

# **First-Person Authority and its Limits**

Adam John Andreotta

Bachelor of Science (Computer Science), Bachelor of Arts (Philosophy, with Honours)



This thesis is presented for the degree of Doctor of Philosophy of

The University of Western Australia

School of Humanities

Discipline of Philosophy

November 2017



## THESIS DECLARATION

I, Adam John Andreotta, certify that:

This thesis has been substantially accomplished during enrolment in the degree.

This thesis does not contain material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution.

No part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of The University of Western Australia and where applicable, any partner institution responsible for the joint-award of this degree.

This thesis does not contain any material previously published or written by another person, except where due reference has been made in the text.

The work(s) are not in any way a violation or infringement of any copyright, trademark, patent, or other rights whatsoever of any person.

The work described in this thesis was funded by the Theresa Symons Philosophy Trust and was supported by an Australian Government Research Training Program (RTP) Scholarship.

This thesis contains only sole-authored work, some of which is being considered for publication under sole authorship

Signature:



Date:

19.7.18



**AUTHORSHIP DECLARATION: SOLE AUTHOR PUBLICATIONS**

---

This thesis contains the following sole-authored work that is currently under review for the journal *Philosophical Studies*.

Details of the work: A 10,000-word manuscript titled 'First-person authority and its limits' was submitted to the journal *Philosophical Studies*, on the 18<sup>th</sup> September 2017. It is still under review.

Location in thesis: Material from chapters 1 and 2 were incorporated into the manuscript.

Signature: 

Date: 19.7.18



## **Abstract**

This dissertation attempts to explain the nature and limits of first-person authority—the thesis that our first-person ascriptions about what mental states we are in are more likely to be true, compared to the ascriptions that others make about our mental states. My central claim is that the limits of first-person authority are the limits of introspection. After offering a general theory-neutral account of what it takes for a process to be introspective I address the question: ‘What mental states can be introspected?’ In so doing I first argue against sceptical accounts of self-knowledge which claim that we cannot introspect our propositional attitudes. I then defend a positive account of introspection for propositional attitudes, a view of self-knowledge called the Transparency Method.



## **Acknowledgements**

First and foremost, I would like to thank my advisors: Miri Albahari and Nin Kirkham. I have greatly benefitted from the constructive and supportive way in which they provided feedback to my work. They were always encouraging, and generous with their time. I learned a lot about how to do philosophy from the example they set. I would also like to thank Miri for allowing me to participate in her ‘Consciousness’ honours seminar, and Nin for allowing me to participate in her ‘Topics in Social and Political Philosophy’ seminar. I greatly benefitted from the discussions that arose in these seminars.

The members of the ‘UWA Humanities Writing Group’, a group I joined in 2014, have contributed profoundly to my personal and professional development at UWA over the past four years. The sessions that our group organised provided me with a place to share some of my early ideas. I would like to thank Deborah Seiler, Guy Kirkwood, Federica Verdina, Louis Marshall, Chellyce Birch, Francesco De Toni, Kelly-Ann Couzens, Jane Whiteley, and Nadia Meneghello for their participation in these sessions.

I would like to extend my gratitude to UWA for providing me with the opportunity to take on the role of submissions editor for the UWA based journal *Limina*. My time as submissions editor gave me an invaluable insight into the world of academic publishing and conference organisation. My own writing style and presentation ability improved enormously from my time spent in the role. I would like to thank Vanessa Karas, Mark Mazzoni, Amy Budrikis, Antonia Naarstig, Alicia Ettlin, Parisa Shams, and Jessica Murray for making my time as submissions

editor a memorable one. I am very proud of the editions we published, and could not have asked for a better team to work with.

I have learned a great deal from the discussions I had with the following people, some of whom read and commented on sections of the dissertation: Michael Schrader, Heath Williams, Cliff Stagoll, Karine Broux, Michael Rubin, Sean Ramsey, Matthew O'Neill, Sam Baron, Clas Webber, Michael Mitchell, Lauren Ashwell, Ryan Cox, Michael Levine, Barry Maund, Michael Barberzat, and Harriet Levenston.

I would also like to acknowledge those who attended the various talks I gave, and asked insightful questions, during the course of my candidature. I would like to thank audiences at the 2016 and 2017 annual *Limina* conference, as well as audiences at the 2016 and 2017 AAP conference. The feedback I received at these talks enabled me to articulate my ideas in a clearer way, and helped to identify the shortcomings in my arguments.

I am indebted to the Theresa Symons Philosophy Trust, who I was fortunate enough to receive a postgraduate research scholarship from. This research was also supported by an Australian Government Research Training Program (RTP) Scholarship. These scholarships allowed me to dedicate the necessary time to complete this manuscript. Additional travel funding from UWA was also granted to me so that I could travel to three AAP conferences—in Canberra, Melbourne and Adelaide, respectively.

Lastly, I would like to thank my family. Without their unabated support and encouragement, the completion of this manuscript would not have been possible.

# Contents

|  |            |
|--|------------|
| <b>Abstract</b> .....  | <b>vii</b> |
| <b>Acknowledgements</b> .....  | <b>ix</b>  |
| <b>Introduction</b> .....  | <b>1</b>   |
| <b>Chapter 1 What is Introspection?</b> .....                            | <b>13</b>  |
| 1.1 Introspection as Inner Perception? .....                             | 15         |
| 1.2 Alternatives to Inner Perception .....                               | 18         |
| 1.3 Terminological Concerns .....  | 22         |
| 1.4 What Introspection is Not.....                                       | 24         |
| 1.5 The Theory-Neutral Account of Introspection .....                    | 28         |
| 1.5.1 The First-Person Condition .....                                   | 28         |
| 1.5.2 The Difference Condition.....                                      | 30         |
| 1.5.3 The Occurrent Condition.....                                       | 33         |
| 1.5.4 Alternative Conditions .....                                       | 34         |
| 1.5.5 Applications .....   | 38         |
| 1.6 The Targets of Introspection .....                                   | 41         |
| 1.7 Conclusion .....   | 48         |
| <b>Chapter 2 First-Person Authority</b> .....                            | <b>51</b>  |
| 2.1 First-person Authority—Terminological Issues .....                   | 53         |
| 2.2 Varieties of First-Person Authority .....                            | 58         |
| 2.3 Infallibility, Self-Intimation, and Incorrigeability.....            | 60         |
| 2.4 First-person Authority: Introspective Justification.....             | 65         |
| 2.5 Objection: Introspection as Explanatorily Inadequate? .....          | 75         |
| 2.6 Conclusion.....  | 78         |
| <b>Chapter 3 Challenging Introspection</b> .....                         | <b>79</b>  |
| 3.1 The Content Externalism Challenge.....                               | 81         |
| 3.2 The Inferentialism Challenge .....                                   | 91         |
| 3.3 Carruthers’ Interpretive Sensory-Access (ISA) Theory .....           | 95         |
| 3.4 Two Objections to the Interpretive Sensory-Access (ISA) Theory ..... | 99         |
| 3.4.1 The Accuracy Objection.....  | 100        |
| 3.4.2 The Sensory Data Objection.....                                    | 102        |
| 3.5 Confabulation and Interpretation.....                                | 105        |
| 3.6 The Ontological Parsimony Argument.....                              | 108        |
| 3.7 Conclusion .....   | 115        |
| <b>Chapter 4 Evidence and Error</b> .....                                | <b>117</b> |
| 4.1 Introspection, Confabulation, and the Varieties of Error .....       | 119        |
| 4.1.1 Self-Deception.....  | 125        |
| 4.1.2 Brute Errors.....  | 129        |
| 4.1.3 Basing Errors.....   | 132        |
| 4.2 The Patterning Argument .....  | 134        |
| 4.2.1 Choice and Confabulation.....                                      | 138        |
| 4.2.2 Judgement and Confabulation .....                                  | 146        |

