things, for instance that I will get wet if I go outside unprotected, and to acting in certain ways, such as bringing an umbrella outside if I do not want to get wet. The important point here is that my commitment to these things entails that *I can be held accountable for failing to live up to the commitment* but entails nothing about *whether or not I will live up to this commitment*. This is a thoroughly normative notion of belief. If someone has a belief, according to Bilgrami, he or she *should* hold other beliefs and he or she *should* perform various actions, completely independently of whether or not he or she *actually does* hold those other beliefs or perform those actions. I stress this point not simply because of the hurdle it poses to understanding the view, but because it is the ultimate source of my criticism.

This thoroughly normative notion of intentionality is complemented by a thoroughly normative notion of agency that begins with and extends the Strawsonian account of agency.[26] Strawson's idea is that the sorts of free actions that a genuine agent

[26] Strawson 1974.

performs cannot be defined in purely descriptive terms. Rather than think that some purely metaphysical difference exists between free actions committed by agents and unfree or coerced actions, Strawson claimed that we needed to look to the evaluative reactive attitudes that we have toward agents given their actions. Free actions, he claimed, are the ones that we justifiably hold agents responsible for via our reactive attitudes of praise and blame. Part of what is important for our reactive attitudes to be justifiable is that they be sensitive to the beliefs and desires of the agent. The actions that an agent performs are not the only important part of the story for our evaluative reactions to be justified. For Bilgrami, with his conception of beliefs as commitments, what beliefs an agent has play a large role in determining the appropriate evaluative reactive attitudes because the actions that an agent should or should not perform are at least constrained in some way by the commitments of the agent. If an agent performs an action that conflicts with the commitments of that agent's belief, then the agent is justly blamed for not fulfilling his or her commitment.

Another important aspect to our justifiably reacting to an agent's behaviors, according to Bilgrami, is that an agent must be aware of his or her commitments. This is not particularly surprising or controversial by itself, but, coupled with this thoroughly normative notion of intentionality, with beliefs as commitments, it has the somewhat surprising result that an agent must have self-knowledge. If an agent lacked self-knowledge, that agent would lack knowledge of his or her commitments thereby making the behavior of the agent inappropriate subject matter for evaluative reactive attitudes. Given the associations between these attitudes and genuine free action on the

Strawsonian picture, without self-knowledge there could be neither genuine free action nor agency.

At this point, it should be clear that self-knowledge is required for agency on Bilgrami's view. This would be an interesting claim, regardless of what conception of belief he was working with. However, it only becomes clear how all of these ideas fit together when we see how Bilgrami describes the conditions under which someone has the kind of commitment he thinks are involved in intentional states. He says:

> To have a commitment, one must be prepared to have certain reactive attitudes, minimally to be self-critical or be accepting of criticism from another, if one fails to live up to the commitment or if one lacks the disposition to do what it takes to live up to it; and one must be prepared to do better by way of trying to live up to it, perhaps by cultivating the disposition to live up to it.[27]

Thus, whenever one is willing to accept criticism for failing to live up to a commitment, either from one's self or from another, one has that commitment. It should be fairly obvious that only agents could meet this requirement of *willingly* accepting criticism and *trying* to do better. Agency, then, is required for having beliefs *qua* commitments and, *a fortiori*, for self-knowledge. The question to ask now is, if we accept all these proposed conceptual connections, what could be said about the specialness of self-knowledge?

On this model, we have beliefs when we are prepared to accept criticism for them. He later asserts that the conditions under which an agent is prepared to accept criticism for a belief are the very same conditions under which someone can sincerely avow a commitment.[28] Thus, whenever someone can sincerely avow a belief, that person in fact has the commitment that, for Bilgrami constitutes the belief. On this picture, then, self-knowledge is somewhat self-guaranteeing as long as an agent is sincere. Because belief is

[27] Bilgrami 2006, p. 226. Italics removed.
[28] Bilgrami 2006, p. 277.

understood as commitment by Bilgrami, there is also reason to think that his self-knowledge relation could never be pointed in another direction. Think again of the game of capture the flag. Suppose that it was the job of one of the players to tie the flag in a tree; he was assigned or volunteered to place the flag. While everyone is capable of knowing who is supposed to tie the flag, only the person who is responsible for tying the flag can feel the pull of the responsibility. To have a commitment, on Bilgrami's account is to be prepared to accept criticism for it. If I am prepared to accept criticism for a commitment, as an agent, it seems I must feel constrained by that commitment. There seems to be little room for thinking another agent's commitments could constrain my actions in the same way, so even if I know what commitments he or she has, I would have to know about them in a different way.

At its core, Bilgrami's account makes self-knowledge special by making belief special. Since he makes having a belief a matter of being reasonably held accountable for commitments, his notion of belief is freed from its place in the causal story many theorists of mind give it. Normally portrayed as physically realized mental states, beliefs are thought to be defined, at least in part, by their causes and effects. For Bilgrami, physical and causal processes are not the most central or most important aspects of mental life. There is no doubt that the aspects of mental life he discusses are interesting. There is certainly a place in the philosophy of mind for the discussion of our ability to form commitments toward the truth of certain propositions and our practice of holding each other responsible for these commitments. The existence and formation of these commitments certainly raises issues about what aspects of mentality we have agential control over. However, I do not think that these issues have much to do with belief or the

epistemology of self-knowledge. Though it has been an important goal throughout the dissertation to bridge naturalist and anti-naturalist discussion of self-knowledge, some views remain a bridge too far.

6. Conclusion

Richard Moran's work has raised important questions about the connections between epistemic agency and self-knowledge. His criticisms of perceptual models of self-knowledge are not thoroughly persuasive, but they do suggest the possibility of overlooked alternatives. Based on the work of Moran as well as Akeel Bilgrami, philosophers concerned with self-knowledge need to be more attuned to the role that deliberation over our mental states might play in our knowledge of those states and vice versa. Even if one ultimately wishes to hold that much conscious deliberation about what to believe is misguided, it remains important to see what role our knowledge, or lack thereof, of our own states plays in our deliberations.

Chapter 6

The Low Road for Belief

> All Christians believe that the blessed are the poor and humble, and those who are ill-used by the world; that it is easier for a camel to pass through the eye of a needle than for a rich man to enter the kingdom of heaven; that they should judge not, lest they be judged; that they should swear not at all; that they should love their neighbor as themselves; that if one take their cloak, they should give him their coat also; that they should take no thought for the morrow; that if they would be perfect, they should sell all that they have and give it to the poor. They are not insincere when they say that they believe these things. They do believe them, as people believe what they have always heard lauded and never discussed. But in the sense of that living belief which regulates conduct, they believe these doctrines just up to the point which it is usual to act upon them… The doctrines have no hold on ordinary believers—are not a power in their minds. – J. S. Mill[1]

1. Introduction

While this passage from Mill could serve several purposes, my attention is drawn to a distinction Mill is making. The distinction is between a living, action-directing sort of belief and another kind of belief which he later calls a "dead dogma". I take it that Mill's purpose in drawing this distinction was to suggest a certain kind of stance for individuals to take with respect to the things they endorse as true. He thinks that we should all try to keep them "alive". Notice here that Mill accepts that the Christians that he discusses really do have the beliefs that they claim to have. I want to suggest another possibility. I want to suggest that people such as Mill describes are self-deceived.[2] When people profess to have a certain belief and act in ways incongruously with that belief, I

[1] Mill 1989, p. 43.

[2] Georges Rey has made a similar suggestion, with slightly different implications (See Rey 2007). He focuses on whether Christians can really believe the sort of metaphysical claims that they profess to believe. For instance, he thinks Christians may not really believe in a God that has the various properties Christian doctrine describes.

suggest that their non-ascriptive behavior reveals their beliefs. The self-ascriptions reveal something else. In the next chapter, I will suggest that the self-ascriptions reveal the meta-beliefs of the speakers. In this chapter, however, I want to focus on what beliefs individuals have in cases of apparent conflict between self-ascribed beliefs and behaviors.

In order to understand self-knowledge, one must examine cases in which self-knowledge appears to fail. What is required are cases in which people have beliefs they believe themselves to lack, lack beliefs they believe themselves to have, or somehow believe the wrong things about their own beliefs. I suggest that the case above, of Christians that do not act as if they believe the things they profess to believe, is such a case. No doubt this case will meet resistance. It would be nice to find a less controversial case of failure of self-knowledge to work with. However, all potential failures of self-knowledge admit some degree of controversy. While most theorists are willing to say that we are sometimes wrong about what we believe, it is hard to find a clear third-personal description of a case in which a person described is uncontroversially wrong about what he or she believes.

In my terms, a failure of self-knowledge is a case in which a person has a belief and a conflicting meta-belief. We can understand a belief and a meta-belief to conflict when the proposition believed contradicts the proposition embedded in the meta-belief.[3] Thus, if I have a belief that P and a meta-belief that I believe that not P, I have a failure of

[3] The situation is more complex than represented here. It is difficult to formulate necessary and sufficient conditions for the sort of conflict I have in mind. There are probably cases that should count as conflicts even though the contradiction is not explicit. There are also worries about content that I cannot address here.

self-knowledge.[4] In presenting a case which is supposed to be a failure of self-knowledge, one has to describe some behaviors of a person in such a way that it would be clear to all that the described person has a conflicting belief/meta-belief pair. The problem is that, as much as philosophers tend to like belief driven explanations, commonplace mental attributions appeal to more mental machinery than just belief to explain behavior. While most of us think that people do get into states of belief/meta-belief conflict, we tend to regard it as an unusual phenomenon. Other explanations may seem more natural for any particular case presented. A belief/meta-belief conflict might explain the Christians' failure to live up to ideals, but so might laziness (perhaps a desire to do what is easy), poor willpower, or a lack of reflection. Furthermore, given enough creativity, it is possible to assign nearly all behaviors—especially non-ascriptive behaviors—explanations at the level of first-order mental states. Indeed self-deception is sometimes framed in terms of a person having conflicting first-order states, one of which is hidden. Thus, though most theorists accept that it is possible for self-knowledge to fail in the way I have described, it is not clear exactly what such a failure would look like from the outside.

To get another example on the table, suppose you saw an animal trainer working with a dangerous looking animal that you cannot identify. The trainer tells you that the creature is completely harmless. Not knowing anything about the animal, you are inclined to take the trainer at his word. However, while working with the animal, he gives it a wide berth, moves his hands slowly and deliberately, and generally acts as if the animal is incredibly dangerous. One possible explanation for the trainer's behavior is that he

[4] While there may be more ways for a person's self-knowledge to fail, it is important to note that I am not counting certain belief/meta-belief arrangements as failures of self-knowledge. For instance, one might believe that P and lack the meta-belief that P, without counting as having a failure of self-knowledge.

believes that the animal is dangerous but he doesn't believe *that he believes* the animal is dangerous. That would be a failure of self-knowledge. However, you can probably think of other mental states that might explain the trainer's behavior. Perhaps he believes that the animal is not dangerous, but he is practicing for a show that will require him to present the animal as dangerous. Or perhaps he only said that it was harmless to try to calm himself down, despite believing that the animal is quite dangerous. Yet another possibility is that he believes the animal is harmless, but behaves as he does due to some non-doxastic mental state—perhaps he is irrationally *afraid* of the creature even though he *believes* it is harmless. We are typically so unprepared to ascribe a belief/meta-belief conflict that silly explanations of behavior—perhaps he enjoys pretending that the animal is a dragon—can seem more likely.

The upshot of these examples is that I do not believe I can offer a description of behavior which everyone will agree results from a belief/meta-belief conflict.[5] At least some of the difficulty comes from not having on hand an account of the scope of behaviors that meta-beliefs can influence. Typically, meta-beliefs and beliefs go together—or so we think—so the possibility exists for any particular behavior that either belief or meta-belief is the source. This can be especially confusing with linguistic behavior, since we tend to not only to credit a speaker with a belief matching any assertion but also assume that the assertion comes from or expresses that belief. So if I say, "it will rain today", then anyone hearing me would likely ascribe to me the belief that it will rain today. If I say, "I believe that it will rain today", the natural belief to

[5] A first person case might be offered, but I'm worried about first person *post hoc* explanations of behavior.

ascribe to me is a meta-belief.[6] Earlier, we have seen that some thinkers locate the security of avowals in the fact that they are expressions of first-order beliefs. I have said that I think this suggestion is not implausible in some cases of first-person avowals. However, another possibility exists. Sometimes, assertions of fact might express or issue from a meta-belief. If I both believe P and believe that I believe P, then my saying, "P" might be due to either to the belief or to the meta-belief. Figuring out the boundaries of influence for these states will have to be done in part by examining behaviors that might result from failures of self-knowledge.

I will refer to behaviors like Mill's Christians or the confusing animal trainer as *conflicting behaviors*. This label does not presuppose that the behaviors do in fact issue from a belief/meta-belief conflict. Instead, I mean only to refer to an individual displaying what appear to be two sets of behaviors that separately would be attractively explained by beliefs that conflict. For now, this can include behaviors that many would be tempted to describe a first-order conflict rather than a belief/meta-belief conflict. As discussion proceeds, the difficulties of determining which behaviors in a set of conflicting behaviors issue from that individual's *beliefs* will come into more focus. In the next chapter, I will revisit the issue of what behaviors can be caused by meta-belief.

Here, I want to explore the idea that a debate between naturalists and anti-naturalists about self-knowledge might boil down to disagreements about the nature of belief. In outline, I think that naturalists tend to have a bottom-up picture of belief while anti-naturalists have a top down conception of belief. What I mean is that naturalists tend to see human belief as a state developed by evolution and contiguous with states in non-

[6] Note that there is also the natural use of "believe" to distance oneself from a claim. I take it that even if the there is such a pragmatic implication in a large percentage of such self-ascriptions, not all such ascriptions are like this.

human animals. On the other hand, I think that anti-naturalists tend to have a picture of belief as distinctively human; they see it as something integral to distinctively human abilities and only (fully) realized by creatures with possession of those distinctly human abilities. Here, I have in mind philosophers that hold that language is required to have belief as well as philosophers that have even stronger requirements.

My argument in this chapter is that the top down approach, what I will call the high road to belief, is in danger of separating belief from explanations of our behavior in a way that would negatively offset whatever success it might have in matching our philosophical intuitions. That includes intuitions about the distinctive character of self-knowledge. I favor a naturalistic, bottom up conception of belief. I think that this low road to belief—in which belief is inextricably tied to more mundane behaviors—makes naturalistic models of self-knowledge more attractive.[7]

Many philosophers that have thought about the nature of belief have suggested a schism in the concept. While there are a variety of non-overlapping distinctions made by philosophers of different backgrounds, I think that the phenomena they seek to explain share a common structure. They all suggest that the belief is either a high road state or a low road state and propose another state or attitude to cover the other. That is, these philosophers have thought that we need both high road and low road states, but there is some disagreement about which sort of states should count as beliefs. In this chapter, I will survey a variety of these distinctions with the intention of arguing that the low road states are beliefs. Here, I leave it open whether there are separate states corresponding to

[7] Though the labels occurred to me independently, they echo the names LeDoux gives to the two pathways he and others have posited for emotional processing. It would be overstating things to assume a tight connection between my distinction and his, but I do think the distinctions are likely to have some connection. See LeDoux 1996.

the high road conception. In the next chapter, I will argue that there are separate high road states and that meta-belief plays a crucial role in their function.