|  |            |
|--|------------|
| 4.2.3 Intention and Confabulation .....  | 155        |
| 4.3 Patterns and Classification .....  | 161        |
| 4.4 Conclusion .....   | 162        |
| <b>Chapter 5 Doxastic Transparency, Rationality, And Introspection.....</b>                                    | <b>165</b> |
| 5.1 Doxastic Transparency .....  | 166        |
| 5.2 Arguments in Favour of the Transparency Method .....   | 174        |
| 5.3 Rationalist Versus Empiricist Approaches to the Transparency Method.....                                   | 186        |
| 5.4 Gertler’s Challenge .....  | 191        |
| 5.5 Transparency and Rationality .....   | 193        |
| 5.6 Conclusion .....   | 200        |
| <b>Chapter 6 Objections to the Rationalist Interpretation of the Transparency Method.....</b>                  | <b>201</b> |
| 6.1 The <i>Homo Philosophicus</i> Objection .....  | 202        |
| 6.2 The Corruption Objection.....  | 207        |
| 6.3 The Epistemic Stance Objection.....  | 212        |
| 6.4 When Judgement and Belief Come Apart: The Matching Problem .....   | 216        |
| 6.4.1 Two Cases of Mismatch .....  | 218        |
| 6.4.2 Possible Responses to the Matching Problem .....   | 219        |
| 6.4.3 Non-Mismatch Explanations .....  | 222        |
| 6.4.4 Mismatch Explanations .....  | 227        |
| 6.5 Conclusion.....  | 236        |
| <b>Chapter 7 Extending the Transparency Method Beyond Belief—Part One: The Scope of Rationalism .....</b>      | <b>237</b> |
| 7.1 The Generality Objection .....   | 238        |
| 7.1.1 Generality Objection One: Perceptual Beliefs .....   | 238        |
| 7.1.2 Generality Objection Two: Propositional Attitudes .....  | 243        |
| 7.2 The Scope of the Rationalistic Interpretation of the Transparency Method .....                             | 248        |
| 7.3 The Prospects of Extension .....   | 258        |
| 7.4 Conclusion .....   | 263        |
| <b>Chapter 8 Extending the Transparency Method Beyond Belief—Part Two: Rationalism versus Empiricism .....</b> | <b>265</b> |
| 8.1 The Passivity Objection.....   | 266        |
| 8.2 Transparency, Empiricism and Rule Following .....  | 270        |
| 8.3 Judgement-Sensitive Belief.....  | 276        |
| 8.4 Transparency and Desire .....  | 281        |
| 8.4.1 Byrne’s Empiricist Approach to the Transparency Method—Desire .....                                      | 282        |
| 8.4.2 The Rationalist Approach to the Transparency Method—Desire.....  | 290        |
| 8.5 Transparency and Intention .....   | 295        |
| 8.5.1 Byrne’s Empiricist Approach to the Transparency Method—Intention .....                                   | 295        |
| 8.5.2 The Rationalist Approach to the Transparency Method—Intention.....                                       | 299        |
| 8.6 Extending the Transparency Method Beyond Belief, Desire, and Intention .....                               | 303        |
| 8.7 Conclusion .....   | 305        |
| <b>Conclusion .....</b>  | <b>307</b> |
| <b>References .....</b>  | <b>313</b> |

*First-Person Authority and its Limits*



## Introduction

Philosophy is a queer pursuit, reckoned, as it sometimes is, to be the most sublime, and sometimes to be the most trivial of human occupations.  
William James ([1903–1904] 1988, p. 3)

Philosophers writing about the topic of self-knowledge have, broadly speaking, focused on three main areas. First, there is the knowledge we have about ourselves—such as our character traits, our personalities, how best to live our lives, and other self-knowledge commonly associated with wisdom. Socrates considered this type of self-knowledge the most important we could achieve, and in Plato’s *Phaedrus* is famously depicted as ignoring Phaedrus’ questions about the local myths until he has followed the Delphic Maxim ‘know thyself’ (1997). Similar perspectives on this type of self-knowledge have been given by more recent philosophers such as Thomas Hobbes, who encouraged his readers to ‘*read thyself*’ ([1651] 1996, p. 8) in order to gain knowledge of the actions and motivations of others; Jean Jacques Rousseau ([1755] 2006), who sought self-knowledge to conquer inequality; and Frederick Nietzsche, who argued that we are often strangers to ourselves—claiming “[e]veryone is furthest from himself”...of ourselves, we have no knowledge’ ([1887] 1998, p. 1).

The second type of self-knowledge that has received prolonged attention is the knowledge we have about the self. Here philosophers have been interested in questions such as ‘What does the ontological self actually refer to?’ (see, e.g., Klein 2014), ‘Do we actually experience any phenomenal quality relating to the self, as in a phenomenal ‘I?’’ (see, e.g., Hume [1739–40] 2000; Prinz 2012), ‘What is the nature of the persisting self over time?’ (see, e.g., Strawson 2009;

Dainton 2008), and ‘Is the self an illusion?’ (see, e.g., Dennett 1991; Albahari 2006).

As difficult, varied, and penetrating as these issues are, they are not the types of self-knowledge that, primarily, I will be concerned with in this dissertation. Instead, I will investigate a third type of self-knowledge that has, unlike the types mentioned above, generally been considered much easier to acquire, as well as less philosophically troubling. This is the knowledge we have of our own mental states: such as our thoughts, sensations, beliefs, intentions, and judgements—the kind of knowledge many think we can acquire by *introspection*. Since this type of self-knowledge is thought to be relatively easy to acquire, it is generally agreed that each of us are in a privileged, authoritative, and unchallengeable position to determine whether we have a headache, what we think about the prime minister’s latest policy, and whether we desire to watch the latest *Star Wars* film. Incorrigeability—the epistemic thesis that one’s own psychological self-ascriptions can never be corrected by another person—appears to be, as Richard Rorty (1970) once put it, the mark of the mental.

Although this way of thinking about self-knowledge and introspection has historically found acceptance in several different philosophical traditions, it is most notably exemplified in the writings of René Descartes, who famously argued that one cannot be mistaken whenever one expresses the thought ‘I am, I exist’ ([1684] 1984, p. 18) (known in the literature as the *cogito*). According to Descartes’ *cogito* argument, even if one is being deceived by an evil demon, and no external world exists—or even any other minds—one can know, with *absolute certainty*, that one is a *res cogitans*: a thinking thing. By extending this

same argument, Descartes thought that there existed a class of mental states that could be known with equal certainty:

it is also the same “I” who has sensory perceptions, or is aware of bodily things as it were through the senses. For example, I am now seeing light, hearing a noise, feeling heat. But I am asleep, so all this is false. Yet I certainly *seem* to see, to hear, and to be warmed. This cannot be false; what is called “having a sensory perception” is strictly just this, and in this restricted sense of the term it is simply thinking ([1641] 1985, p. 19).

Although there is controversy over whether we should interpret passages such as this one as expressing Descartes’ view that we have infallible access to our mental states, it is less controversial to posit that Descartes viewed self-knowledge as highly reliable, immediate, and acquired in a way that is different from the way in which each of us achieve other types of knowledge.

This view of self-knowledge was also held by philosophers after Descartes. John Locke, for instance, advocates the strong claim that it is ‘impossible for any one to perceive, without perceiving that he does perceive. When we see, hear, smell, taste, feel, meditate, or will anything, we know that we do so’ ([1690] 1975, p. 9). And David Hume, who is well known for his scepticism, agreed with Descartes and Locke about the security of our self-knowledge, as he states in the following passage.

For since all actions and sensations of the mind are known to us by consciousness, they must necessarily appear in every particular what they are, and be what they appear. Every thing that enters the mind, being in *reality* a perception, ’tis impossible any thing shou’d to *feeling* appear different. This were to suppose, that even where we are most intimately conscious, we might be mistaken ([1739–40] 2000, T 1.4.2.7; SBN 190)

This way of thinking about self-knowledge has not just been limited to those authors writing in the western tradition. Similar views can be found in ancient

Indian, Chinese, and Aztec writings.<sup>1</sup> According to these views, the knowledge of our own mental states is privileged, authoritative, and achieved in a way that is different from the way that others, from the third-person point of view, can achieve knowledge of our minds.