If there really is a schism in the concept of belief, then all theorists will face a choice about which states to call beliefs. Such a choice might have dramatic consequences for the project of this dissertation. If there were, say, one sort of states that we had privileged access to and another that we did not, then whether we have privileged access to our beliefs will vary from philosopher to philosopher according to what kind of states they consider beliefs. Of course, as a naturalist, I am skeptical about any kind of privilege that would threaten naturalism, but for a naturalistic rejoinder to self-knowledge based challenges to naturalism to succeed, we must first be clear about exactly we are supposed to have privileged access to. If I can establish that there is no such access to beliefs, some progress will have been made.

While I acknowledge that I am setting this debate up to be partly terminological, I think the issue is deeper than terminology. Naturalists about the mind view it as progress when mental phenomenon can be explained in scientifically plausible terms. If there are aspects of self-knowledge—in the broad sense—that can be so explained, then it is crucial to progress that we identify those aspects and try to come up with those explanations. Drawing terminological distinctions can exaggerate difficulties if done one way and engender fruitful research programs if done another.

2. The Low and High Roads

In this section, I will examine the distinction between what I will call the high road and low road conceptions of belief and the evidence that both conceptions pick out real mental phenomenon. Later, I will argue for the low road conception of belief.

The word 'belief' is used across many sub-disciplines in philosophy. There is very little agreement on the nature of belief. Perhaps the least controversial claim that can be made is that belief is a kind of propositional attitude. Of course, there is little agreement among philosophers as to the nature of propositional attitudes or even on the nature of the propositions that belief and other attitudes are supposed to be directed towards. Another point of agreement is that beliefs can be used in explanations of human behavior. The picture is typically that humans represent the world to be a certain way via their beliefs and act to change the world from the way they believe it to be in accordance with desires that it should be different. The picture is of one unified system acting in accordance with one set of information and one set of goals. The same set of propositional attitudes is supposed to explain all of the individual's behavior. In practice, we are apt to ascribe beliefs to people on the basis of any sort of behavior that they perform.

It is also common practice to ascribe beliefs to animals on the basis of their behaviors. This point has divided philosophers. Some have held that only humans may legitimately be said to have beliefs. On the face of it, this claim is surprising because of the behavioral similarities between humans and non-human animals. If a person walked over to a bowl from which he or she had previous been seen to get food and, finding it empty, emitted a non-linguistic whine, we would not hesitate to say that that person believed there would be food there. If very similar behavior were performed by a dog,

many would likely also claim that the dog's behavior was explained by a belief. It seems implausible that there is no common cause of the dog and human behaviors. Any conception of human belief as something other than the kind of mental state that explains the dog's behavior faces a prima facie puzzle. How can the dog perform behaviors without beliefs that are caused in human beings by beliefs and other propositional attitudes? There may be answers to this question, but it must be answered by anyone that would deny propositional attitudes to non-human animals.[8]

Human beings also perform various behaviors that non-human animals do not. We use language to convey our thoughts, we ascribe mental states to ourselves and others, and we deliberate about what to believe. We explain these behaviors by appeal to propositional attitudes as well. Some philosophers, including many that deny animals propositional attitudes, play up the importance of these sorts of behaviors as essential to the possession of belief.

To take some examples from the extremes, compare Daniel Dennett to Donald Davidson in the quotations below. Dennett is famously willing to attribute intentional states even to systems less complex than animals. He writes:

> All there is to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy, and hence all there is to really and truly believing that p (for any proposition p) is being an intentional system for which p occurs as a belief in the best (most predictive) interpretation.[9]

Whereas Davidson thinks that language is essential for having intentional states. He claims:

[8] For instance, it may be that there is simply something additional in the human's mental states that qualify them as genuine beliefs even though they are largely similar to the mental states of the dog.
[9] Dennett 1987, p. 29. Italics removed.

> We identify thoughts, distinguish among them, describe them for what they are, only as they can be located within a dense network of related beliefs. If we really can intelligibly ascribe single beliefs to a dog, we must be able to imagine how we would decide whether the dog has many other beliefs of the kind necessary for making sense of the first. It seems to me that no matter where we start, we very soon come to beliefs such that we have no idea at all how to tell whether a dog has them and yet such that, without them, our confident first attribution looks shaky.[10]

Davidson has in mind an example due to Norman Malcolm of a dog that might be described as thinking that a cat went up a tree. Davidson stresses the implausibility of the dog possessing concepts like CAT and TREE and thus the implausibility of the dog having beliefs about cats and trees in general or *this* cat and *that* tree in particular.

Much has been written in favor of each philosopher's position and many philosophers have staked out positions in between. While I do not want to occupy a position as extreme as Dennett, I think it critically important to recognize some non-human animals as having beliefs. My primary reason is that I think we share mental states with non-human animals which cause them to behave in ways similar to us. Since we employ belief in explanations of even mundane human behavior, animals doing the same sorts of things because of the same sorts of states must also count as having beliefs. In order for our ascriptions of belief to humans to be even approximately right, we have to grant beliefs to non-human animals. I will draw out this argument in sections four and five.

In the next section, I review various attempts that philosophers have made to distinguish beliefs from similar mental states. I think that this will show some confusion over whether to take the low road or the high road in understanding belief. Examining the

[10] Davidson 1982, p. 98.

different attempts made by philosophers to handle what I call the belief schism should be illuminating for my purposes later on in the chapter.

3. The Belief Schism

Several philosophers have argued purely on conceptual grounds that the concept of belief is ambiguous between two distinct mental state concepts which do different work in establishing behavior (or at least that the word, 'belief' is often used to cover two different kinds of states) . While it would be nice to think that these philosophers have stumbled onto the true structure of the mind via *a priori* means, the truth is that the distinctions do not overlap perfectly and any of them may still turn out to conflict with the mind as eventually described by science. Empirical psychology has also begun to take seriously the idea that we may not have a singular reasoning system. Dual process and dual systems theories in psychology suggest the possibility that science will recognize something like the distinctions made by philosophers. Each distinction seems aimed at explaining how our talk about beliefs seemingly picks out different sorts of mental states at different times. In each case, there is a state which corresponds roughly to my low road and high road conceptions of belief. That is, in each case, there is a state which is largely held to be responsible for mundane behavior of the sort we share with non-human animals and another state that is held to be responsible for some distinctly human behaviors. What is interesting, however, is that there is no real agreement on whether the low road or high road states should be classed as beliefs.

3.1. Belief and Opinion

Daniel Dennett, admirably enough, has an explanation for why some philosophers disagree with him about beliefs and other intentional states. In response to worries about our ability to change our minds, Dennett makes a distinction between belief and opinion. He thinks that some philosophers are confusing beliefs and opinions when they talk about belief.

> Changes of mind are a species of judgment, and while such judgments arise from beliefs and are ultimately to be explained by one's beliefs, such judgments themselves are not beliefs—when such judgments are called occurrent or episodic beliefs, this is a serious misnomer—but acts, and these acts initiate states that are also not states of belief, but of something rather like commitment, rather like ownership… I suggest that we would do quite well by ordinary usage if we called these states *opinions*, and hence distinguished opinions sharply from beliefs.[11]

Dennett derives his distinction in part from the work of Ronald de Sousa, who distinguishes beliefs from acts of assent.[12] De Sousa stresses the difference between partial and flat out acceptance of a proposition. Dennett's distinction also incorporates this. However, Dennett makes language a central point of difference between belief and opinion. He writes, "Now why do we have opinions at all? And why do we have them while animals don't? Because we have language."[13] Dennett's idea is that opinions allow humans to collect true sentences while beliefs are a matter of weighted credence in propositions, which are non-linguistic entities.

There is one other feature of note with Dennett's distinction. Dennett thinks that opinions are voluntarily formed and represent the mechanism by which we change our minds about something. That is, opinions, but not beliefs, are under voluntary control. Interestingly, Dennett also suggests that an epistemology concerned with what one ought

[11] Dennett 1978, pp. 303-304
[12] De Sousa 1971.
[13] Dennett 1978, p. 306.

to *believe* is misguided and ought instead to be framed in terms of what one ought to opine.

3.2. Belief and Acceptance

Many philosophers have distinguished between belief and acceptance. In L. J. Cohen's version of the distinction, belief is a disposition to feel that a proposition is true and acceptance is taking a proposition to be true for reasoning and argument.[14] Acceptance is always seen as a flat-out notion, much like Dennett's opinions and is usually thought to be under direct control. One major difference is that acceptance can be used to cover cases in which someone willfully adopts a proposition for the purposes of action or reasoning with no regard whatsoever for its truth. For instance, a lawyer may accept that his or her client is innocent because of a responsibility to defend that client and without any regard for the fact of the client's innocence or guilt. Of course, the reason that one accepts a proposition can have to do with the propositions seeming true or with consideration of evidence. It merely need not be tied to a concern for truth.

Others have discussed acceptance, but the distinction has been drawn in multiple different ways.[15] Cohen explicitly states that he does not think he is using 'belief' and 'acceptance' as technical terms. He thinks he is clarifying the meaning of the words in standard non-philosophical usage, but his larger concern is the use of the concepts he delineates to make sense of a variety of philosophical puzzles.[16] For instance, he gives

[14] See Cohen1992.
[15] See Schwitzgebel 2010, entry 2.5.
[16] Cohen 1992, p. 4.

accounts of both self-deceit and *akrasia* as cases of belief and acceptance conflict with one state dominating the other.

3.3. Belief and Alief

Tamar Gendler has noted a distinction between belief and what she calls 'alief'. She provides a series of examples in which she thinks it clear that a person has a certain belief but performs some action or other that seems to conflict. For instance, she talks about walking across a very high transparent structure:

> How should we describe the cognitive state of those who manage to stride to the Skywalk's center? Surely they believe that the walkway will hold: no one would willingly step onto a mile-high platform if they had even a scintilla of doubt concerning its stability. But alongside that belief there is something else going on. Although the venturesome souls wholeheartedly believe that the walkway is completely safe, they also alieve something very different. The alief has roughly the following content: 'Really high up, long long way down. Not a safe place to be! Get off!!'[17]

This is an interesting sort of case. Many fears are such that they persist despite a person's attempts to believe that the object of fear is not really dangerous. Sometimes this is described in terms of intellectually believing or knowing something, but not believing it on a gut level. I am, for instance, moderately afraid of flying. Having a decent grasp on statistics, when confronted with the data about airplane safety, I am forced to admit that they are perfectly safe. Yet still, when I am in a metal box hurtling through the air at several hundred miles an hour, I feel some fear. The question is, what do we believe in these cases and what, if any, other sorts of mental states explain the apparent conflict? Gendler takes it for granted in all cases she discusses that the people described have the beliefs associated with intellectual understanding. However, she recognizes that there

[17] Gendler 2008, p. 635.

must be some sort of mental state with something like representational content responsible for what she calls belief-discordant behavior.

Interestingly, Gendler claims that cases of belief/alief mismatch are manifestly different than cases of self-deception. She writes:

> The mismatch runs two directions: unlike in cases of self-deception, the subjects in our cases show no reluctance to endorse explicitly the belief with which their behavior fails to accord. And unlike in cases of self-deception, their behavioral responses do not result from some deliberate or quasi-deliberate process of misrepresentation.[18]

The second point involves thinking of self-deception as a something done intentionally by the self-deceived. I discuss failures of self-knowledge rather than self-deception partially to avoid that connotation. The first point, however, is interesting. On her way of describing the cases, the people involved would have to think that they had a belief that conflicted with their behavior. The people going across the walkway would have to believe of themselves that they believed it was dangerous or at least not believe of themselves that they believe the bridge is safe. Gendler is right to think that this is implausible. However, it is worth noting that if we located belief where she locates alief it would be much more natural to describe the case as one of self-deception or at least as a failure of self knowledge.

3.4. Dual Processes and Dual Systems

While not as clear on the philosophical implications for mental states like belief, dual process accounts have recently covered the same sort of ground as the above philosophers. Dual process theories posit separate cognitive reasoning mechanisms which are each used in different situations. In recent years, some philosophers and psychologists

[18] Gendler 2008, p. 639.

have proposed that there are separate reasoning *systems* which carry out these separate reasoning processes. The difference between a dual system theory and a dual process theory is actually quite significant for my purposes here. Two distinct processes may use the same information to arrive at different results. That is, multiple reasoning processes can be recognized without generating the belief schism. However, if there are distinct reasoning systems, it is possible and perhaps even likely that those systems would have distinct stores of information. In short, the contention that there are both high road and low road states is much more plausible under a dual system theory than a dual process theory, though not excluded by a dual process theory and not entailed by a dual system theory.

Following Frankish and Evans, I will quickly review the history of dual process theories in psychology.[19] Frankish and Evans identify at least four different lines of research that arrived at something like a dual process theory on independent grounds. These theories were separately created to explain results related to observed differences in implicit and explicit learning, discrepancies between behavior and introspective reports in deductive reasoning, discrepancies between social behaviors and stated attitudes, and observed differences in intuitive and reflective decision making.[20] In each of these lines of research, subjects displayed behaviors that conflicted with verbal reports of the subjects mental states. In other words, they displayed the kind of conflict that I described above. That separate research programs arrived at the same sort of conclusion is a point in favor of dual process views. That the dual processes cover a range of cases of

[19] See Frankish and Evans 2009.

[20] For implicit and explicit learning, see Reber 1993. For deductive reasoning, see Evans 1977. For social attitudes, see Smith and Collins 2009. For decision making see Reyna 2004.

reasoning while showing similarities with each other is a point in favor of dual systems theories. I will use the research in social cognition as a representative example.

There has been a dramatic shift in explicitly endorsed attitudes by whites toward racial relations in the last century in the United States. Since Barack Obama has been elected president the popular media has asked with various degrees of seriousness whether the election signaled the end of racism in America. Of course it is obvious that explicit racism still exists in individuals, but the hope of many seems to be that racism is now a small minority position and that social attitudes have changed such that racism as a dominating factor in the lives of most black Americans is or would soon be at an end.