However influential this way of thinking about self-knowledge has been, there has, in recent decades, emerged several different empirical discoveries that have led some authors to question just how justified we are in holding this view. These include, but are not limited to, the development of psychoanalysis (see, e.g., Freud [1916-1917] 1966), studies into the prevalence of self-deception (see, e.g., Trivers 2011), research into the nature of cognitive dissonance (see, e.g., Egan, Santos, and Bloom 2007), and a greater appreciation of the extent to which the unconscious plays a role in our lives (see, e.g., Wilson 2002).

In addition to these empirical data, the emergence of certain philosophical doctrines, such as content externalism—the view that the meanings of our thoughts are determined externally and ‘not in the head’ (see Putnam 1975)—have led some philosophers (see, e.g., Boghossian 1989) to doubt whether we do have privileged access, or first-person authority, to our own mental states.

While such considerations have motivated some authors to reconsider the view that our self-ascriptions are *always* reliable, others have taken more radical positions. Some authors have denied that there is *any* special, or unique, way in which we achieve such self-knowledge. One of the best-known examples of this rejection came in the mid-twentieth century from Gilbert Ryle (1949), who argued that we acquire self-knowledge of our beliefs, intentions, and desires in much the same way that others learn about our own minds—that is, by gathering

---

<sup>1</sup> For a discussion about how these various cultures theorised about the mind, see Carruthers (2011, pp. 30–31). Carruthers (2008) argues that ‘Cartesianism’—the view that if one believes one is in a mental state, then one is in that mental state—is universal to the human species.

evidence from behaviour, speech, and other external phenomena. Since we lack any special way of achieving such self-knowledge, he thought that the only sense in which we have *first-person authority* or *privileged access* to our own psychology is in the sense that we have *more* data in our own case, rather than having access to a different *kind* of self-knowledge. This focus on external behaviour was consistent with the doctrine of analytic behaviourism—which is present in the work of Ryle—which focused on third-person verifiable data, rather than on anything internal.

Although analytical behaviourism is no longer a widely held view amongst contemporary philosophers and psychologists, there remains opposition to the idea that each of us stand in a privileged, or authoritative, position with respect to the way in which we acquire knowledge of our mental states. Peter Carruthers (2011), for instance, argues that the way in which we come to have self-knowledge of most of our mental states is no different in principle from the way in which we achieve knowledge of the mental states of others. He cites empirical data from psychology and neuroscience to support this view. Alison Gopnik (1993) advocates a similar view. She cites evidence from developmental psychology, to argue that the knowledge of many of our mental states is achieved in an interpretative way, rather than by a direct introspective method. Daniel Wegner (2002) and Michael Gazzaniga (1998), who also focus on evidence from psychology and neuroscience, draw similar, though less radical, conclusions. Both are sceptical of the idea that we have the kind of special access to our own psychology that we typically take ourselves to have.

Daniel Dennett (1991) offers a much broader scepticism, and suggests that when we take ourselves to be introspecting we are ‘actually engaging in a

sort of impromptu theorizing—and we are remarkably gullible theorists’ (1991, pp. 67–68). And Eric Schwitzgebel (2012b) enlists various empirical data to support the thesis that our introspective reports are prone to error—which leads him to reject the claim that introspection is as reliable as it is generally taken to be (see also Hurlburt and Schwitzgebel 2007).

What are the implications of accepting such scepticism, and why does it matter? It matters, in my view, because this scepticism has the potential to radically change the way that we view first-person authority—the thesis, broadly speaking, that our self-ascriptions about what mental states we are in are more likely to be true, compared to the ascriptions that others make about our own minds. If the sceptical accounts are correct, then the common-sense intuition that there is an asymmetry between the way in which we know our minds, compared to the way in which we come to know the minds of others, may have to be jettisoned. This intuition is not just held by professional philosophers, but is rather one that is accepted by most human beings. Denying that first-person authority exists, therefore, would have a profound effect on the way that we conceive of ourselves in relation to other people.

One of the problems that confronts us, with respect to addressing such scepticism, is that there is no universally agreed upon conception of first-person authority that is found in the literature. For instance, if one equates first-person authority with infallibility—the psychological thesis that we cannot be wrong about the nature of our own minds—then first-person authority may need to be rejected, on the grounds that we sometimes make mistakes in our psychological self-attributions. However, we do not need the results from psychology or neuroscience to show that such a thesis is false. It is, after all, part of our

everyday common-sense understanding of our own psychology, in our current age, that we can sometimes make mistakes in our own self-ascriptions. If we do not accept that such an account of first-person authority is accurate, however, then what *should* it mean to say that each of us has first-person authority?

The central claim that I will defend in this dissertation is the thesis that introspection can explain first-person authority—in a way that does not require infallible knowledge of one’s mental states. What is introspection, however? William James stated that ‘the word introspection need hardly be defined—it means of course, the looking into our own minds and reporting what we there discover’ ([1890] 1981, p. 185). This description gives the impression that introspection is a kind of inner perception, and that first-person authority arises from our ability to utilise this faculty. Although such a view has its defenders, it has been met with criticism in recent years, with views such as the transparency method, the self-shaping view, and neo-expressivism, all attempting to explain first-person authority without appealing to a special inner mechanism of self-detection.

Given that such views deny that self-knowledge is acquired by inner perception, should we still classify them as introspective? In my view, we should. In this dissertation, I will argue that the word ‘introspection’ should be seen as a theory-neutral term that can be used to describe the process by which a subject can acquire knowledge of her mental states. However, since not all processes should be described as introspective, a set of criteria is required to distinguish introspective from non-introspective self-ascriptions. The contemporary literature on self-knowledge lacks such a well-defined distinction, which impedes our ability to understand not only the kinds of scepticism that I

listed above, but also the nature, and limits, of first-person authority. In what follows, I offer such a set of criteria by offering what I will call the *theory-neutral account* of introspection.

If I am right that introspection, in this theory-neutral sense, can explain first-person authority, then we are able to answer the question ‘What are the limits of first-person authority?’ by answering the question ‘What mental states can be introspected?’ For instance, if one can introspect one’s own sensations, then one will have first-person authority with respect to one’s sensations. If one cannot introspect their height or age, then one will lack first-person authority with respect to their height or age.<sup>2</sup>

Such a construal of first-person authority makes it clear just how certain sceptical approaches to self-knowledge can potentially undermine first-person authority. If one cannot introspect a certain type of mental state, then one cannot have first-person authority with respect to that mental state. Consider Carruthers (2009, 2010, 2011), for example, who defends the position that we cannot introspect most our own propositional attitudes. He claims:

[p]hilosophers...are virtually united in thinking that there is introspection for judgments and decisions, just as there is for perceptual and imagistic states...No doubt this is partly because some philosophers are unaware of the relevant empirical evidence and other empirical considerations. But it is also because philosophers’ views tend to be much more driven by intuitions than by empirical evidence. And there is no doubt at all that we have a powerful intuition of the existence of introspective access to our own judgments and decisions. I shall argue, however, that this is a *mere* intuition, without any rational ground (2010, pp. 82–83).

---

<sup>2</sup> This is, of course, not to say that one will not be able to know what one’s height or age is if one cannot introspect it.

Since Carruthers denies that we cannot introspect most of our propositional attitudes, he would deny that we have first-person authority, in the way that I construe it. In this dissertation, I will argue that such a position, as well as positions like it, are not supported by the empirical evidence. I will argue that the empirical evidence cited in support of such a position is inclusive, and does not show that we cannot introspect our propositional attitudes.