Timothy Wilson summarizes the shifting attitudes from in the later part of the twentieth century as follows:

> In 1942 only 2 percent of southerners and 40 percent of northerners believed that whites and blacks should attend the same schools, whereas by 1970 these percentages had increased to 40 percent and 83 percent respectively. In a 1997 Gallup poll, 93 percent of whites said they would vote for a qualified black candidate for president, compared with 35 percent in 1958. Sixty-one percent said they approved of interracial marriage, compared with 4 percent in 1958.[21]

It seems fair to say that racism is no longer the social norm that it once was. However, alongside these stated attitudes, social psychologists have made some interesting discoveries. While there have been undeniable improvements in opportunities minorities enjoy, many individuals that profess acceptance of racial equality still display a pattern of seemingly racist behavior. This behavior is implicit and unconscious: engaged in despite an apparently conscious rejection of racism. Psychologists have been rightly concerned with determining the patterns of such behavior and on coming up with methods to disrupt the patterns to generate individuals that live up to the egalitarian ideals they endorse.

[21] Wilson 2002, p. 188.

However, what is sometimes noted but not often enough addressed is that the phenomenon is difficult to describe in terms of the psychology of the individuals that display the behavior.

A recent survey of the research on this implicit and unconscious form of racism provides some idea as to the types of behaviors displayed.[22] Much of the research to date has been about the attitudes of white Americans toward black Americans. While there are still overt racists that publicly and privately espouse non-egalitarian attitudes, the behaviors I will describe below have been systematically demonstrated in a significant portion of whites that otherwise appear to be sincere in their claims that they believe blacks and whites are equal.

Standard labels for the following behaviors are implicit racism or aversive racism. It is called aversive because people that display these behaviors avoid blacks and situations in which their commitment to racial equality might be tested. This contrasts with more overt forms of racism in which racists are quite comfortable getting into the similar situations and displaying their racist attitudes. Aversive racists display many signs of discomfort when in close proximity to blacks. They avoid eye contact and act nervous. Aversive racists are much more likely to help whites than blacks when they appear to need assistance. Interestingly, aversive racists typically fail to help blacks by avoiding a request for help rather than rejecting the request (as more overt racists do). In situations in which aversive racists have some measure of control over the opportunities of blacks, they are much more likely to deny them opportunities. For example, aversive racists will award jobs or housing to white candidates but not black candidates that are otherwise

[22] See Dovidio and Gaertner 2004. See also Wilson 2002, pp. 188-194.

identical. All of these behaviors are done seemingly unconsciously and subjects are likely to deny that they engage in them.

The simplest explanation for the apparent clash of attitudes and behaviors would be that subjects were lying; claiming to endorse popular views that they do not really share. However, this explanation is challenged by more than the fact that aversive racists appear to be sincere in endorsing racial equality.[23] It is challenged by the fact that, when race is made salient to the aversive racist, he or she will typically not only cease displaying many racist behaviors, but demonstrate behaviors that seem best understood as *over*compensating. For instance, whereas they might avoid assisting normally, once the race of the person requesting assistance is made salient, an aversive racist will provide more assistance for a black person than they would help a white person in the same situation. The most compelling explanations of these behaviors have proven to be dual process theories.

Social psychologists have also, in a completely separate line of research relied on dual process theories to explain results related to persuasion. Researchers have found that in certain situations, people seem to respond more to the way a message is delivered and the context in which it is received than the content of the message. Yet, in other situations, subjects seem able to ignore the influences of these factors in favor of evaluation of the content. The preferred explanations of this sort of phenomenon have been dual process theories. Smith and Collins summarize as follows:

> "For example, someone might agree with a persuasive message delivered by an attractive, likable source simply because of the positive feelings aroused

[23] Apparent sincerity is pretty strong on its own. Aversive racists, for instance are markedly different from members of the KKK, who sometimes present egalitarian views to save themselves embarrassment in public life and admit their real beliefs in private rallies. An aversive racist would be thoroughly horrified at a KKK rally.

> by the source, or agree with a majority's position because in the individual's past experience, majorities have usually turned out to be correct. Heuristic processing is the default processing mode, but people perform *systematic* processing when they have the motivation, time, and cognitive capacity allowing for more effortful processing. This involves the active, effortful scrutiny of all relevant information, requiring considerable cognitive capacity and potentially leading to more enduring attitude change. Systematic processing (to the extent that it occurs) is assumed to take place in addition to and simultaneously with heuristic processing, rather than replacing it.[24]

In short, the most fruitful explanations of how people are persuaded by various messages involve distinguishing between processes that are automatic and heuristic and processes that are active and systematic. Aversive racists can also be understood as reasoning differently at different times, tending toward racist behaviors when relying on automatic and heuristic processes and being overtly egalitarian when engaging in active systematic reasoning.

It has gradually become clear that the characterizations of reasoning processes in various dual process theories were remarkably similar. This is true across disciplines in psychology as well as across research programs in social cognition. While there is no clear consensus as to which features each process has, it is a reasonable gloss to say that one set of processes can be described as typically non-conscious, heuristic, and passive, while another can be described as roughly conscious, systematic, and active. Given the similarities between processes, some theorists have begun to think that there are actually two systems which can be employed for a variety of processes. Following Stanovich, theorists that posit these systems call them System 1 and System 2.[25] System 1 is the unconscious automatic system and System 2 is the conscious active system.

[24] Smith and Collins 2009, p. 198.
[25] See Stanovich 1999.

It seems likely that each system would use different information. For instance, the research on aversive racism above seems to make the most sense if certain groups are represented via some sort of unconscious prejudicial stereotype in addition to some sort of consciously egalitarian commitment. In a sense, each system has a different state playing the belief role, where "belief role" means roughly the state which stores information to be used in determining a course of action. When people's actions appear to be based on some information, we attribute to them a belief with that informational content. If a dual system theory of mind is correct, the common usage of the word belief may be ambiguous between System 1 and System 2 states because different states may influence different behaviors at different times. There is currently a flurry of work being done to try to characterize the systems and describe a mental architecture that would support two systems.[26]

There are worries surrounding the coherence of a dual systems view of the mind and development of any such view is tied to the active research program surrounding them. Though I will propose a view that has much in common with dual systems theorists and probably counts as a dual systems view, my only concern here is to point out the similarities between dual systems theories and the philosophical proposals surrounding beliefs. Psychologists advocating dual-process theories are, I think, covering the same sorts of cases as the philosophers mentioned above. Unfortunately psychologists seem even more split than philosophers about the role of belief in these cases, with some thinking that beliefs are system 1, some system 2, some both, and none clear on exactly how belief fits in.

[26] Frankish 2009 and Carruthers 2009.

3.5. Mind and Super Mind

Recently, Keith Frankish has been explicitly advocating for a dual system model of the mind. Interestingly, though he thinks system 1 and system 2 make use of different states, he is willing to call both states beliefs. He argues that there are two strands in the folk-psychological concept of belief. His division is roughly as follows:

> Strand 1 Beliefs: Representational states that are primarily formed passively and unconsciously, feature degrees of confidence in non-linguistic content, and result in functional dispositions to behave or form other mental states.
>
> Strand 2 Beliefs: Representational states that are primarily formed consciously and actively, are flat-out bets on the truth of certain linguistic contents, and result in an occurrent state involved in conscious action.

It is hard to assess whether Frankish is right about the folk-psychological concept of belief. In some of these cases, one can imagine the same state being at various time each of these things—as in the conscious/unconscious or occurrent/dispositional dimensions. With respect to those properties, it is not clear to me that folk psychology would recognize different states. Others features listed can appear as disagreements about the nature of belief—as in the flat-out vs. partial debate. Even, if he is right about both the concept of belief and the mind having two systems I worry that the folk-psychological concept is hopelessly conflating distinct states. It seems to me that if a concept conflates two things that need to be distinguished, it is a faulty concept.

However, Frankish also realizes the need to carefully delineate the different states he claims are picked out by different strands in our concept. For him, System 1 is the mind and System 2 is the supermind, built on top of and from components in System 1. System 2 states are superbeliefs and differ quite a bit from beliefs. His superbeliefs are actually a variety of acceptance, and, he thinks that these states can fulfill Strand 2 of a

folk-psychological concept of belief.[27] Regardless of whether this is correct, it is clear that Frankish thinks that the System 1 states deserve to be called beliefs.

3.6. Takeaways

We have seen that philosophers have made a variety of suggestions for separating out low and high road notions of belief. What we have not seen is any agreement over whether belief is a high road or the low road state. Dennett, in talking of opinions, and those philosophers that discuss acceptance would take the low road for belief. Gendler would clearly take the high road and pass off the low road to her aliefs. Frankish, despite his sensitivity to this sort of question, actually hedges quite a bit, saying that there are two types of states, but that both fall under different strands of the folk-psychological concept of belief. In short, different philosophers have embraced distinctions between different states that they take to be mistakenly lumped together as beliefs. However, these philosophers have shown little agreement over which of these systems states are beliefs. In the next section, I will argue that for philosophy of mind and psychology, it makes sense to consider low road states to be beliefs. For now, I leave it open whether someone like Dennett is right that there is a non-belief state behind our more complex behaviors or whether someone like Frankish is right that there is some way to distinguish kinds of belief. My point of disagreement is with philosophers such as Gendler that want to discount the more basic states as genuine beliefs.

4. Low Road Beliefs

[27] See Frankish 2004, especially chapter 5.

I do not want to argue that it is essential to the concept of belief that beliefs be part of explanations of behavior, though, for what it is worth, I find that a plausible idea. I am not concerned with how folk psychology characterizes belief. Rather, I want to argue that belief, as a theoretical concept for psychology and the philosophy of mind, should cover at least those states that explain the mundane, low road behaviors shared by humans and animals. In part, I hold this due to thinking that both psychologists and the folk will use the word 'belief' to pick out mental states that explain human behavior no matter what philosophers do. Thus, I am concerned somewhat with the theoretical role played by belief in folk psychology; to the extent that I care about the role of belief in folk psychology, I care about what it is employed to explain. My thought is that the word 'belief' will be—and should be—employed in empirical psychological theories to do what it does in folk-psychological theories. By this, I mean that psychologists will be looking for belief based explanations of human behavior and will in all likelihood use the word 'belief' to pick out some of the states countenanced in their explanations. If philosophers were to insist that nothing in the mind as described by a mature empirical psychology could be properly called belief because nothing described had some property or properties essential to the folk concept of belief, psychologists would rightly brush aside their concerns and continue using the word.

Of course, the mind might turn out to be so very different than the folk-psychological concept that even the word 'belief' is abandoned by psychologists the way the word 'phlogiston' was by chemists. However, despite much worry over eliminativism in the literature, empirical studies of the mind have largely continued to use common mental terms and theorists seem to view themselves as studying the very states and

processes that the folk have long discussed.[28] Eliminativism seems unlikely to be broadly accepted outside of philosophy, even if it eventually has a resurgence within it. If I am right and psychologists continue to use terms like 'belief' come what may, philosophers will have to choose between insisting on their individual conceptual analyses and ease of communication with psychologists. At some point it becomes pointless to fight for the use of a term on conceptual grounds. However, I do not think that the issue is purely terminological.

In the subsections below, I discuss belief based explanations for human and non-human animal behavior. The ultimate goal here is to establish that belief based explanations of behavior are appropriate for a wide range of behaviors. I argue that such explanations are appropriate for non-human animals as long as their behaviors require information sensitive representations to explain them. Similarly, I argue that we often employ belief-based explanations for human behaviors when all we have evidence for is information sensitive representational states. I argue that this is adequate in the case of humans as well. This means adopting a low road conception of belief.

4.1. Animal Beliefs

We have already seen philosophers such as Dennett and Davidson disagree over whether or not animals have beliefs. Although much empirical work will have to be done to specify which, if any, animals other than humans can perform various complex behaviors, it seems fairly safe to suppose that many non-human animals have minds. It seems that we need to posit at least that much to explain animal behavior. While some behaviors may be reduced to simple stimulus-response mechanisms, much animal

[28] See Churchland 1981 and Stich 1983.

behavior can only be explained in terms of acquired representations bearing information about the animal's environment. As Hilary Kornblith puts it:

> The elaborate behavior of ravens in distracting a hawk so as to steal her egg is not a simple response triggered by some environmental condition. While the behavior is straightforwardly explained by appealing to beliefs and desires, no one has ever offered an explanation of such complex behaviors in terms that obviate the need for representational states. Nor is the case of stealing an egg an unusual one in the animal behavior literature. What we see is a wide range of behavior, in a wide range of different species, that has straightforward explanations in terms of beliefs and desires, and no competing alternative explanations. In circumstances such as these, a reluctance to endorse the available explanatory scheme does not seem cautious; rather, it seems unmotivated.[29]

In short, the theory that animals have beliefs is useful and no reasonable alternative explanation has been offered.

One might object that attributions of beliefs can be useful in the absence of alternatives without making it the case that the thing discussed actually has beliefs. For instance, Davidson points out that, "someone might easily have no better or alternative way of explaining the movements of a heat-seeking missile than to suppose the missile wanted to destroy an airplane and believed it could by moving in the way it was observed to move."[30] The missile, *pace* Dennett, does not really have such beliefs, so why think that animals must?

The trick here is that the missile's movements can be explained in non-representational (or at least non-mental) terms, with the proper understanding and willingness to elaborate the details. We know what those explanations look like. Kornblith's point is that there is no corresponding non-mental explanation for the sorts of animal behaviors he outlines. So some non-human animals must, while the missile need

[29] Kornblith 2002, p. 42. Foot note omitted, but see Allen and Bekoff 1997 for more discussion.
[30] Davidson 2001, p. 101.

not, have some mental representations. The issue is whether to think of these mental representations as beliefs and desires or in some other terms. To be cautious—more cautious than Kornblith thinks is warranted--one might grant that the sorts of animal behaviors mentioned above result from the animals having minds, but withhold judgment about the structure of their minds. Perhaps the mental states of these animals fall short somehow of being genuine beliefs.

I think that there is a good case to be made for regarding these states as genuine beliefs. As I noted above, there is no consensus on the necessary and sufficient conditions for a state counting as belief. I want to argue that many non-human animals have representational mental states which causally influence their behavior and are sensitive to acquired information. To be clear, I am not arguing that any of these properties is necessary for a state's being a belief or that these properties are sufficient for a state counting as a belief. I am not attempting a conceptual analysis of the concept of belief. What I do want to argue is that, given these properties, we will have to appeal to the mental states of non-human animals to explain their behaviors in the same way that we appeal to beliefs to explain the behaviors of human beings. Given the common structure of the psychological explanations of the behavior of both humans and non-human animals, there ought to be at least a prima facie case for regarding the states posited by those explanations to be fundamentally similar.

That some non-human animals have representational mental states is not widely questioned. A more live issue is how far down the phylogenic chain we would have to look to find organisms without metal representations. At the far end, single-celled organisms, plants, and insects, for instance, seem implausible candidates for mental

representation. Though some species of these organisms can respond to environmental stimuli in ways complicated enough to warrant the term 'behavior', none requires a representational explanation.

Jerry Fodor has proposed that the dividing line between representational systems and non-representational systems is their ability to respond selectively to non-nomic properties.[31] Nomic properties, or properties that enter into lawful relations, can be, in theory at least, detected and responded to by non-representational mechanisms. For an organism's behavior to be influenced by non-nomic properties, by contrast, the organism must somehow generate a representation of that property. Consider the property of being a dangerous object. Presumably, there is no simple mechanism that can be applied to get a device to respond to all dangerous objects. Organisms that can represent objects as having that property, however, can behave in a lawful way toward objects that it represents as dangerous.

I for instance, might behave in a very similar manner toward any number of devices that I represent as bombs. You just could not have a simple detection mechanism responsive to bombs—not to all bombs, at any rate. A bomb can look like anything and can be made out of a wide range of materials, including a wide range of different explosive materials. But the property of being a bomb can fit into laws describing my behavior via the intermediary of my mental representation of that property. I can even come to represent that some object is a bomb (and run away from it) when it does not look anything like any bomb I've ever seen or thought of.

[31] See Fodor 1986.

Fodor's distinction is certainly close to the correct distinction, though I have no vested interest here in whether he gets the details exactly right. The important point is that representational states are required to respond to properties that cannot be responded to more directly via simple detection of that property. The complexity of non-human animal behavior is more than sufficient to conclude that animals have representational states. While it is quite difficult to fit linguistic terms onto the representations of animals, it is clear that they do have states that represent.

I have already argued that non-human animals display behaviors caused by representational mental states. To that point, I add the fact that non-human animals are able to update at least some of these representations in light of new or changing information. This point includes the obvious abilities to learn changing features of their environments, but also much more complicated learning. For instance, Eric Saidel has argued that we should attribute belief to animals that can learn to, "form new associations of goals and the means to achieve those goals," and that we have ample evidence for this in some non-human primates.[32]

So non-human animals can have representational mental states that change in accordance with information they are exposed to. What more could we require before granting that these animals are capable of belief? I think it would be unmotivated to require the full range of behaviors and abilities that human beings can display. However, I can think of several possible requirements that do not seem out of the question. One commonly agreed upon feature of belief, as noted above, is that beliefs are supposed to be propositional attitudes. So we should consider whether the mental states of non-human

[32] See Saidel 2009.

animals can be rightly thought of as propositional attitudes. Another possibility is that human belief revision is more complex than that of non-human animals, involving reflection and deliberation.

It is beyond the scope of this dissertation to explore the nature of propositions or propositional content in any detail. However, I think that the issue of whether non-human animal mental representations are propositional attitudes can be separated into a few sub issues. One issue is whether these mental states have propositional content. Another is whether these states can rightly be called attitudes. Both issues are complex. I cannot say anything here that will be especially convincing in favor of thinking that non-human animal mental states have propositional content. For what it is worth, I am not overly concerned with the issue because I think it will be easily satisfied on some accounts of propositional content and that any account ruling out such states as having propositional content will threaten the propositional content of human belief as well. I am more concerned about the second issue.

Whether or not the mental states of non-human animals can rightly be called attitudes is also somewhat complex. A traditional gloss on belief is that it is the holding true of a proposition. Above, I conceded that it is not clear whether non-human animal mental states have propositional content. So it is not clear that a non-human animal mental state could be a holding true of a proposition. However, there is a clear sense in which non-human animals hold the contents of their representations as true. Namely they act as if their representations of reality corresponded to reality. This suggests that their representations are the right sorts of attitudes to count as beliefs. Perhaps something more is required. In particular, perhaps we should require that for an animal to hold something

true in the sense required for belief, it must be sensitive to some sort of reasoning process. I argued above that non-human animal representations are sensitive to acquired information. As far as that goes, I think it adds to the idea that animals are capable of the proper attitudes. However, if the reasoning process is meant to be something more self-conscious or deliberate, then the issue moves to the potential of reflection or deliberation as a requirement for belief.

Since it is clear that no non-human animals have anything remotely as advanced as the natural language capabilities of humans, it is clear that no non-human animals can undergo a process of reasoning quite like the explicit linguistic reasoning that human beings are capable of. But it is not clear how much reasoning non-human animals can do. Certainly there is experimental and anecdotal evidence that some non-human animals can problem solve. But it is not clear how similar their problem solving processes are to the processes of human reasoning. It is beyond our current understanding to speculate on the experiences of non-human animals in problem solving. Thus, non-human animals can certainly meet some sort of minimal reasoning criteria, but it is just not clear how much reasoning or deliberation they engage in.

In summary, I think that there is a good prima facie case for thinking of some non-human animal mental states as beliefs. I think that that case is made all the stronger by considering the likely overlap between human and non-human animal mental states. In the next subsection, I will be discussing human behaviors that seem to result from states that meet at least my minimal criteria for belief, but might not meet the requirements of some philosophers to count as beliefs. Since we tend to use belief ascriptions in explanations of behaviors of this sort, I argue that philosophers taking the high road risk

undermining the role of belief in human behavior. In short, if beliefs are distinctively human, they probably do a lot less than we think they do.

4.2. Mundane Human Behavior

The purpose of this section is to explore belief ascriptions in the behaviors of mundane human behaviors. While belief based explanations of behavior are extremely common in everyday human life, considerations like those above might prompt the same sort of worries about our attempts to explain human behavior that we had about explaining the behavior of simpler mechanisms. That is, we may appeal to psychological explanations as convenient and understandable explanations even when there are more simplistic or mechanistic explanations that better explain the behavior.

We appear to be predisposed to understanding mentalistic explanations of behaviors. Long before a child can understand the mechanisms involved a thermostat, he or she can be told that it's a box that changes the temperature whenever it *believes* that the temperature is too hot or too cold. That will give the child a pretty solid basis for predicting the behaviors of the heating and cooling systems. For Daniel Dennett, predictive success is all that matters for belief, even in the case of a thermostat. The possibility that I want to consider here is our talk of beliefs as explanations for ordinary human behavior is sometimes similar to the explanation of a complicated mechanism to a child. Humans, of course, do have genuine beliefs and desires and those beliefs and desires can factor into our behavior. I am not considering a global eliminativism or disregard for belief. Rather, I am proposing that we have adopted belief-level explanations for a wider range of human behavior than is actually warranted.

I am proposing the possibility of an overreach in our belief based explanations with some level of sympathy for it. Over intellectualizing is a common problem in human inquiry into human nature. I suspect we do sometimes attribute complex thoughts and motives to ourselves and others when there are simpler explanations to be had. However, I ultimately want to reject this idea. I want to defend the claim that many of our mundane behaviors are to be explained in terms of our beliefs, just as common discourse suggests.

As with non-human animal mentality, I think that the fact that mundane human activities are controlled by representational states which are sensitive to information should be enough to consider those activities as explained by beliefs. Part of my reasoning is that I think the nature of belief is a bit up for grabs. We can allow some revision to any pre-theoretical—or overly theoretical—conceptions of belief in order to preserve its behavioral explanatory role. Consider a simple case. Suppose I want a glass of water. I get up, make my way to the cabinet, open it, take out a glass, fetch a couple of ice cubes from the freezer, and some water from the tap. How should we understand these behaviors?

The shortest and most likely explanation to be offered would be that I got a glass of water because I was thirsty. As slightly more mentalistic answer might be, as I suggested, that I got a glass of because I wanted one. This desire can only work to produce my behavior, however, if combined with the right beliefs. And in this case, it seems that there are a number of beliefs that must be factored in. I have to have beliefs about the locations of the glass and the ice cubes. I have to have beliefs about the faucet as a source of water and the nature of its use. These beliefs control and explain my

behavior. Any shorter references to my thirst or my desires are, at best, shorthand for those states working in concert with my beliefs.

It may be objected that I have these beliefs only tacitly or implicitly or that they should not count as genuine beliefs for some reason. I point out that without acquired information, it would be impossible for me to perform these behaviors. I learned all about glasses and ice cubes and faucets at some point in my life and I must have stored mental representations of that knowledge in order to make use of that information. Further, unlike other cases which I will consider later, I would claim to believe these things, if asked, and I would, if reflecting, attribute my behavior to these beliefs.

I take our explanations of behavior such as this to be fundamental to belief. I am not claiming that these behaviors are part of the essential properties of belief. Rather, I am claiming that, whatever else belief is or does in the human mind, it regulates our behaviors. If we deny explanations of behavior such as the ones above, we risk a radically revisionary view of the human mind, with unnamed information bearing representations playing a crucial role in our behaviors and belief playing some other role, if any. I find it plausible that these information bearing states are the same sorts of things found in non-human animals and I am inclined to call them beliefs in order to have some way of explaining why humans or non-human animals do the things that they do.

4.3. Humans and low road beliefs

Suppose the careful cognitive ethologist comes up with a new label for animal states that allow for psychological explanations of animal behavior without anthropomorphizing prematurely. In practice this would be very difficult. This would separate researchers in cognitive ethology from their commonsense understanding of

mental causes and effects in trying to understand animal minds. At best, it seems likely that researchers would provide new labels for the animal states but rely on a presumed similarity between those states and the beliefs and desires of humans to attribute them.

If the consequences for the study of animal minds does not generate enough worry to pressure high road minded philosophers to grant that animals have beliefs, worries about explaining human behavior should. Under this suggestion, given the ever increasing appreciation for the overlaps between human and non-human animal minds, behaviors common to animals and humans might cease to be genuinely explainable in terms of beliefs and desires. Suppose, for example, that it is not a raven, but a human being, attempting to distract a hawk to steal an egg. If you like, suppose that this human being has become lost in the woods and is worried about starving or perhaps this person is a scientist needing to study the egg. For either the human or the raven, suppose, we have only non-verbal and non-linguistic behavior to support our mentalistic explanations. No one would hesitate to describe the human being as believing the eggs were in such and such a location or desiring to fool the hawk. Why is this safe for the human, but not for the raven given the similarities of behaviors?

Of course, one might say that we know human beings are capable of having beliefs and we do not know the same for ravens. This misses the point. If we have inadequate evidence for the raven's behavior being produced by beliefs, we seem to also have inadequate evidence for a person's behavior being caused by a belief. No doubt a human being might have produced those behaviors via a certain set of beliefs and desires. Yet, if we consider some alternative to belief, say anilief, to explain the raven's behaviors, then aniliefs may also explain the man's behaviors.

Sometimes when we act, we act deliberately with accompanying experiences of elaborate, conscious reasoning. But we do not always act like this. Sometimes, we act without conscious reasoning, without any inner monologue or visualizations, without framing the problem at hand in linguistic terms. Yet, for all this lack of reflection, we still manage to do some pretty complex things, such as distract a hawk to steal an egg or drive to work during rush hour. If we associate beliefs only with self-conscious and deliberate linguistic reasoning, we face a serious question about whether we should attribute our mundane and unreflectively produced behaviors to beliefs. This problem is made all the more salient if one is already positing some non-belief but belief-like mental state for non-human animals.

There are good reasons for thinking that humans and non-human animals, especially mammals and most especially other primates, have minds similar to humans. I do not think one has to tell a complicated just-so story about the evolution of animal minds to argue that the minds of humans and non-human animals have all evolved from the simpler minds of our genetic ancestors. Of course, human minds have evolved to a level of complexity beyond any other currently living animals, at least in some respects. So there is room to think that beliefs did not enter the picture until the later stages of human evolution. But evolution does not typically throw away working solutions to problems. If early ancestors of humans had minds comparable to the minds of contemporary non-human animals, with belief like states that enabled their complex behaviors, there is no obvious reason we would have lost those states in favor of beliefs. Rather, there is every reason to suppose that we still have states of that sort in common with non-human animals. If humans have more complex abilities, those abilities may

result from evolutionarily new mental states and processes. If beliefs are evolutionarily new in this sense, then the most reasonable thing to conclude is that they are only responsible for distinctly human behaviors, leaving the mundane behaviors we share with non-human animals to be caused by more primitive mental states and processes that we mistakenly conflate with belief.

This picture is not entirely without merit. As I have pointed out, there is a potentially high cost to be paid in undermining our belief-based explanations of much human behavior. However, the cognitive ethologist ultimately has to explain why human beings can talk, carry out complicated plans, or engage in sophisticated reasoning. I am sure many would find it intuitive to think that our having beliefs, and non-human animals lacking them, explains our abilities relative to theirs. After all, when a human being says, " I am going to throw this rock into that tree to distract the hawk", we tend to view those linguistic behaviors as being caused by beliefs and desires just as much as we do the behavior of throwing the rock.

The key point is that if this structural picture is correct, one in which evolutionarily old mental states operate in non-human animals with some additional newer states in humans, there is a worry about where to assign the beliefs. It is difficult to see why the new states would have taken over jobs done by the old states in humans. But if those older states are not beliefs and they are still active in humans, then those inclined toward the high road run the risk of invalidating commonplace explanations of behavior. Those that prefer the low road for belief have, it seems to me, a variety of more attractive options.

One possibility is that beliefs and desires operate slightly differently in human beings than they do in other animals because we have sophisticated abilities. Perhaps language use as a skill is entirely separate from belief and beliefs combined with language can explain why we can do more than other believers. Or perhaps there is some other cognitive mechanism or mechanism that was added that explains the increased uses humans have for their beliefs. This sort of strategy would recognize non-human animal belief and deny that new belief-like states have evolved in humans. In essence, it would preserve our current explanations without positing too many mysteries, though there are admittedly missing details.

Even if one accepts that there are evolutionarily new belief-like states in human beings separating us from other animals, the low road theorist is, I think, in a better position with respect to our ordinary explanations of behavior. There is the possibility of admitting the new states as a different kind of belief or as beliefs with a structure not possible in non-human animals. Ultimately, I will propose something like this strategy. In the next chapter, I argue that meta-beliefs, found in humans but not in (most) non-human animals, yet structurally similar to the sorts of beliefs that non-human animals do have, are essential for some of our distinctively human abilities. The high road theorist cannot help himself or herself to the same strategies. Any attempt would come out as understanding the older states to be beliefs in a metaphorical way, and our talk of mundane human behavior being explained or caused by beliefs would be made out to be as metaphorical as our ascriptions of belief to a thermometer.

Of course, it might turn out that there are older and newer states in humans and that the older states are too dissimilar to the newer states to warrant calling both beliefs.

This last possibility is the one which causes the clearest dilemma. As I have argued, the high road theorist would assign the term 'belief' to the newer states and potentially lose explanatory power over mundane behaviors. In this case, the low road theorist would assign the term 'belief' to the older states. This would result in a similar loss of explanatory power. In my assessment, having to posit new mental states to explain distinctively human behaviors would be a smaller disruption to our mental explanations for behaviors. No doubt, some philosophers would disagree with me on this point. However, I also find the low road theorist's other options attractive. I hold out hope that recognizing the complex behaviors of non-human animals as being caused by their beliefs will enable us to stick pretty close to our current explanatory picture of mental causation.