In addition to arguing against sceptical accounts of self-knowledge, and analysing first person authority in terms of introspection, I will also attempt to contribute to our understanding of the question ‘What types of self-knowledge can we introspect?’ I will do so by defending a view commonly referred to in the literature as the transparency method—a view I argue should count, on my theory-neutral definition, as a form of introspection. This is a view that is opposed to the traditional conception of introspection (such as James’), which construes it as a form of inner perception. According to the transparency method, one does not need to ‘look inside’ one’s own mind to acquire knowledge of one’s own mental states. Rather, one can attend to the content of what the mental state in question is about. According to the transparency method, one could know whether one believes that snow is white by answering the question ‘Is snow white?’

Drawing upon and expanding on recent work by Richard Moran (2001, 2012) and Matthew Boyle (2009, 2011), I will argue that the transparency method is not only a plausible way in which to explain the self-knowledge we can have of our mental states, but it can also account for human rationality—a feature of self-knowledge that I will argue should not be, but often is, discounted from accounts of self-knowledge. I will also argue that the application of the

transparency method is wider than some think, meaning I think it is applicable to other types of propositional attitudes (apart from belief) such one's desires, intentions, wishes, hopes, and so on.

I structure the dissertation in the following way. The first block of chapters, 1–2, will focus on conceptual issues pertaining to the nature of introspection and first-person authority. In chapter 1, I give a novel theory-neutral account of introspection. There, I seek to determine what it means for a process to be thought of as introspective (as opposed to what theory of introspection is correct). In chapter 2, I integrate the theory-neutral account of introspection into an account of first-person authority. There, I argue for a view called introspection justification. On this view, a subject has introspective justification for believing that she is in mental state *M*, if and only if that subject has justified her belief that she is in mental state *M*, by introspecting that she is in mental state *M*. I will argue that while other senses of first-person authority may be devised, the introspective justification account of first-person authority is superior. The central claim I argue for is that introspection can explain first-person authority.

In the second block of chapters, 3–4, I address scepticism about introspection. In chapter 3, I consider two distinct challenges to the view that we can introspect our own propositional attitudes (e.g., our beliefs, desires, intentions, hopes, fears, wishes and so on). First, I discuss the philosophical doctrine of content externalism. This is the doctrine which says that in order to have knowledge of one's own thoughts, one needs to go beyond what is 'in the head'—meaning that one would need to rely on third-person verifiable evidence, or observe one's own environment, in order to know one's own thoughts. I argue

there is no conflict between content externalism and one's ability to acquire introspective knowledge of one's own thoughts. In the second part of the chapter, I consider inferentialism—the thesis that the way in which we achieve knowledge of our own propositional attitudes is no different in kind from the way in which we acquire knowledge of the minds of others. I examine one version of inferentialism—Peter Carruthers' Interpretive Sensory-Access (ISA) theory. According to this theory, we cannot introspect most of our propositional attitudes. After an exposition of the view, I raise several problems with it. In chapter 4, I consider the empirical evidence that is cited in support of the ISA theory. I argue that the evidence is inconclusive, and does not support the theory.

The third block of chapters, 5–6, seek to offer a positive account of introspection for propositional attitudes. In chapter 5, I defend the transparency method. I argue that it can offer a plausible account of how one can know what one believes. I also challenge recent attempts to account for transparency without appealing to rational agency. In chapter 6, I offer replies to various objections that have been recently made to this view.

The fourth block of chapters, 7–8, are focused exclusively on extending the transparency method beyond belief. In chapter 7, I consider arguments given by philosophers who think that the transparency method *cannot* be extended beyond belief. I argue that the reasons typically given for upholding such a position are untenable. In chapter 8, I show how the transparency method can be extended to other mental states such as desire and intention. I also show how one might go about extending it even further. As with the attitude belief, I argue that rationality cannot always be eliminated from the process of achieving self-knowledge of such mental states, contrary to what some authors think.

In addition to arguing that introspection can explain first-person authority, the other main claim that I will defend is the thesis that rationality cannot be eliminated from our explanation of self-knowledge. Although it is tempting to think of first-person authority as simply the result of one's ability to 'glance inside' to a realm in which only the subject has access, I will argue that such a position cannot be the whole story. To accurately reflect our ability, as human beings, to hold mental states for reasons, we also need to invoke rationality. It will be my aim to show how rationality, contrary to what is often claimed, can be understood as an epistemic notion.

## Chapter 1

### What is Introspection?

The problem with introspection is that it *acquiesces* in the illusion that there is an inner eye that sees.  
Daniel Dennett (2017, p. 185).

William James wrote that ‘the word introspection need hardly be defined—it means, of course, the looking into our own minds and reporting what we there discover’ ([1890] 1981, p. 185). This description gives the impression that introspection is a kind of inner perception, and that privileged access to our own mental states arises from our ability to utilise this faculty. The inner perception theory, however, has been met with criticism in recent years with views such as the transparency method (see, e.g., Moran 2001; Byrne 2005a), the self-shaping view (see, e.g., McGeer and Pettit 2002), and neo-expressivism (see, e.g., Finkelstein 2003; Bar-On 2004) all attempting to explain privileged access/first-person authority without reference to inner perception.<sup>1</sup> Given that these views deny that privileged self-knowledge is acquired by inner perception, there is a question whether we should still classify them as ‘introspective’.

In this chapter, I will argue that the word ‘introspection’ should be understood as a theory-neutral term—one broad enough to encompass several different views that attempt to explain self-knowledge, including some that deny inner perception. However, since not all ways of achieving self-knowledge should be described as introspective, a set of requirements is needed to distinguish introspective from non-introspective self-ascriptions. The current literature on self-knowledge lacks such a well-defined distinction, which impedes our ability to understand the nature of the privileged access/first-person

---

<sup>1</sup> For a more comprehensive list of these rival views, see Eric Schwitzgebel (2014).

authority debate.<sup>2</sup> This chapter will put forward such a set of requirements. I will call this the theory-neutral account of introspection.<sup>3</sup>

Although I think there are several benefits of having a theory-neutral account of introspection—some of which will be explained as we proceed—the main benefit that I will be interested in explicating, in this chapter and the next, is how introspection, broadly construed, can help to explain the nature of first-person authority. This chapter will specify what it takes for a process to be classified as introspective. Then, in the next chapter, I will argue that introspection, in the theory-neutral sense, can explain first-person authority. It thus becomes important that we are explicit about what features a process must possess in order to count as introspective.

This chapter will proceed as follows. I begin, in §1.1, by introducing the inner perception view. In §1.2, I look at alternative accounts of self-knowledge to the inner perception view. In §1.3, I consider some terminological issues which relate to the etymology of the word ‘introspection’. In §1.4, I state how I think introspection should *not* be characterised. In §1.5, I present a theory-neutral account of introspection. In §1.6, I consider what types of self-knowledge can be known by introspection.<sup>4</sup>

---

<sup>2</sup> This chapter and the next will attempt to explain why this is. Although I use the terms ‘first-person authority’ and ‘privileged access’ interchangeably here, I will argue, in the next chapter, that it is useful to think of them as distinct.

<sup>3</sup> It is worth pointing out that when I talk about self-knowledge, I am talking about the justified true beliefs that one has about oneself. What a theory of self-knowledge seeks to explain is how it is that one comes to possess such beliefs. One may or may not invoke introspection to explain how one can have such self-knowledge. The inner perception view, for example, is a view that does invoke introspection to explain self-knowledge. The question I seek to address in this chapter is ‘Should rival accounts to the inner perception view also be thought of as ‘introspective’ ways of achieving self-knowledge?’

<sup>4</sup> In asking the question ‘What types of self-knowledge can be known by introspection?’ I am asking a question about the source of such self-knowledge.

## 1.1 Introspection as Inner Perception?

A distinct feature of human cognition is that we can know, and when needed, report, our current psychological mental states such as our sensations, beliefs, desires, and intentions.<sup>5</sup> This faculty engenders a wide range of cognitive abilities, many central to our lives, such as decision-making, moral deliberation, and long-term planning.<sup>6</sup> Such self-knowledge also facilitates many of the cooperative tasks we undertake, which require us to successfully inform others of what we desire, intend, believe, and so on.