5. Low Road Beliefs and Self-knowledge

The low road conception of belief complements a naturalistic model of self-knowledge. In previous chapters, it has been pointed out that naturalistic models of self-knowledge admit the possibility of systematic error and undermine the intuitions that self-knowledge is direct. If one takes the high road for beliefs, these can seem like fundamental problems. However, if we view human beings as having the same sorts of mental states grounding their mundane behaviors as non-human animals, it becomes quite plausible that we could be systematically wrong about these sorts of states. Low road states often operate below the level of conscious awareness and they are likely outside of conscious control. A good deal of the pressure that anti-naturalists put on naturalistic models of self-knowledge disappears when one takes the low road for belief.

This is not to say that taking the low road removes all mysteries. No doubt many would respond to my comments here by saying that I have simply moved the interesting problems of self-knowledge. Even if human beings have unconscious beliefs controlling a portion of their behaviors, some behavior, it seems, is controlled deliberately in light of our conscious attitudes. Making low-level states into beliefs seems to make naturalistic models of self-knowledge plausible models of our knowledge of these low road states. But there are questions equally deserving of examination which we also might think of as problems of self-knowledge. For instance, when our behaviors are consciously controlled, what sorts of mental states are active? Are they beliefs of some form? If so, one might think those beliefs are the ones that demonstrate the special features of self-knowledge. Even if not, I might be accused of some sleight of hand. Many, if not all, of the theorists I have been arguing against have some sort of high road conception of belief. If their conception locks onto something mental, even if not to belief, there are genuine concerns surrounding our knowledge of those states. This line of criticism is not entirely unfounded, but there are several things that I would like to point out to deflect it somewhat.

First, if we can model any part of our self-knowledge with confidence, we have made progress. If we knew, for example, we had an inner scanner that registered our first-order beliefs and outputted meta-beliefs, we would be in a position to ask real questions about the successes and failures of this mechanism. Discovering patterns of error could provide a great deal of insight into how such a scanner works, how it evolved, and how to improve upon it, both first-personally and third-personally. So what if there are remaining puzzles about our access to other states and processes of our mind? In science and

philosophy, great progress is sometimes made simply by being clear about what the big mysteries are and shifting smaller mysteries onto bigger ones.

Second, it is possible that progress in understanding self-knowledge in the expanded sense has been held up by the continued conflation of low road belief with the sort of complex state picked out by the high road conception of belief. If the high road does lead to a real sort of state, acknowledging a separation between these states and low road states might be very useful. One possibility, which deserves some exploration—though I am not especially optimistic about it—is that these high road states display a genuine infallibility and that we have only thought we could be wrong with respect to them by conflating them with beliefs. I can see some intuitive appeal, for instance, in holding that conscious states are infallibly known and that those conscious states which we unreflectively call beliefs display the sorts of features that anti-naturalists have been mistakenly attributing to all beliefs, even unconscious ones. Many of the proposed divisions between System 1 and System 2 states also hold promise. It should be no surprise that our knowledge of unconscious, automatic, heuristic processes and the states that underlie them is different than our knowledge of conscious, voluntary, systematic processes and the states that underlie them. If the puzzles surrounding self-knowledge of high road states can be located in their being conscious, or being voluntary, then we will have reduced one mystery about the mind to another.

In the final chapter of this dissertation, I outline a new model of self-knowledge. This model is, first and foremost, a model of our knowledge of beliefs conceived of as low road states. However, I do not ignore the questions raised above about the nature of high road states. I propose that at least some of our distinctively human abilities are

influenced by our self-knowledge and that meta-belief plays a role in conscious reasoning and some linguistic behavior.

Chapter 7

A New Conception of Self-Knowledge

1. Failures of Self-Knowledge

I began the previous chapter with the claim that understanding self-knowledge would involve understanding cases in which self-knowledge failed. I defined failures of self-knowledge as cases of conflict between belief and meta-belief. In order to study failures of self-knowledge, it was necessary to have some understanding of the nature of belief. Otherwise, there could be no agreement about which cases counted as failures of self-knowledge. Mental ascription is a messy enterprise and for any set of behaviors we observe, there are a variety of mental ascriptions that we can use to attempt a psychological explanation. So I deferred discussion of failures of self-knowledge in favor of answering questions about the nature of belief.

My take on belief was that beliefs are mental states responsible for even the most mundane behaviors and of a kind possessed by many non-human animals. I understand beliefs to explain how both humans and animals are able to act on past information, react to novel situations, and solve problems. Granted, human beings can do all of these things better than most non-human animals in most situations. However, I do not view the increased abilities of human beings as resulting from a fundamentally different mental structure than non-human animals. I think that both humans and non-human animals have minds with beliefs and desires. In the previous chapter, I introduced a framework for thinking about belief which contrasted low road and high road conceptions of belief. I argued in favor of the low road conception there and continue to use it here.

I am sure that some readers will not have been convinced that I came to a proper understanding of belief in that chapter. However, armed with the low road conception of belief, I can now examine cases in which I think self-knowledge fails. Though some might regard me as failing in my stated goal of providing a theory of our knowledge of our own beliefs, I am most definitely discussing our knowledge of certain aspects of our mental lives. There can be little doubt that the kinds of cases I am concerned with represent some sort of breakdown of ordinary folk-psychological behavior explanations. Discussion of what is going on in such cases should prove valuable even if I am wrong about the role of belief in these cases. And, since understanding these cases means understanding the mental states and processes that regulate and explain them, this is a key to a kind of self-knowledge, even if the reader does not agree that it is of the kind advertised.

Unfortunately, having some clarity on the belief side of the failure of self-knowledge is not enough. In order to pinpoint a failure of self-knowledge, we must also have some idea what role meta-beliefs play. I now turn to this issue. First, I approach the problem from a conceptual standpoint. Then, I move to possible cases in which there we see genuine conflict between belief and meta-belief to see what can be learned from those cases, both in terms of delineating the functions of belief and meta-belief and for theories of self-knowledge.

2. Functions and Failure

In this section, I attempt to isolate the functional role of meta-belief in failures of self-knowledge. A few preliminary notes are in order. The first is that I assume that it is

possible for self-knowledge to fail. This may beg the question against certain views of self-knowledge, but making it helps to lay out my views. Second, I discuss a variety of hypothetical cases of belief, meta-belief, and behavioral explanation. These cases are meant to help elucidate my ultimate view. If any are found objectionable by the reader, they may be ignored.

2.1. First-Order Belief

Before examining the role of meta-belief, I should be clear about what role I think first-order belief has in human mental life. An easy summary of my views here is as follows: First-order beliefs allow human beings to do just about everything that they do. In other words, meta-belief—and by implication self-knowledge—is not necessary for most behaviors performed by human beings. One reason that I hold this is that I think that most of what human beings do is not distinctively human. Where we share behaviors with non-human animals that lack the concepts necessary to represent their own mental states, we should not be surprised that we can accomplish those same behaviors without relying on meta-belief. Theorists of the mind should always be on guard against over-intellectualizing, especially in the case of human beings. This means, among other things, not focusing on distinctively human behaviors to the exclusion of other human behaviors, not assuming that behaviors are distinctively human on the basis of *a priori* reasoning, and looking for the simplest explanations of human behaviors even when there are more complicated mentalistic explanations that seem natural.

However, even if we focus on distinctively human behaviors, I would maintain that first-order beliefs allow humans to do just about everything distinctively human that

they do. I cannot hope to come up with an exhaustive list of distinctively human behaviors, or to provide comprehensive mental explanations for those distinctively human behaviors I do examine. However, if meta-belief plays any significant role in human abilities and behaviors, I do think that it will be in influencing more or less distinctively human abilities and behaviors. There is very little reason to think that most non-human animals have self-knowledge or even much ability to form representations about the mental states of themselves or others. Of course, humans may use their self-knowledge to do things that they could do, or that other animals could do, without it. But this would be unexpected from an evolutionary perspective. It is unusual for evolution to find new ways to solve problems it has already solved.

Consider language use and mind-reading, two human abilities which are currently being studied to discover whether any non-human animals also possess them. So far as we know, humans are the only animals with any real facility for language. Recent studies into various non-human animal communication has shown that some animals have more elaborate communication systems than we initially supposed, some of which demonstrate features in common with human language such as simple syntax.[1] But despite our burgeoning appreciation for the linguistic abilities of non-human animal, humans remain unrivaled in terms of linguistic abilities.

Some have held that language use requires self-knowledge.[2] But this rests on an overly intellectual picture of human mentality. One thing that quickly becomes clear when we examine actual human language is that it is infected with mental terms and terminology. So while the ability to use language will not require mental concepts, the

[1] See various articles in Bekoff, Allen, and Burghardt 2002, pp. 255-322.
[2] See Mellor 1977.

ability to speak English or any actual human language will. This brings me to discussion of another set of distinctively human abilities. There has been an active research program for several decades examining whether various non-human animals have a theory of mind. The results are somewhat inconclusive.[3] Humans, however, do seem to make explicit use of represented information about the minds of others. We are capable of acting as if it is the case that P and Bob thinks that not-P. So let us suppose that mindreading, or explicit representation of the minds of others, is a distinctly human ability. Is it one that requires meta-belief or self-knowledge? There is a sense in which such an ability clearly requires meta-belief. Namely, in order to explicitly represent the mental states of others, one must be able to have beliefs about beliefs (or other such states). However, since I have been using 'meta-belief' to mean self-directed mental state ascription, it is not so clear that mind reading require meta-belief in this sense. Some might see a conceptual connection between knowing the minds of others and knowing one's own mind, but I am inclined to think that the two need not come together. It isn't entirely clear to me that a creature needs to have any sort of self concept to have beliefs about the minds of others. At best, it might be reasonable to maintain that it would be able to understand its own mind in the same way as it understands the minds of others. That is, it would see itself in third personal terms as a creature with such and such mental states. Perhaps it could even privilege itself in certain ways, caring more about its own mental states without ever conceiving of them as its own.

2.2. Meta-belief and Rationality

[3] See Fitzpatrick 2009, for a discussion of mind-reading abilities of non-human primates.

In an earlier chapter, I discussed Sydney Shoemaker's argument that a wholesale failure of self-knowledge is impossible. His argument is lengthy and complicated, so I will not reproduce it here. However, I pointed out then that a key role in that argument was played by a premise about the role of belief and assertion. The crux of the supposed inconsistency in the idea that a rational intelligent person could have a belief without knowing that he or she had it, stemmed from the fact that a person who believed a certain proposition would be able to assert it. That is, Shoemaker held that a person that believed P should be able to assert P. He tried to establish from this that any person who held a belief that P would also be able to assert, "I believe that P", hence functioning as if he or she had self-knowledge. I think that this move does not work. However, there is something interesting about the assumption that Shoemaker made. It deserves some closer scrutiny.

Suppose that a person that had a belief that P was always able to assert that P. Would it be possible for belief and meta-belief to conflict? It seems to me that such a conflict could occur, but that it would not be very stable. Since human beings can represent themselves as having mental states, and since a human being is likely to notice when he or she asserts a proposition, it would be difficult to maintain a conflicting meta-belief. Let's suppose that a person believes P and believes of himself that he believes not P.As soon as that person had occasion to assert P, the potential would exist for noticing something odd. Believing of himself that he believed not-P, he would be able to assert, "I believe not-P". Conjoining these assertions would result in assertion of a sentence in the form of Moore's paradox: P and I believe not-P. It has long been held that there is something wrong with asserting a Moore-paradoxical sentence, though whether the

problem is logical, conceptual, or psychological is still a topic of discussion. Whatever the status of the problem with such an assertion, it does seem safe to say that a person who found himself asserting such a sentence, or even in a position to think it true, would notice the oddity. So far, this is much in line with Shoemaker's discussion. Let me now part ways with him.

Shoemaker goes on to argue that noticing such an oddity would prevent the believer from asserting those Moore-paradoxical sentences. Furthermore, he thinks the speaker would be able to reason out that he ought to assert that he believes P, not merely that P is true. Such a person, Shoemaker claims would come to ascribe beliefs to himself in a functionally identical manner to the rest of us, effectively reducing the possibility of a failure of self-knowledge to an irrational rarity. My take on Shoemaker's reasoning is that he fails to establish a functional identity between this sort of person and the rest of us. I take it that most of us do not typically find ourselves asserting or being prepared to assert Moore-paradoxical statements. If anyone were to find themselves in such a position and to use that fact to adjust their meta-beliefs or their self-ascriptions of belief, they might talk like us, but they would function very differently.

Though I think Shoemaker fails to establish a conceptual connection between rationality and self-knowledge as he hoped to, I do think he showed that, under the following assumptions, a failure of self-knowledge would be, in a sense, unstable. First is the assumption already mentioned, that first-order belief would guarantee the ability to assert the believed content. Second, and also held by Shoemaker, is that a normal rational person would notice the oddity of asserting or being in a position to assert a Moore-paradoxical sentence. Third, we would have to assume that a person would respond to

noticing this oddity by an adjustment in either belief or meta-belief. I find the second assumption very plausible. The third, I think is debatable, but I also find plausible, given the first two assumptions.[4] It is the first assumption that I find most questionable.

Shoemaker does not say too much to defend this idea that belief would guarantee the possibility of asserting the believed content. One might maintain that such a feature was an essential property of belief. However, this is entirely incompatible with my low road conception of belief. I think that non-human animals that lack linguistic abilities can still have beliefs. It would not do to hold that an essential property of belief was that the believer be able to assert the believed content. Perhaps one might allow for animal beliefs but insist that given the ability to use language at all, it would be guaranteed that a believer could assert the content of his or her beliefs. However, I do not think that this is well motivated either. Even if it happened that human beings were capable of asserting whatever they believe, why should we think that more primitive language users would be able to? It seems perfectly conceivable that a primitive language user, perhaps one that evolved the capacity for belief long ago but language only recently, might have beliefs it could not form into language or even beliefs cut off from its linguistic abilities.

I propose to reject the assumption under discussion. If we instead hold that a believer can have a belief without being able to assert the content of that belief, we will create the possibility of a stable failure of self-knowledge. A stable failure of self-knowledge will be one in which a person has a belief such that they are not able to assert its content and also has a meta-belief with conflicting content. There are a variety of ways

[4] This is pure speculation, but given my proposals to come later in the chapter, I think that a person in such a situation would revise their meta-belief to be in line with the assertion stemming from the first-order belief because I think the correction would come from conscious reasoning that has more of a role in regulating meta-belief.

someone could get into such a state and various possibilities for how stable such a state might be.