Although this may be familiar, and uncontroversial, several basic questions arise from these preliminary remarks. First, how do we come to have knowledge of these facts about our own psychology—that is, what is the source of this self-knowledge? Second, is the way that we gain knowledge of our mental states different from the way that others, from the third-person perspective, gain knowledge of them? And third, how reliable are these self-ascriptions? Could one erroneously believe that one supports liberalism, or that one intends to travel to Egypt in the summer? One way to answer these questions is to appeal to the notion of *introspection*—the cognitive faculty responsible for the private, secure, and immediate way that we have knowledge our own mental states.

---

<sup>5</sup> What it is for a subject to have knowledge of a proposition is a matter of some controversy amongst contemporary epistemologists. Ever since Edmund Gettier's seminal paper, 'Is Justified True Belief Knowledge?' (1963), many philosophers have found the once popular explanation of knowledge as 'justified true belief' untenable. Contemporary philosophers still hold that justification, truth, and belief are central to knowledge, however. Jennifer Nagel describes the current orthodox view of knowledge as follows: 'in order to know a proposition *p*, an agent must not only have the mental state of believing that *p*, but various further independent conditions must also be met: *p* must be true, the agent's belief in *p* must be justified or well-founded, and so forth' (2013, p. 281). When I refer to 'knowledge' in this work, it is this sense that I shall have in mind. For an alternative view, which posits knowledge as a mental state, see Timothy Williamson (2000).

<sup>6</sup> As some philosophers (see, e.g., Davidson 1982; Heil 1992) deny that non-human animals, or very small children, have complex mental states such as beliefs or intentions—because they don't have language—I will, in order to avoid entering into such controversy, focus solely on the introspective capacity of adult human beings. In what follows, nothing I will discuss here will bear directly on this issue. Other philosophers (see, e.g., Dennett 1991, pp. 77–78), make the distinction between belief and opinion, and suggest that while animals *can* have beliefs, they cannot have opinions.

What is introspection though? As I mentioned, James thought the term itself could be straightforwardly defined—noting that it was ‘the looking into our minds and reporting what we there discover’ ([1890] 1981, p. 185). Now, while this provides us with a basic understanding of what is typically meant by the term ‘introspection’, it does not tell us much about how the process can yield self-knowledge, how reliable it is, or what its nature is.<sup>7</sup> For instance, should we think of introspection as a form of inner perception—as is suggested by James’ usage of the words ‘looking’ and ‘reporting’? If we want to adhere to a strict etymological interpretation, then it seems like we should. The term ‘introspection’, after all, originates from the Latin word *introspicere*: which literally means ‘to look into, look at, examine, observe attentively’. The word comprises of *intro*, which means ‘in, on the inside, within’; and *specere*, which means ‘to look at’.<sup>8</sup>

According to some philosophers, this is exactly how we *should* characterise introspection: as a form of inner perception. John Locke—one of the earliest proponents of this view—said ‘though it be not Sense, as having nothing to do with external Objects; yet it is very like it, and might properly enough be call’d internal Sense’ ([1690] 1996, p. 105).<sup>9</sup> More recent philosophers, such as David Armstrong (1968) and William Lycan (1996), have followed Locke in this contention. On this view, human beings possess a faculty of inner sense that is capable of giving one knowledge of one’s own mental states.<sup>10</sup>

---

<sup>7</sup> Assuming introspection can yield self-knowledge, of course.

<sup>8</sup> See Harper (2017).

<sup>9</sup> This quotation from Locke was brought to my attention from a paper by Angela Coventry and Uriah Kriegel (2008, p. 224).

<sup>10</sup> The inner sense view is not the only approach to self-knowledge that can be considered as a form of inner perception. For a rival view, see Brie Gertler (2012a), who defends a view called the ‘acquaintance approach’.

Although these authors stress the similarity between inner sense and outer sense (e.g., visual and auditory perception), there are, of course, some important key differences to note. One difference is that inner perception requires no inner eye (see Moran 2001). Another difference is that inner perception doesn't require better or worse lighting conditions—as visual perception does (see Hacker 2013). Despite these differences, proponents of the inner sense view hold that inner perception is the right way to think about introspection; and that one can determine what one believes, intends, desires, feels, and so on, by *looking inside*—or as Lycan puts it, by using one's 'internal monitor' (1996, p. 17).<sup>11</sup>

This type of account can provide us with answers to the preliminary questions that were posed above as follows. On the inner perception view, self-knowledge is privileged because one's inner sense can only deliver knowledge of one's own mental states—e.g., my inner sense cannot tell me that another person is in pain.<sup>12</sup> It is, thus, a process different from the way that one comes to have knowledge of another person's mental states. And like outer perception, such as seeing and hearing, it can be thought of as generally reliable—though susceptible to breakdowns.<sup>13</sup>

While the inner perception approach may, in the end, provide us with a true account of introspection, we should *not*, in my view, equate introspection with inner perception. That is, our *pre-theoretical* notion of introspection should be broader. The aim of this chapter will be to put forward a pre-theoretical account

---

<sup>11</sup> This view of introspection is held by proponents of the higher-order perception (HOP) theory of consciousness. On this view, consciousness is explained in terms of a relation between a conscious state, and a higher order representation of that state. According to Wesley Sauret and William Lycan, this is the view that 'a mental state is conscious just in case it is itself represented in a quasi-perceptual way by an internal monitor, scanning device or attention mechanism' (2014, p. 363).

<sup>12</sup> One could still know that someone else is in pain, of course. But one would need to use their outer sense.

<sup>13</sup> For discussion about why the possibility of breakdowns may be problematic, see Akeel Bilgrami (2006, pp. 121–123).

of introspection that is broad enough to encompass multiple views of self-knowledge.

The question I answer here can be stated as follows: ‘What does it mean to say that a process is introspective?’ To get a sense of the kind of question I am asking, it is helpful to think of the sorts of questions one would ask before constructing a theory of free will or colour. One would ask, for example, ‘What do we mean when we say that a person has free will?’ or ‘What does it mean to say that an object has a certain colour?’<sup>14</sup> Once such preliminary questions have been answered, then we can assess further questions such as ‘Does anyone ever exercise free will?’ or ‘Does any object have colour?’<sup>15</sup> What I propose to do here is the same for introspection. My claim is that the inner perception view is only *one* theory of introspection—one of several. I will articulate a pre-theoretical account of introspection, in order to see what other views of self-knowledge can be thought of as introspective too.

## 1.2 Alternatives to Inner Perception

As the inner perception view remains a controversial position in the literature, alternative positions have been developed in the last few years. In this section, I will mention briefly a number of these views in order to show that some authors have attempted to explain how we can have self-knowledge *without* reference to

---

<sup>14</sup> The process is not always this simple, however. In the case of colour or free will the semantic and metaphysical debates can inform each other. For instance, theoretical reasons may force a semantic shift away from our what we meant pre-theoretically by free will or colour.

<sup>15</sup> Such questions are the kinds typically associated with traditional conceptual analysis (see Frank Jackson 1998). I am not seeking to radically revise the conception of introspection, as a revisionist might attempt to do. Manuel Vargas (2005) gives an example of revisionism in his paper ‘The Revisionists Guide to Responsibility’. Here, Vargas argues that we should moderately revise the concept of moral responsibility.

inner perception.<sup>16</sup> It is not my aim to defend any particular view in this chapter, nor provide an exhaustive list of rival views. Rather, I mention a select few in order to provide a sense of the kinds of differences that exist among the various alternatives to the inner perception view. I will then address the question: ‘What does it take, for a process that yields self-knowledge, to count as introspective?’

The first view I will mention is the *transparency method*. Its proponents include Gareth Evans (1982), Richard Moran (2001), Alex Byrne (2005a), Akeel Bilgrami (2006), and Matthew Boyle (2011). The view’s defining characteristic can be expounded by looking at the following passage from Evans:

in making a self-ascription of belief, one’s eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me, ‘Do you think there is going to be a third world war?’, I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question ‘Will there be a third world war?’ I get myself into the position to answer the question whether I believe that *p* by putting into operation whatever procedure I have for answering the question whether *p* (1982, p. 225).

What is noteworthy about this passage is Evans’ claim that it is *outward phenomena* that must be attended to. This can be contrasted with what proponents of the inner perception view say—namely, that one should focus on *inward phenomena* in order to know what one believes. This point is elaborated by Boyle who says:

I can know various aspects of the nature, content and character of my own mental states by attending in the right way, not to anything “inner” or psychological, but to aspects of the world at large (2011, p. 226).

---

<sup>16</sup> As I mentioned above, most authors would grant that the inner perception view is form of introspection, regardless of whether or not they accept the view.

What Boyle is suggesting in this passage is that it is the content of the mental state that should be attended to, if one wants to have knowledge of one of their own mental states, rather than the psychological mental state itself. Although I will examine this view in some more detail in the forthcoming chapters, we can get a basic idea of how the procedure is supposed to work by considering Evans' example about the third world war. Suppose I wanted to know whether I actually do believe that there will be a third world war. On the transparency method, one can find out whether one does believe this by attending to a question about the world (something not *psychological*, as Boyle says above)—and make up one's mind about the question. Proponents of this view claim that doing so will give one transparent access to one's own mental states. The general idea is that one can determine whether one believes that *P* by answering the question 'Is *P* true?'

This contrasts with the inner sense view—which *does* require one to detect something psychological. According to this view, if I want to know whether I have a certain belief—e.g. 'Do I believe that there will be a third world war?'—I must detect, or observe, an inner psychological state—such as a feeling that I do currently hold that belief.

The second alternative to the inner perception view I will look at is best referred to as the *self-shaping view*. It is defended by Victoria McGeer and Phillip Pettit (2002). On this view,

self-knowledge of one's beliefs [in human beings] comes about, not as a result of having a special insight into whether one has the dispositions that make one a believer that *p* or that *q*, but as a result of having a special ability to develop the dispositions that make one a believer that *p* or that *q* (2002, p. 293).

The special ability that McGeer and Pettit are referring to here is the ability of human beings to commit themselves to certain ways of thinking and acting, with respect to certain propositions. On this account, I do not learn of my belief that there will be a third world war by *detecting* any mental phenomena. Rather, I do so by cultivating certain dispositions, or commitments, that pertain to the proposition ‘there will be a third world war’.

The third alternative to the inner perception view that I will mention is best referred to as *neo-expressivism*.<sup>17</sup> This is a view defended by Dorit Bar-On (2004, 2009) and David Finkelstein (2003). According to neo-expressivism, one does not come to know about the contents of one’s own mind in terms of *detecting* or *observing* certain mental phenomena; but rather, by *expressing* one’s own mental states.<sup>18</sup> According to Bar-On, one’s self-avowals ‘constitute a certain class of *expressive acts* in which a subject gives articulate vent, in speech or in thought, to present mental states’ (2009, p. 69). This means that when one asserts that one is in a mental state, or has a certain thought—e.g., when one says ‘I am excited’—one is not merely detecting an existing mental state, but is expressing that mental state.

Other views, still, combine various ideas that feature above. Akeel Bilgrami (2006) combines the transparency method and the self-shaping view. Eric Schwitzgebel (2012a) defends the thesis that introspection shouldn’t be conceived of as a *single process*—for example, as *either* the inner perception view or the transparency method, but rather as a *plurality of processes* that could include some, or even all, of the views mentioned above.

---

<sup>17</sup> Neo-expressivism differs from simple expressivism—a view that is sometimes attributed to Wittgenstein—in that it allows for truth-conditions to be adopted. On a simple expressivism view, expressions are not truth-evaluable.

<sup>18</sup> Bar-On and Falkenstein are both opposed to the inner sense view. See Bar-On’s (2004) chapter 3; and see Finkelstein’s (2003) chapter 1.

Some argue that the way in which we know our own minds is no different in principle from the way we know the minds of others. Gilbert Ryle (1949) famously argued that there is no special way in which we know our own minds. He said that self-knowledge of our mind is ‘not attained by...introspection’ (1949, p. 149). And Peter Carruthers has recently argued that ‘self-knowledge of most forms of thought doesn’t differ in kind from knowledge of the thoughts of other people’ (2011, p. xii). On this view, knowledge of our own minds is acquired by inference—just like it is when it comes to the knowledge that one can have of the mind of another. Call this view *inferentialism*.<sup>19</sup>

I have, of course, not been attempting to provide an exhaustive overview of the competing accounts of self-knowledge that have been explored in the recent literature. Instead, I have shown from this small sample of views the extent to which some philosophers have sought to distance their own views from the inner perception view. Before assessing the plausibility of these views, I will first address the question of whether we should or should not consider each of these views as introspective. To do this, I will propose a set of criteria that will allow us to distinguish introspective from non-introspective processes.

### **1.3 Terminological Concerns**

Before offering a set of criteria which will allow us to distinguish introspective from non-introspective processes, I will first raise some concerns with attempting to characterise introspection as broadly as I intend to do. These concerns will be largely terminological. Recall, as was noted above, that the term ‘introspection’

---

<sup>19</sup> I go into more detail about inferentialism in chapter 3. As I understand it, this is the view that there is no significant asymmetry between the way in which we acquire knowledge of our own minds, and the way in which we have knowledge of the minds of others. See Cassam (2014, chapter 11) for another recent defence of this view.

originates from the Latin word *introspicere*, which means to look inward. Given this, some authors have expressed concern with using the word ‘introspection’ as a neutral term that can be divorced from inner perception. Akeel Bilgrami, for example, says

it is misleading to use the term “introspection”...as a neutral term intended to be synonymous with “self-knowledge” in general....The verb “to introspect” suggests a kind of cognition...that tilts the usage in favor of the perceptualist [inner perception view] (2006, p. 38).

Another is Schwitzgebel, who says, with specific reference to the transparency method, ‘transparency accounts stress the outward focus of our thought in arriving at self-ascriptions, calling such...[views] accounts of “introspection” strains against the etymology of the term’ (2014). And others such as Bar-On have characterised the contemporary debate about self-knowledge as one between ‘introspectionist’ views (by this she means the inner sense view) and ‘alternative’ views, which include the transparency method and neo-expressivism (2004, p. 22).

In my view, these are not strong reasons to equate inner perception with introspection. I think that if we are careful to specify just what is meant by introspection, then we can arrive at a theory-neutral account that is both sensitive to the intuitions that inner sense theorists hold, yet is, at the same time, broad enough to include other views that are not characterised in terms of inner perception.

I do not wish to give the impression that all philosophers have equated introspection with inner perception. In militating against this way of construing introspection, I follow several other philosophers. Moran, for example, says

‘identifying introspection with a kind of perception is a substantive interpretation of *immediacy* and not equivalent to it’ (2001, p. 11). Charles Siewert says ‘while the term “introspection” seems practically inevitable, we can elect to use it—as I will—without committing ourselves to the existence of an “inner sense”’ (2012, p. 129). And Fred Dretske says:

“introspection” is just a convenient word to describe our way of knowing what is going on in our own mind, and anyone convinced that we know—at least sometimes—what is going on in our own mind and, therefore, that we have a mind and, therefore, that we are not zombies, must believe that introspection is the answer we are looking for. I, too, believe in introspection (2003, p. 7).