2.3. Stable Failures of Self-knowledge

In the previous section, I showed that a failure of self-knowledge would be unstable if we were always in a position to assert the contents of our beliefs. In this section, I will continue using the idea of a stable failure of self-knowledge in the hopes of better understanding belief/meta-belief conflict. I am operating on the assumption that there is some psychologically plausible way for a person to arrive at a state of belief/meta-belief conflict that could exist for a continuing period of time. So the target is to describe a psychologically plausible, stable failure of self-knowledge. I will refer to this target as a 'stable conflict' or as a 'stable failure'.

One way to get into belief/meta-belief conflict that is compatible with just about every fallible model of self-knowledge is to be unreflective. An unreflective person might have their beliefs change without realizing it or arrive at certain meta-beliefs in some illicit way which might be corrected upon reflection. This sort of failure of self-knowledge would still be importantly different from cases of poor reflection which should not count as failures of self-knowledge. For instance, I would not class as failures of self-knowledge beliefs that a person has but have never become the subject of meta-belief. The interesting cases are ones of conflict, not failures to notice.

This brings me back to the earlier point about the stability of failures of self-knowledge. If simple reflection or introspection can undo a belief-meta-belief conflict, then it isn't particularly stable. It is essentially on a par with the cases above in which the

conflict is resolved when the believer notices the discrepancy in his assertions. So the failures of self-knowledge which are of genuine interest are ones in which a person has a belief that is inaccessible to simple reflection or introspection, that belief does not allow for the assertion of its content, and the person has a conflicting meta-belief.

A genuinely stable failure of self-knowledge like this seems much harder to get into. We should ask what belief and meta-belief would have to be like in order for a person to get into such a state. For the belief/meta-belief conflict to survive reflection or introspection, there would have to be a certain lack of conflict between the results of that reflection and the meta-belief. Similarly, there would have to be a lack of conflict between the meta-belief and the believer's assertions. So a stable failure of self-knowledge would have to involve assertions in line with meta-belief rather than belief. The first feature we can claim for meta-belief, then, is a role in the production of assertion. Let me elaborate on the nature and strength of this claim.

First, I am claiming not just that meta-beliefs have a role in some assertions. Most philosophers would probably be willing to grant them a role in the assertion of self-ascriptions such as, "I believe that…". I am claiming that, at least in failures of self-knowledge, meta-belief plays some role in the assertion of propositions that are not mental (or at least in the assertion of sentences that do not appear to have mental content given their surface grammar). So, in a case in which a person believed P and had the meta-belief that not P, the situation would be stable only if that person would assert both, "I believe not P" and "Not P". If the believer was inclined to assert, as we might expect given the conflict between belief and meta-belief, both "P" and "I believe not P", the situation would not be stable. Anyone prepared to assert both of these things would be in

a position to notice the oddity of the Moore-paradoxical conjunction and there would likely be some sort of resolution to the conflict.

One might hold—and it is possible to read parts of Shoemaker in this way—that the believer's assertions are actually always controlled by first-order belief. In that case, we should accept our conflicted believer, who believes P and believes of himself that he beliefs not-P, should be prepared to assert both, "P" and "I believe P". If that were the case, it would not be at all clear what the meta-belief was doing or why we should think that the believer even has such a meta-belief. If we are going to take seriously the possibility of even a single failure of self-knowledge, we must understand the conflicted believer's belief about his own beliefs as a genuine mental state with causes and effects. It is difficult to see how a person who talks and acts as if P in all regards should be considered to have a genuinely conflicting meta-belief at all. So a stable failure of self-knowledge would require a person whose assertions were all in accordance with the meta-belief.

I am not here claiming that all assertions are controlled by meta-belief. That is a hypothesis worthy of some consideration, but it is beyond my purposes here. I am endorsing a more restricted claim that sometimes meta-belief can play a role in assertion. Specifically, my claim is that sometimes, when a person has a meta-belief, that person will assert the content of the self-ascribed belief as if it expressed a first-order belief. It seems to me hard to avoid thinking this if we accept the possibility of a stable failure of self-knowledge.

One point in favor of this idea comes from considering old discussions about the so called transparency of belief. One interpretation of the work of Gareth Evans is that

the grounds for believing P are the same as the grounds for believing of oneself that one believes P.[5] The idea here is that in order to find out whether or not one believes P, one engages in an outward looking examination of the evidence for the truth or falsity of P. If the results of the examination are that the evidence seems to you to indicate P, then it is reasonable to believe P and reasonable to believe that you believe P. The fundamental idea is that you see through the belief, right to its content, and knowing that you have the belief that P is dependent on understanding the reasons for believing P in the first place. Some have pointed out that the close connection here between believing P and believing that one believes P could provide the basis for a model of self-knowledge (or something like it).[6] However, in accordance with the transparency metaphor, the direction of this connection is always suggested to go in one direction: belief to meta-belief. In suggesting that a meta-belief that one believes P could result in assertions that P, I am in effect proposing a connection in the other direction. It might be thought of as a false transparency. Whereas a normal belief can be looked through, like a window, to the content outside, a conflicted meta-belief is like a painting mistaken for a window.

If, as I am contending here, a stable failure of self-knowledge can result from both non-mental assertions and self-ascriptions being compatible with the meta-belief, we have some understanding of what the meta-belief does in these cases at least. And, armed with my low road conception of belief, we can have some idea what beliefs do in cases of conflict as well. I have argued that various non-verbal behaviors result from beliefs. We can suppose that these sorts of behaviors continue to result from beliefs even in failures of self-knowledge. So, if a person has a belief that P and a meta-belief that not-P, he or

[5] See Evans 1982.
[6] See Fernandez 2003 and 2005 and Williams 2004 and 2006.

she will assert not-P and that he or she doesn't believe P, but exhibit at least some non-verbal behaviors as if P were true. That is, some of the person's non-verbal behaviors will be explainable on the basis of that person having a belief that P. I make no claim yet about which non-verbal behaviors will be in line with the belief that P, nor any claim that all non-verbal behaviors are explainable in terms of the belief that P.

Just as the stability of the conflict would be threatened by the conflicted believer noticing that he or she would assert the conflicting "P" and "I believe not-P", the stability would be threatened by the conflicted believer noticing that he or she asserted "not-P" or "I believe not-P" while acting as P were true. Here, however, there are at least two reasons that I see why the situation might be more stable. The first is that the conflicted believer may be less likely to notice their non-verbal behaviors conflicting with their assertions. The second is that a noticed conflict may be rationalized in other terms.

It would be difficult to say whether anyone can fail to notice their own assertions. However, we are all accustomed to failing to notice our non-verbal behaviors from time to time. We might not realize that we tap our fingers on the desk, adjust our watches, or scratch our beards—my own personal weakness—unless it is pointed out by another. Of course, these behaviors are not plausibly explained by beliefs. My first point is simply that it is easier to be unaware of what one's body is doing than it is to be unaware of what one is asserting. However, I think the point can be carried further. For instance, if one is lost in thought about some philosophical problem, one might walk into another room, pour a drink, and walk back, without being aware of what one is doing. Unlike the nervous twitch examples, this sort of behavior is complicated enough to require a large number of beliefs. Yet, in such cases, we are not aware of these behaviors.

Further, we may be aware of our behaviors without seeking any explanation for them, or we may attempt to explain them without appeal to belief. Even when I am aware of walking to the next room and pouring a drink, I do not think about the role all of my beliefs play in that behavior. We recognize beliefs as playing a role in these behaviors when we are being reflective, but the most complex explanation likely to come from common sense is something like, "I fixed a drink because I was thirsty." If I am not paying careful attention to the role my beliefs have in producing behaviors, I am unlikely to notice a conflict. If don't attempt to explain my behaviors, then I won't explain them in relation to a belief that conflicts with anything I'd assert or anything that I take myself to believe.

If there was any noticed tension, I could look for a non-belief rationalization of the behavior. Mental explanations are always tricky and we may come up with a convincing one that doesn't conflict with our meta-beliefs. We see something like this tendency in the phenomenon of confabulation, or the offering of *post hoc* rationalizations for things that they cannot otherwise understand.[7] Initially, confabulation was understood to be tied to certain mental disorders. These disorders produced odd behaviors or odd emotional responses. Roughly, the idea was that confabulators came to believe odd things in order to explain their own behavioral and emotional responses in various situations. For instance, people with Capgras' syndrome tend to claim that their loved ones have been replaced by impostors, possibly to explain the lack of affection they feel when looking at their loved ones.[8] The disorder interrupts the emotional response normally connected to sight, but the afflicted individual makes sense of this condition by believing

[7] See Hirstein 2006, for an attempt to understand the philosophical implications of this phenomenon.
[8] See Hirstein 2006, pp. 12-13 and 114-122.

something extraordinary. Recently, confabulation has been seen less as a rare feature of unusual disorders and more as a relatively common phenomenon.[9]

The phenomenon I am suggesting here may or may not be considered a kind of confabulation. Essentially, I am suggesting that if someone noticed that he or she was behaving in a way that seemed to conflict with their ideas about their own beliefs, they might do something like confabulate an explanation that does not involve challenging their ideas about their own beliefs. For instance, suppose I believed that it was not raining and that I believed of myself that I believed it was raining. My belief that it was not raining might result in my leaving my umbrella at home. However, rather than think that I was wrong about what I believe, I might instead come to think that I like walking in the rain; why else would I leave my umbrella behind if I knew it was raining? I don't mean to suggest that there is a failure of self-knowledge every time we forget our umbrellas. In all likelihood, there is a simpler explanation in a case like this. I only discuss it to bring out the structure.

To summarize, I have been considering what a psychologically plausible, stable failure of self knowledge would look like in order to get a better understanding of the functions of belief and meta-belief in failures of self-knowledge. My considerations suggested that in order for the failures to be stable, certain behaviors, including both assertion and self-ascription, would have to comport with the meta-beliefs. First-order beliefs, on the other hand, will be more likely to manifest themselves in non-verbal behaviors. I will now turn to considering how someone might get themselves into a state which would count as a stable failure.

[9] See Wilson 2002.

3. Meta-belief and Belief Deliberation

I now understand a stable failure of self-knowledge to be a case in which a person has a conflict between a belief and a meta-belief, in which the belief primarily influences non-verbal behaviors and the meta-belief controls relevant assertions and self-ascriptions. I have argued that it is psychologically plausible that such a state would be stable. However, nothing I have said so far has indicated how a person might arrive in such a state. In this section, I will examine a few possibilities.

A first possibility to consider is that one might possess a belief, know that one has it, and have the belief change without noticing. Suppose a young college student comes to believe that anyone can be successful in America. Years of experiences might unseat this belief. This college student might see case after case of hardworking and talented people fall into trouble with drugs or debt or plain bad luck. If this college student, while young and idealistic, made patriotism central to his or her identity, that belief that anyone can make it in America might change without this former idealist realizing it. Perhaps this student would eventually become a senator and this senator would continue to say that he or she believes anyone can make it in America (and believing that he or she believes that) while sending their own children to expensive private schools and ensuring that they have every advantage in life. I think many of us are familiar with situations roughly like this in ourselves. Reflecting on some long held attitude, we come to think not merely that our old belief was false, but that we had stopped having that belief some time ago. The picture here is one in which belief changes without our awareness of that change. Exactly

how such a change might escape awareness is a worth further investigation, but I will say no more here.

If this is one way in which someone might enter into a failure of self-knowledge, then we should consider the reverse direction. Instead of looking at a case in which a belief changes without being noticed, we might look at a case in which a person takes his or her beliefs to change when they have not. In this sort of case, one starts out with a belief and knows that one has it. Perhaps I believe that it is raining and I believe that I believe it is raining. If I mistakenly think that my belief changes, that means my meta-belief will become that I believe that I believe it is not raining. In this new conflicted state, I would say that it was not raining and say that I believe it is not raining, but I may do things such as bring my umbrella or put on rain boots. But what would have made my meta-belief change without changing my belief?

In the first sort of case, one in which a belief changes without the corresponding meta-belief change, we saw the belief change in light of information that warranted a change though the believer did not recognize that information as warranting a change. Here, we might expect to see a belief stay the same while the meta-belief changes if the believer comes to think a change is warranted as a result of deliberating over some new evidence. That change may simply not come. Here, the picture is one in which the believer assumes that all of his or her behaviors will reflect the outcome of his conscious deliberations, yet the underlying belief does not change in light of those considerations. It is hard to see why a belief that it is raining might resist evidence that it has stopped raining, but it is not hard to imagine cases in which information warrants a rational change that is painful or radical resulting in a failure of first-order belief change that the

believer assumes has occurred. For instance, when a loved one dies, people often talk of not really believing that that person is gone until some bit of evidence finally prompts them to grieve. Like all the examples considered here, this is a hard case to ascribe mental states accurately, but I find the possibility of a failure of self-knowledge here intriguing. In this sort of case, a person could say, "So-and-so is dead" but many of their actions and attitudes are not in accordance with this truth. What may be interesting about this case is that the intransience is so strong that it can resist explicitly noticing the conflict in behavior. I am suggesting that there may be similar failures of self-knowledge in which the conflict goes unnoticed or is explained away.

Both kinds of cases so far are ones in which the underlying beliefs are not responding to explicit reasoning or deliberation. In the first case, the belief changes unconsciously. Some sort of belief updating happens below the level of conscious awareness and the change is never noticed or stored as a meta-belief. We all must accept that beliefs can update without conscious deliberation. I am suggesting that failures of self-knowledge can occur when such changes are unnoticed. Similarly, the second sort of case involves beliefs failing to change as a result of conscious deliberation. It is natural to assume that our beliefs will change in response to our deliberations. However, if they do not, we may not notice as long as some of our behaviors, especially our assertions and self-ascriptions, are in accordance with that warranted change.

Given the separation between deliberation and first-order belief change described above, one might worry that first-order beliefs are never sensitive to conscious deliberation. I view that as something of a live possibility. If our beliefs update entirely due to unconscious processes, our deliberations would match the results of those

processes as long as both the processes and the deliberations were rational and responding to the same evidence. A failure of self-knowledge could then occur whenever one or the other process was irrational or cut off from the evidence employed by the other. However, the structure of failures of self-knowledge that I have proposed is compatible with deliberation influencing belief in normal cases. This picture would be one of a unified system which throws up roadblocks to prevent conscious deliberation from making certain changes.

4. Failures of Self-Knowledge

In this section, I want to consider a few different mental phenomena which may fit my model for failures of self-knowledge. Some—or all—may turn out not to fit once they are fully understood. However, I think that cases such as these require thought, specifically thought along the lines of self-knowledge and the nature of belief. We stand to learn much if we can understand the minds of people in cases like these.