I agree with all three philosophers: we should not equate introspection with inner sense. What should we equate it with, then? In the passage directly above, Dretske claims that the word ‘introspection’ is a just convenient way of describing the process of learning about what is going on in our own mind. But surely not all ways of knowing what is going on in one’s own mind should be characterised as ‘introspective’. If I know that I am angry by catching my reflection in a mirror, or I know that I intend to leave the cinema because I find myself reaching for my coat, it does not seem like I have used introspection. Thus, there are ways of knowing about my own mental states that do not seem best described as ‘introspective’.<sup>20</sup>

#### **1.4 What Introspection is Not**

I have suggested that a pre-theoretical account of introspection should not be equated with inner perception. I have also noted that not all ways of learning

---

<sup>20</sup> To reiterate: introspection is one way that one may come to have self-knowledge. I grant there may be others.

about the contents of one's mind should count as introspective. To help us formulate a set of criteria that will allow us to distinguish introspective from non-introspective processes, I will say a bit more about what introspection is *not*.

As a starting point for this discussion, it will be useful to briefly mention Gilbert Ryle's account of self-knowledge—one of the better-known views that is opposed to the inner perception model of introspection. On Ryle's account,

[t]he sorts of things I that can find out about myself are the same as the sorts of things I can find out about other people, and the methods of finding them out are much the same (1949, p. 155).

What is most striking about this passage is Ryle's denial of the claim that there is any significant asymmetry between the way that one knows about one's own mental states, and the way that one knows about the mental states of others. On Ryle's account, I would, for example, learn that I am *angry* by looking in a mirror; that I am *jealous* of Samantha's success because Samuel tells me I look jealous; that I am nervous because I am shaking; and so on.<sup>21</sup> Since such ways of learning about my own mind are the same ways that another person who is interested in learning about my mind would use, it doesn't seem appropriate to characterise such processes as introspective.

I have not said much about the kind of asymmetry that would be entailed by an introspective process, but intuitively, an introspective process is one which would be significantly different from the one used to acquire knowledge of others' mental states. At a minimum, then, any theory that says that the process by which one learns about one's own mind is the same as the process that one learns about the minds of others, as Ryle contends, should not be classified as an

---

<sup>21</sup> I do not mean to suggest that even if one were to possess an inner sense, one would be unable to do these things.

introspective process. Let us try to be more specific. What else is introspection not? Carruthers suggests one answer. He thinks that if one has acquired knowledge of a mental state by introspection, then that process will *not* be an ‘*interpretative process*’ (2010, p. 76). By this, he means that introspective self-knowledge does not require one to interpret one’s own behaviour or speech. For example, suppose that I attribute to Tony the intention to leave Trish’s birthday party. One way I might do this is by observing Tony reaching for his car keys; recall that he was recently yawning; and hear him avow the phrase ‘it is getting late’. In attributing this intention to Tony, I interpret his behaviour and speech, and then conclude that he intends to leave the party. Carruthers thinks such a process does not occur with respect to an introspective process.

In addition to this negative claim about introspection, Carruthers thinks introspection entails the following positive claim:

introspection is a higher-order process, issuing in awareness or knowledge of (or at least beliefs about) the occurrence of token mental states. (On some accounts introspection needn’t always be reliable, any more than external perception is.) When I introspect a feeling of anger, for example, I become aware of that feeling, and come to know (or at least believe) that I am angry (2010, p. 76).

While I think that Carruthers’ positive and negative characterisations of introspection are to be preferred over views that say that introspection is a kind of inner perception, they remain overly theoretical. In the passage quoted above, Carruthers’ positive characterisation of introspection is well-suited to higher order theories of consciousness (see, e.g., Rosenthal 1986), but remains too theory specific.

In order to give a more general account of introspection, I will now state the conditions that I think a process must possess in order to be characterised as introspective. Although there is no general consensus amongst contemporary philosophers about what a process *must* possess in order to qualify as introspective, there are a number of features that are considered essential by some.<sup>22</sup> Schwitzgebel (2014) considers six conditions that have been recently discussed: (a) the mentality condition—that introspection only yields knowledge about mental states; (b) the first-person condition—that introspection only gives knowledge about one’s own mind, and no one else’s; (c) the temporal proximity condition—that introspection only delivers knowledge of current mental states; (d) the directness condition—that introspection gives direct knowledge; (e) the detection condition—that introspection involves the detection of pre-existing mental states; and (f) the effort condition—that introspection requires some sort of effort.<sup>23</sup>

According to Schwitzgebel, while most philosophers wouldn’t call a process introspective unless it satisfied (a)-(c), some wouldn’t call a process introspective unless it also satisfied (d)-(f). For example, some philosophers (see, e.g., Gertler 2011a) think that an introspective process will also need to meet condition (e), because they think introspection involves the detection, by some sort of internal scanner, of pre-existing mental states. If we were to accept this way of construing introspection, then we would have to deny that the transparency method and neo-expressivism are forms of introspection, because they do not involve inner perception.

---

<sup>22</sup> This is a point made by Schwitzgebel (2014).

<sup>23</sup> I should state that Schwitzgebel himself does not accept all of these conditions.

In my view, while the correct theory of introspection may end up consisting of some, or all, of the conditions listed above by Schwitzgebel, I do not think that all are required in a pre-theoretical account. It is important to note that the question I am addressing here is not ‘What conditions exist in the correct theory of introspection?’ or ‘What conditions exist in the inner perception view?’ I want to determine what conditions are required to count as *introspective*—regardless of the theory.

In what follows, I will propose three conditions a process must possess in order to be considered introspective. I will then discuss why some of the other conditions that Schwitzgebel lists here should not be included as part of a theory-neutral account of introspection.

## **1.5 The Theory-Neutral Account of Introspection**

In my view, there are just three features a process must possess in order to be classified as introspective. These are best referred to as the following:

- (i) The First-Person Condition
- (ii) The Difference Condition
- (iii) The Occurrent Condition

Each can be expounded as follows.

### **1.5.1 The First-Person Condition**

The first condition I will propose says that a process is introspective only if it reveals knowledge about one’s own mind, and *not* anyone else’s. If I have the capacity to introspectively know my own mental state *M*, then this process, if introspective, would *not* give me knowledge of any other minds. For example, if

I able to learn that I have a headache by using introspection, then I cannot learn about someone else's headache by this same process: introspection is, in this sense, exclusive. This is not to say I *cannot* learn that I have a headache by some other means, such as when I observe my own behaviour, or examine a brain scan. But since these methods are the same that someone from the third-person perspective would use to learn that I have a headache, they should not be thought of as introspective.<sup>24</sup>

Exclusivity is something that had been disused by several recent philosophers. Dan Zahavi, for instance, in a discussion about the problem of other minds writes 'the first-personal givenness of the mind of the other is inaccessible to me' (2005, p. 154). Zahavi is not claiming that one cannot have knowledge of another's mind. What he is describing is the sense in which knowledge of our own minds can be exclusive or unique. Sydney Shoemaker, similarly, talks of self-acquaintance—the view that the essence of the mind is that it has 'special access' (1996, p. 27) to its mental states.<sup>25</sup> What I think is important about what both Zahavi and Shoemaker refer to here is the sense in which some self-knowledge is restrictive. If there exists a way in which I can have knowledge about my own mind that can only give me knowledge, then I cannot know about the mind of another in the same way. One does not of course need to agree that it is true that such a way of achieving knowledge exists. But, what I do think we need to recognise is that such a characteristic is essential to

---

<sup>24</sup> There are some difficult cases that appear to complicate this condition, however. Susan Dominus (2011), for example, reports of a case concerning a set of two conjoined twins, both with separate brains, who appear to share some conscious experiences. And Schwitzgebel (2014) suggests that science fiction has provided us with some examples, such as telepathy, that might present us with a case where someone has direct access to the mind of *another*—a mind that is not one's own. My own view is that it is unclear how we should interpret these cases. With respect to the conjoined twins case, it may be that the twins share the same *type* of mental state, but not the same *token* mental state. In the telepathy case, it may be that a sophisticated form of mind reading is being used, and thus the first-person condition is not jeopardized.