4.1. Phobias

Phobias have proven resistant to certain kinds of evidence. Consider a person that has a fear of spiders. Reading about the lack of dangers posed by most spiders does little to reduce the effects of the phobia. Suppose an arachnophobe reads extensively about spiders. He learns that most spiders bite humans only when provoked, that most spiders are not able to pierce human skin with their bites, and that there are very few kinds of poisonous spiders in his or her area. Given these facts, it seems rational to me to conclude that spiders are not dangerous. Suppose it seems that way to the arachnophobe also. This

arachnophobe is then able to say, “I believe spiders are harmless”, and “Spiders are not at all dangerous”. Yet, as is typical for phobias, our arachnophobe friezes up in terror at the sight of a small house spider.

How should we understand the mind of the educated arachnophobe? My suggestion is that the arachnophobe has a range of beliefs about spiders which are insensitive to certain kinds of evidence, even though he thinks his beliefs will be sensitive to that evidence. It is hard, of course, to *say* exactly what beliefs the arachnophobe are. I have in mind beliefs with contents similar to “Spiders are dangerous”, but representing them in English has the potential to create inappropriate connotations. The important thing is to see the potential structure that I have proposed for failures of self-knowledge. The first-order beliefs are in some sense irrational and unresponsive to reasoning. Yet the believer still feels as if his or her learning and deliberating is working normally.

We may be tempted to understand this case in different terms. We may be inclined to think that the arachnophobe really does believe that spiders are harmless. The phobic behaviors we would then be explained in terms of non-doxastic states. We could say that, despite his rational beliefs, he has a *fear* of spiders. Here, it seems that we would understand fear to be a non-doxastic attitude toward a kind of object, spiders in this case. This is not an unnatural idea. In fact, this is much how I introduced this case earlier in this section. A couple of points should be made. The first is that there is a sense in which this fear is *irrational*. To suggest that attitudes should be rationally related to evidence is to suggest that those attitudes are rooted in beliefs. Additionally, we should be careful in dismissing the cognition involved in emotions. The phobic behaviors are only triggered by very specific stimuli (spiders) and can be triggered by various sense modalities and

indirect evidence, such as being told there is a spider nearby. They produce a range of behaviors that can vary situationally. And though they cannot be easily unlearned, they can eventually be extinguished. All of this suggests that phobias involve mental representations. Whether these representations should be considered beliefs is, as with most cases I have discussed, not entirely clear.

4.2. Implicit Biases

There is a wide literature on implicit bias in human reasoning. I do not mean to suggest here that all implicit biases should be understood in the same way or that each is a failure of self-knowledge. However, some show signs of exhibiting the structure of such a failure. For instance, consider cases of implicit racism.[10] As mentioned last chapter, this is a phenomenon in which people that sincerely describe themselves as without prejudice seem to demonstrate prejudicial behaviors. They also seem quite sincere in claiming not to notice their behaviors as prejudicial.

Here, again, we have behaviors—complex behaviors involving mental representations—which conflict with the believer's assertions and self-ascriptions. We can see the implicit racist as having deliberated about racial equality, decided that the evidence was in favor of it, yet failing to change their first-order beliefs. Perhaps these racist beliefs result from some old evolved mechanism which prompted us to favor our close kin over competing groups. Perhaps they are resistant simply because they were learned early and beliefs formed early are more resistant to change. The reason that they are maintained is unimportant. The important point is that they are irrationally maintained

[10] See Dovidio and Gaertner 2004 and Wilson 2002.

though the believer consciously weighs the evidence against them. However, the implicit racist is not aware of any kind of irrationality, because he or she takes herself to have updated those beliefs.

Just as with phobias, it is possible to view this case as one in which the implicit racist really does have the beliefs he or she thinks she has. One might try to hold that the implicitly racist behaviors result from something like an emotional response to members of the other race. Here as well, it is worth noting the complexity of the representations that must be involved and the amount of cognition required. Just as with phobias, I worry that non-doxastic descriptions will gloss over important features of the mental causes of these behaviors.

4.3. Religious Belief

I opened the last chapter with a passage from J.S. Mill citing the apparent inconsistencies in the professed attitudes and manifest actions of many Christians. I suggested at that time that I thought such a case might be a failure of self-knowledge. Now that I have elaborated the structure of failures of self-knowledge as I understand it, it is worth revisiting the case. There are at least two kinds of failure of self-knowledge that might be associated with religious belief.

The first, which I took to be the phenomenon Mill was noticing, is when religious people fail to live up to the commitments they should have in light of the beliefs they profess. For instance, Christians are supposed to believe that there is nothing noble about being rich. Many Christians will claim to posses some belief like this, yet some that profess this belief work hard to cheat their neighbors out of hard-earned money. Or, if

you have a higher opinion of the average Christian, it still seems that many work to obtain money for its own sake rather than facing up to potential obligations to the poor.

Georges Rey has also discussed religious belief as a failure of self-knowledge.[11] The cases he is interested in are cases in which the professed religious beliefs seem irrational given the evidence available to the believer. Suppose, for example, that you thought there was plenty of evidence against the existence of an all powerful God and that this evidence was available to the religious and non-religious alike. The idea would be that no one really *believes* in an all powerful God. Your rationality just won't allow it. However, there are social, political, or, moral reasons to believe what Christians are supposed to believe so that you can be a member of the church. So you are able to convince yourself that you believe in an all powerful god, though in reality, you believe no such thing.

These cases each have the right sort of structure to count as failures of self-knowledge on my view. I do not propose that all religious people lack self-knowledge in one of these ways. I find it an intriguing way to understand apparent hypocrisy on the part of some religious individuals. There are, of course, other potential explanations. Rather than appeal to an emotional explanation here, a better competitor theory would probably involve failing to understand the implications of one's beliefs. I do not think that such an explanation would adequately explain the whole range of behaviors, but this is not the place to settle the issue.

4.4. Intellectual Belief

[11] See Rey 2007.

Finally, I want to consider the possibility that we can understand the bizarre intellectual commitments of some philosophers as failures of self-knowledge. Various philosophers have held that knowledge is impossible, that motion is impossible, and that everything that exists is mental. These views are really the tip of an iceberg of strange theses held by philosophers over the centuries. But some of these ideas seem to rationalize bizarre actions. If you thought knowledge was impossible, why would you attempt to learn things? On what basis would you expect the rooms in your house to be a certain way? If you thought motion was impossible, why would you attempt to move? There is a famous anecdote of Dr. Samuel Johnson attempting to refute Bishop Berkeley's idealism by kicking a large stone. How could anyone successfully believe anything other than that action involved one physical body striking another?

My suggestion is that philosophers may, through complex and abstract deliberation, come up with convincing arguments for all sorts of ideas. However, there may be some ideas that, good argument or not, we are not designed to believe. Even if one had good reason to think that motion was impossible, it would be impossible not to try to move from place to place and, moreover, to think of yourself and describe yourself as so moving. Some philosophical commitments, we might say, exist only in philosophical discussion.

That's not to say that these beliefs are irrational. It would be all too easy to caricature some philosophers as attempting to talk themselves out of good, old common-sense. However, there is renewed interest in the idea that, for instance, there is really only one thing and that distinct existences are illusory.[12] The arguments for this position are

[12] See Schaffer 2008.

intriguing. However, even if my assessment of the arguments was that they were clearly sound, I do not think I could truly accept the conclusion in the sense of coming to believe it. My beliefs about other things being separate from me are so fundamental that I would continue to talk, think, and act as if there were separate and distinct things, unless I was being exceedingly and deliberately careful. And even then, I doubt I could stay consistent for very long.

5. Self-knowledge as Unity of Reasoning

The reader may or may not at this point accept my conception of failures of self-knowledge. There have been a lot of points of potential disagreement. My conception depends upon a low road conception of belief, which is certain to be contentious. It also depends on some reasoning about the role of belief and meta-belief in human behavior. While I do not take firm stands on what roles these states have in all behaviors, I do hold that in stable failures of self-knowledge first-order assertion can be controlled by meta-belief. This is a significant thesis. Since part of my argument for it relies on assuming that stable failures of self-knowledge are possible, one could easily attempt to argue in the other direction. Indeed, Sydney Shoemaker's famous argument for the conceptual connection between self-knowledge and rationality does essentially that. Finally, even if my conception of failures of self-knowledge and my arguments for it are unproblematic, one might not think any of the cases I have described really fit that conception. I hope that the reader is at least still open to my conception, if not outright persuaded. But, regardless of how successful I have been to this point, I ask the reader's indulgence as I

attempt to move from my conception of a failure of self-knowledge to a sketch of a positive theory of self-knowledge.

Before I discuss a theory of the epistemology of self-knowledge, I want to first address this issue of the function and value of self-knowledge. In earlier parts of the dissertation, I pointed out that it is possible to think that most of human behavior results from our first-order beliefs. We have a tendency to over intellectualize human behavior, and this might extend to our understanding of ourselves. We should be careful to ask what we can't do simply by having beliefs and desires. For what purpose, would we additionally need to know what our beliefs and desires are?

The picture created by my discussion above is one in which our beliefs at least sometimes fail to respond to our deliberations about what to believe. This can create a stable conflict between belief and meta-belief. In turn, this can create a situation in which we sometimes do not act in accordance with what we take ourselves to believe. Worse, such a person would make assertions and self-ascriptions in line with that reasoning, causing some of their non-linguistic behavior to conflict with their stated commitments. Thus, a person that fails to possess self-knowledge can become a sort of oblivious hypocrite, talking a good game, but failing to make good on the talk.

Avoiding this situation is, I propose, the primary value of self-knowledge. The self-knower is assured of a unity of talk and action as well as a unity of thought and action. Whether or not one's beliefs are actually sensitive to one's deliberations, a person without conflict between belief and meta-belief functions as if they are. In short, even if the picture we have of ourselves as rational actors is flawed, we are not hindered by accepting it so long as we know what we really believe.

Now, I turn to the epistemology of self-knowledge. I see two fundamental questions that need to be addressed. One is the question of modeling how meta-beliefs are formed. The other is a question about the scope of failure and the possibility for improvement.

In order to draw conclusions about the justification of our meta-beliefs, we must have some idea what the connection between belief and meta-belief is like. Unfortunately, my discussion of the failures of self-knowledge can be interpreted in one of two ways. The first is as indicating interference with the normal operation of a connection between belief and meta-belief. This is compatible with virtually any model of self-knowledge. On this understanding, exploring failures of self-knowledge can tell us lots of interesting things about the mind, but cannot provide an account of how we form meta-beliefs. However, there is a more interesting approach to take. We might interpret failures as breakdowns of the mechanisms in successful instances rather than as interference from outside processes and mechanisms. Ultimately, there are a lot of empirical questions that have to be asked and answered before too much progress can be made here. However, it seems worthwhile to say something about what kind of model this second approach could yield.

Since failures of self-knowledge can be seen as cases in which first-order beliefs change without deliberation or cases in which they remain the same despite deliberation, we have some reason to doubt the effectiveness of deliberation on our first-order attitudes. Rather than positing mechanisms which interfere with the connection between deliberation and first-order belief, we could consider the possibility that there is no direct connection, even when there is no conflict between belief and meta-belief. On this view,

we are all radically wrong in supposing that our conscious deliberations matter to our first-order beliefs.

Yet, this radical mistake need not result in total disaster. A few plausible assumptions will ensure a fairly regular correspondence between belief and meta-belief. First, we must assume that meta-belief is typically sensitive to conscious deliberation. So, when you consider whether P and it seems true, you may not succeed in believing P, but you will succeed in believing that you believe P. I do not mean to equate meta-belief with consciously held attitudes. I merely suggest a closer connection between them than between deliberation and belief. Second, we should assume that beliefs are formed and maintained rationally or in accordance with relevant information. Third and finally, we should assume that whatever processes form and maintain beliefs have access to the same information that we have access to in our conscious deliberations. If we make those assumptions, then our beliefs will form and maintain themselves rationally in parallel with our conscious deliberation and there will be little conflict. In fact, we can more thoroughly say when conflicts will occur. Conflicts would occur when there was a breakdown in the rationality of the first-order processes, conscious deliberation, or both. Additional conflicts may occur if there is a lack of shared information.

This is an interesting picture of self-knowledge. I think that it comports well with contemporary work in psychology. Specifically, it fits into models of the mind as being composed of two different types of systems.[13] However, further development of it will require more time and more consideration of current and future empirical work. One objection that may come to mind is that this makes our conscious deliberations nearly

[13] See Frankish 2004 and Frankish and Evans 2009.

epiphenomenal. But this is a complicated issue. As it stands, I propose that conscious deliberation does, or at least can, influence meta-belief. Since I hold meta-belief is at least sometimes responsible for some linguistic behaviors, I have not taken away all causal relevance from conscious deliberation. And there are still important questions to be addressed about the mental causes of voluntary versus involuntary behavior and consciously controlled versus unconscious behavior. There is some reason to think that we can make ourselves act more in accordance with our conscious deliberations as long as we apply care and effort. Detailing this process is part of the empirical and conceptual project still ahead.

Finally, I turn to the ameliorative question. Given my exploration of the failures of self-knowledge, how can we do better? In part, answering this question also depends on future empirical work. Having some understanding of the structure of failures is the first step to identifying the conditions under which actual failures may occur. But, crucially, this discussion also points to a need to reevaluate the role of deliberation in our mental lives. It does not suggest that we cease deliberating, but it does suggest that we not treat our deliberations as settling our beliefs. We must be open to the possibility that we have hidden attitudes which might influence our behaviors against our rational judgments. Further, we must consider the possibility that actually coming to possess the rational belief in a situation may involve more than arm chair reasoning. We may have to study the results of empirical psychology to find out if we can get our behaviors to match the results of our best reasoning, how to do it when it is possible, and how to mitigate the drawbacks if there are times when it is not possible or not practical. Whether or not you have been convinced by my model of these phenomena as belief and meta-belief, I think

you will agree that the task assigned by the Delphic oracle can now be seen to be just that much more difficult. I leave it to the reader to decide whether it is still worthwhile.

Bibliography

Allen, Colin, and Marc Beckoff. (1997): *Species of Mind*: *The Philosophy of Biology and Cognitive Ethology*. MIT Press.

Armstrong, David. (1968): *A Materialist Theory Of Mind*. Routledge & Kegan Paul.

Audi, Robert. (1994): "Dispositional Beliefs and Dispositions to Believe." Noûs 28.4: 419-434.

Audi, Robert. (1993): *The Structure of Justification.* Cambridge University Press.

Bar-On, Dorit, and Douglas C Long. (2001): "Avowals and First-Person Privilege." Philosophy and Phenomenological Research 62.2: 311-335.

Bar-On, Dorit. (2004): *Speaking My Mind: Expression and Self-Knowledge*. Oxford University Press.

Bekoff, Marc, Colin Allen, and Gordon M. Burghardt. (2002): *The Cognitive Animal*. MIT Press.

Bilgrami, Akeel. (1998): "Self-Knowledge and Resentment" in *Knowing Our Own Minds.* Ed. Crispin Wright, Barry C. Smith, and Cynthia Macdonald. Oxford University Press. pp. 207-242.

Bilgrami, Akeel. (2006): *Self-knowledge and Resentment*. Harvard University Press.

Boghossian, Paul A. (1989): "Content and Self-Knowledge." Philosophical Topics 17.1: 5-26. Reprinted in Boghossian, Paul. *Content and Justification.* Oxford University Press. 2008. pp. 139-158.

Boghossian, Paul. (2008): *Content and Justification.* Oxford University Press.

Borel, Brooke. (2009): "Mind-Reading Tech May Not Be Far Off". *Popular Science.* URL = <http://www.popsci.com/scitech/article/2009-06/mind-reading-tech-way>

Bonjour, Laurence. (1980): "Externalist Theories of Empirical Knowledge." Midwest Studies in Philosophy 5: 53-74.

Brandom, Robert. (1995): "Knowledge and the Social Articulation of the Space of Reasons." *Philosophy and Phenomenological Research*, Vol. 55, No. 4. pp. 895-908

Burge, Tyler. (1979): "Individualism and the Mental". *Midwest Studies in Philosophy* ; pp. 4:73-121.

Burge, Tyler. (1988): "Individualism and Self-Knowledge". *Journal of Philosophy*. Vol. 85, pp. 649-663.

Burge, Tyler. (1996) "Our Entitlement to Self-Knowledge." Proceedings of the Aristotelian Society 96: 91-116.

Carruthers, Peter. (1996): *Language, Thought, and Consciousness*. Cambridge University Press.

Carruthers, Peter. (2009): "An Architecture for Dual Reasoning". in *In Two Minds: Dual Processes and Beyond.* Oxford University Press. pp. 109-129.

Carruthers, Peter and Peter K Smith. Eds. (1996): *Theories of Theories of Mind*. New York: Cambridge University Press.

Cassam, Quassim. Ed. (1994): *Self-Knowledge*. Oxford University Press.

Chaiken, Shelly and Yaacov Trope, Eds. (1999): *Dual Process Theories in Social Psychology.* The Guilford Press.

Cohen, L. J. (1992): *An Essay on Belief and Acceptance.* Clarendon Press.

Davidson, Donald. (1982): "Rational Animals". *Dialectica: International Journal of Philosophy of Knowledge.* 36: 317-328. Reprinted in *Subjective, Intersubjective, Objective*. Oxford University Press. Pp. 95-105.

Davidson, Donald. (1984): "First Person Authority." Dialectica: International Journal of Philosophy of Knowledge 38.(1984): 101-112. Reprinted in *Subjective, Intersubjective, Objective*. Oxford University Press. 2001. pp. 3-14.

Davidson, Donald. (1987): "Knowing One's Own Mind." *The Proceedings and Addresses of the American Philosophical Association,* 60, pp. 441-58. *Subjective, Intersubjective, Objective*. Oxford University Press. 2001. pp. 15-38.

Davies, Martin and Tony Stone, Eds. (1995a): *Folk Psychology*. Cambridge: Blackwell.

Davies, Martin and Tony Stone, Eds. (1995b). *Mental Simulation*. Cambridge: Blackwell.

de Sousa, R. (1971): "How to Give a Piece of Your Mind: or, the Logic of Belief and Assent". *Review of Metaphysics*, 1*25*52-79.

Dennett, Daniel. (1989): "True Believers". In *The Intentional Stance.* MIT Press 1989. pp. 13-35.

Dennett, Daniel. (1978): "How to change your mind" in *Brainstorms.* MIT Press. pp. 300-310.

Descartes, Rene. (1993): *Meditations on First Philosophy.* Trans. Donald A. Cress. Hackett Publishing.

Dovidio, J. F., & Gaertner, S. L. (2004). “Aversive racism”. In M. P. Zanna (Ed.), Advances in experimental social psychology (pp. 1 – 52). San Diego, CA: Academic Press.

Elga, Adam. (2005): "On Overrating Oneself---and Knowing It." Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 123.1-2. pp. 115-124.

Evans, Gareth. (1982): “Self-Identification”. In The Varieties of Reference. Ed. John McDowell. Oxford University Press.

Evans, Jonathan St. B. T. (1977): “Toward a statistical theory of reasoning.” *Quarterly Journal of Experimental Psychology*, 29, 297-306.

Evans, Jonathan St. B. T. and Keith Frankish, Eds. (2009): *In Two Minds.* Oxford University Press.

Ferdnandez, Jordi. (2003): “Privileged Access Naturalized”. *The Philosophical Quarterly.* Vol. 53, No. 212.

Ferdnandez, Jordi. (2005): “Privileged Access Revisited”. *The Philosophical Quarterly.* Vol. 55, No. 218.

Fitzpatrick, Simon. (2009): “The Primate Mindreading Controversy: a case study in simplicity and methodology in animal psychology”. In *The Philosophy of Animal Minds.* Oxford University Press.

Frankish, Keith and Evans, Jonathan St. B. T. (2009): “The Duality of mind: An historical perspective” in *In Two Minds: Dual Processes and Beyond.* Oxford University Press. pp. 1-32.

Frankish, Keith. (2004): *Mind and Supermind.* Cambridge University Press.

Frankish, Keith. (2009): “Systems and Levels: Dual-system theories and the personal-subpersonal distinction”. in *In Two Minds: Dual Processes and Beyond.* Oxford University Press. pp. 89-109.

Fricker, Elizabeth. (1998): “Self-Knowledge: Special Access versus artifact of grammar—a dichotomy rejected” in *Knowing Our Own Minds*. Ed. Crispin Wright, Barry C. Smith, and Cynthia Macdonald. Oxford University Press. pp. 155-206.

Gendler, Tamar Szabó. (2008): "Alief and Belief." *Journal of Philosophy* 105.10. pp. 634-663.

Gertler, Brie. Ed. (2003): *Privileged Access.* Ashgate.

Gertler, Brie. (2011): *Self-Knowledge.* Routledge.

Gettier, Edmund. (1963): “Is Knowledge Justified True Belief?”. *Analysis* 23: pp. 121-123.

Gibbons, J. (1996): “Externalism and Knowledge of Content”, *Philosophical Review* 105: 287-310.

Goldman, Alvin. (1979): “What is Justified Belief?” in *Justification and Knowledge*. Ed. George Pappas. Dordrecht. pp. 1-24.

Goldman, Alvin. (1993): "The Psychology of Folk Psychology" *Behavioral and Brain Sciences*. 16: 15-28.

Goldman, Alvin. (1995): "Epistemology, Functionalism, and Privileged Access," *Behavioral and Brain Sciences*. 18: 395-398.

Goldman, Alvin. (2006): *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading.* New York: Oxford University Press.

Goldman, Alvin. (2008): “Immediate Justification and Process Reliabilism,” in Quentin Smith, ed., *Epistemology: New Essays*, pp. 63-82. New York: Oxford University Press.

Gopnik, Alison. (1993). “How we know our minds: The illusion of first-person knowledge of intentionality”. *Behavioral and Brain Sciences*, 16, 1-15, 90-101

Gopnik, Alison. (1999): *The Scientist in the Crib: What Early Learning Tells Us About the Mind.* Harper Collins.

Gopnik, Alison and Meltzoff, Andrew N. (1997): *Words, Thoughts, and Theories.* MIT Press.

Hatzimoysis, Anthony. Ed. (2011): *Self-knowledge.* Oxford University Press.

Heal, Jane. (2004): “Moran’s *Authority and Estrangement*”. *Philosophy and Phenomenological Research*. Vol. LXIX, No. 2: 427-432.

Hess, E. H. (1975): "The role of pupil size in communication" In *Scientific American* .233: 110-119.

Hirstein, William. (2006): *Brain Fiction: Self-Deception and the Riddle of Confabulation.* MIT Press.

Klein, Peter. (1999): “Human Knowledge and the Infinite Regress of Reasons”. In *Philosophical Perspectives* 13, Epistemology: 297-325.

Kornblith, Hilary. Ed. (1994): *Naturalizing Epistemology.* MIT Press.

Kornblith, Hilary. (2002): *Knowledge and its place in nature.* Clarendon Press.

LeDoux, Joseph. (1996): *The Emotional Brain: the mysterious underpinnings of emotional life.* Simon and Schuster.

Locke, John. (1975): *An Essay Concerning Human Understanding.* Edited by P.H. Nidditch. Clarendon Press.

Ludlow, Peter and Norah Martin, Eds. (1998). Externalism and Self-Knowledge. Stanford: CSLI Pub.

Lycan, William G.(1995): "Consciousness as Internal Monitoring, I (Vol. 9: AI, Connectionism and Philosophical Psychology)." Nous-Supplement: Philosophical Perspectives 9: 1-14.

Martin, M. G. F. (1998): "An Eye Directed Outward". In *Knowing Our Own Minds.* Ed. Crispin Wright, Barry C. Smith, and Cynthia Macdonald. Oxford University Press. pp. 99-121.

McDowell, John. (1998): "Response to Crispin Wright". In *Knowing Our Own Minds.* Ed. Crispin Wright, Barry C. Smith, and Cynthia Macdonald. Oxford University Press. pp. 47-62.

Mellor, D. H. (1977): "Conscious Belief." Proceedings of the Aristotelian Society 78: 88-101.

Mill, J.S. (1989): *On Liberty and Other Writings*. Edited by Stefan Collini. Cambridge University Press.

Moran, Richard. (1997): "Self-Knowledge: Discovery, Resolution, and Undoing." European Journal of Philosophy 5.2: 141-161.

Moran, Richard. (2001): *Authority and Estrangement*. Princeton University Press.

Nuccetelli, Susana. Ed. (2003): *New Essays on Semantic Externalism*. MIT Press.

Plato. (2002): *Five Dialouges: Euthyphro, Apology, Crito, Meno, Pheado.* Trans. G.M.A. Grube, 2nd edition, revised by John Cooper. Hackett Publishing.

Putnam, Hilary. (1981): *Reason Truth, and History.* Cambridge University Press.

Reber, A. S. (1993): *Implicit Learning and Tacit Knowledge.* Oxford University Press.

Reyna, V.F. (2004). How people make decisions that involve risk. A dual-processes approach. Current Directions in Psychological Science, 13, 60-66.

Rey, Georges. (2007): "Meta-Atheism: Religious Avowal as Self-Deception". In *Philosophers Without Gods*. Ed. Louise Antony. Oxford University Press. pp. 243-265.

Ryle, Gilbert. (1949): *The Concept of Mind.* University of Chicago Press.

Saidel, Eric. (2009): "Attributing Mental Representations to Animals". In *The Philosophy of Animal Minds.* Oxford University Press.

Schaffer, Jonathan. (2008): "Monism", *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/monism/>.

Schwitzgebel, Eric. (2010): "Belief", *The Stanford Encyclopedia of Philosophy* (Winter 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2010/entries/belief/>.

Shoemaker, Sydney. (1988): "On Knowing One's Own Mind." Nous-Supplement: Philosophical Perspectives 2. Atascadero: Ridgeview, 1988. 183-209. Reprinted in Shoemaker, Sydney. (1996): *The First Person Perspective and Other Essays*. Cambridge University Press. pp. 25-49.

Shoemaker, Sydney. (1990): "First-Person Access." Nous-Supplement: Philosophical Perspectives 4: 187-214. Reprinted in Shoemaker, Sydney. (1996): *The First Person Perspective and Other Essays*. Cambridge University Press. pp. 50-73.

Shoemaker, Sydney. (1994a): "Lecture I: The Object Perception Model -- Self Knowledge and 'Inner Sense'." Philosophy and Phenomenological Research 54.2: 249-269. Reprinted in Shoemaker, Sydney. (1996): *The First Person Perspective and Other Essays*. Cambridge University Press. pp. 201-223.

Shoemaker, Sydney. (1994b): "Lecture II: The Broad Perceptual Model -- Self Knowledge and 'Inner Sense'." Philosophy and Phenomenological Research 54.2: 271-290. Reprinted in Shoemaker, Sydney. (1996): *The First Person Perspective and Other Essays*. Cambridge University Press. pp. 224-245.

Shoemaker, Sydney. (1994c): "Lecture III: The Phenomenal Character of Experience -- Self Knowledge and 'Inner Sense'." Philosophy and Phenomenological Research 54.2: 291-314. Reprinted in Shoemaker, Sydney. (1996): *The First Person Perspective and Other Essays*. Cambridge University Press. pp. 246-268.

Shoemaker, Sydney. (1995): "Moore's Paradox and Self-Knowledge." Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 77.2-3: 211-228. Reprinted in Shoemaker, Sydney. (1996): *The First Person Perspective and Other Essays*. Cambridge University Press. pp. 74-93.

Shoemaker, Sydney. (1996): *The First Person Perspective and Other Essays*. Cambridge University Press.

Smith, Eliot R. and Collins, Elizabeth C. (2009): "Dual Process Models: A social psychological perspective". in *In Two Minds: Dual Processes and Beyond.* Oxford University Press. pp. 197-217.

Stannovich, Keith E. (1999): *Who is Rational? Studies of Individual differences in reasoning".* Lawrence Elrbaum Associates.

Stich, Stephen P. and Shaun Nichols. (2003): Mindreading: an integrated account of pretence, self-awareness, and understanding of other minds. Oxford University Press.

Stich, Stephen. (1983): *From Folk Psychology to Cognitive Science: The Case Against Belief.* MIT Press.

Strawson, P. F. (1974): "Freedom and Resentment" in *Freedom and Resentment and other Essays*. Methuen Publishing Ltd.

Wegner, Daniel. (2002): *The Illusion of Conscious Will*. MIT Press.

Williams, Bernard. (1973): "Deciding to Believe" in *Problems of Self.* Cambridge University Press.

Williams, John N. (2004): "Moore's Paradoxes, Evans's Principle and Self-Knowledge". *Analysis.* 64.4: pp. 348-53.

Williams, John N. (2006): "In Defense of an Argument for Evans's Principle: A Rejoinder to Vahid". *Analysis.* 66.2: pp. 167-70.

Williamson, Timothy. (2000): *Knowledge and Its Limits.* Oxford University Press.

Wilson, Timothy. (2002): *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Belknap Press.

Wilson, T. D., & Dunn, E. (2004): "Self-knowledge: Its limits, value, and potential for improvement." *Annual Review of Psychology, 55,* 493-518.

Wright, Crispin. (1998): "Self-Knowledge: The Wittgensteinian Legacy". In *Knowing Our Own Minds.* Ed. Crispin Wright, Barry C. Smith, and Cynthia Macdonald. Oxford University Press. pp. 13-46.

Zimmerman, Dean. (2002): "'Persons and Bodies': Constitution without Mereology?." Philosophy and Phenomenological Research 64.3: 599-606.