<sup>25</sup> By acquaintance, I do not mean to suggest a form of dualism here. I take such a notion to be compatible with some forms of physicalism.

introspection. Anyone who denies that we can sometimes know our own minds in such a way, must deny that we can sometimes introspect.

This discussion elucidates one part of James' definition—namely, the part that says that introspection is the looking into one's own mind and reporting what one discovers. If we are careful in our formulation of this condition, and focus on the first-person ability, we can avoid any commitment to a self-scanning process, which is controversial in the literature. The acknowledgement of this condition leads us to the next. If introspection can only reveal knowledge of one's own mind, then it follows that introspective self-knowledge must be *different* from other ways of acquiring knowledge. How might this be so?

### **1.5.2 The Difference Condition**

The second condition that I will propose says that introspective self-knowledge is acquired in a way that is *different* from the way that one acquires other kinds of knowledge—such as the way that one acquires knowledge of one's environment, or knowledge of the mental states of others. This condition follows from the first-person condition which says that introspection can only give us knowledge about our own minds.

In what sense, then, is introspection different from other ways of acquiring knowledge, such as from perception or testimony? There are several ways in which introspection might be different. Declan Smithies and Daniel Stoljar, for example, consider both *psychological* and *epistemological* versions of what they call the 'difference thesis' (2012, p. 5). Psychological versions of the difference thesis attempt to show how introspection is psychologically different from other ways of acquiring knowledge (e.g., perception, testimony, and so on).

And epistemological versions attempt to show how introspection is epistemologically different from other ways of acquiring knowledge.

In my view, such differences are still too theoretical for a theory-neutral account of introspection. For example, the inner perception view of introspection may give a different answer to the question of how introspection is psychologically different, from other way of achieving knowledge, compared to the account a rival view might give. The difference I seek to account for in the difference condition is a difference that would be common to all theories of introspection.

The sense in which I think introspection is *different* from other ways of acquiring knowledge, that is common to all forms of introspection—and follows from the first-person condition—is that it is knowledge of a categorially different kind.<sup>26</sup> If introspection is a process that can only give one knowledge of one's own mental states, then it follows that introspection must be a process that is categorially different from the way that I learn about my environment (by perception); or the way that I learn about the minds of others (by testimony). Unlike these ways of acquiring knowledge, introspection does not require one to use *observation*, in the sense pertaining to sense perception (e.g., visual, auditory, and so on); or to *gather evidence*, in the sense that one reaches a conclusion by interpreting third-person verifiable evidence that one has gathered. Introspection is knowledge of a categorially different kind.<sup>27</sup>

To say that introspection is knowledge of a different kind, is to say more than just that it is *different* from other ways of acquiring knowledge. It is to say

---

<sup>26</sup> I follow Moran here who seeks to account for a form of knowledge that is 'categorially different in kind and manner' (2001, p. xxxi).

<sup>27</sup> One might think that the transparency method, with its focus on external phenomena, requires one to gather third-person evidence—thus bringing it into conflict with the difference condition. This thought should be resisted. In chapter 5, I go into more detail about why the transparency method is a form of introspection.

that introspection is significantly different. If I need to observe my own behaviour, or hear myself speak, in order to have knowledge of what I believe or intend, then in some respects this process may be different from the way that someone else might learn about my own beliefs or intentions. One difference is that I would have more data in my own case. I would be able to observe my own actions and avowals constantly, since others are only around me for some of the day. Although this would be *a* difference, it is not a difference *in kind*. It is the same process that others would use to acquire knowledge of my mental states. Views that deny that there is any significant asymmetry between the way one acquires self-knowledge and the way that others, from the third-person perspective, gain this same knowledge should be thought of as antithetical to introspection. Ryle (1949) and Carruthers (2011), for example, stress the similarity between the way we know our own minds and the ways we know the minds of others (such as observing behaviour). Carruthers, for instance, argues that the knowledge of ‘most of our own kinds of thinking...is no different in principle from the knowledge that we have of the mental states of other people’ (2011, p. xi). Because his view doesn’t conform to the difference condition, we can say that his view should not be counted as introspective.

The first-person condition and the difference condition cannot be all there is to introspection however, as we have left open the fact that some thoughts—retrospective thoughts about mental states—may be considered part of the reach of introspective knowledge. This is something we do not want to say. This leads us to the final condition to be considered.

### 1.5.3 The Occurrent Condition

The third and final condition I will propose says that introspective knowledge can only give one knowledge of one's *current, or very recently occurred*, mental states or thought processes. Although this point may seem trivial, it is an essential part of the theory-neutral account of introspection. This is because we need to distinguish between thoughts or utterances such as: 'I am now in pain' and 'I now believe that *War and Peace* is a long novel' from our memories of having such mental states such as: 'I was in pain last week' and 'I believed *War and Peace* was a long novel when I was in high school'. If I can introspect that I am in pain, it is of my current pain. The recollection, or memory, of this pain may be current, but it is not the original feeling of pain.

Because the predicate 'is current' is vague, we should not expect a precise explanation here. We do, after all, speak of 'the current year', 'the current administration' and 'the current stock prices'—phenomena that have very different time durations. The sense of current that I wish to capture in the occurrent condition is the sense that pertains to the *present moment of experience*. Even this, however, is not without controversy. As Terry Horgan and Uriah Kriegel note, '[i]t is not a trivial question what a "present" phenomenal experience is' (2007, p. 126). One suggestion that they consider is the so-called 'specious present'. This phrase, popularized by Williams James, refers to the present flow of experience which is, as Horgan and Kriegel point out, an '*episode* [of experience] lasting 2–3 seconds' (2007, p. 140 note 10).

David Armstrong, alternatively, has described an even shorter duration: an 'introspective instance'—the term he uses to describe the 'smallest unit of time discernible with respect to inner experience' (1968, p. 105). According to

Armstrong, it may be possible for an episode of pain to be over, before the speaker's sentence declaring 'I am in pain' is even finished. The 2–3 second window of time described by the specious present would thus be too long.

Despite this controversy, I do think that something like the specious present gives us a useful way of distinguishing between our current experiences, and our memories of certain experiences. While not being *exactly* precise, it allows us to say why my utterances such as 'I am now in pain' is current; whereas my memory of being in pain last week is a *current memory* of a previously experienced mental state

#### **1.5.4 Alternative Conditions**

According to the theory-neutral account of introspection: introspection is (1) the process by which one gains knowledge of one's own, and only one's own, psychological mental states; (2) a process *different* in kind from alternative ways of acquiring knowledge about the world; whereby one can only (3) gain knowledge of one's *current* mental states. Before moving on, I will consider three of the other conditions that Schwitzgebel mentions above with respect to introspection—conditions that I do not think need to be considered part of the theory-neutral account.<sup>28</sup> These are best referred to as the directness/immediacy condition; the effort condition; and the detection condition.

The directness/immediacy condition characterises introspective self-knowledge as distinct because it is acquired in a *direct* or *immediate* way. Georges Rey, for example, uses the term 'introspection' for the 'distinctive, comparatively more "direct" way we seem to have knowledge of our own states'

---

<sup>28</sup> Schwitzgebel does not necessarily accept these alternative conditions.

(2013, p. 260). Moran, conversely, seeks to account for the ‘immediacy’ of introspective self-knowledge in wholly negative terms, in the sense that it is ‘awareness that is not inferred from anything more basic’ (2001, p. 11). There is certainly some truth to what Rey and Moran say here, but I think there are some problems with attempting to build such requirements into the theory-neutral account of introspection.

Firstly, if ‘directness’ is understood in the sense that introspective self-knowledge is easier, or quicker, to achieve than non-introspective knowledge—as Rey’s description seems to suggest—then introspective self-knowledge will not always be more direct. For example, if someone asks me ‘Do you believe that The Spanish Civil War began in 1936?’ I will typically need to pause for a few moments before I can give my answer, even though I have used introspection. This process appears to be less direct, in terms of time and effort, than the process by which I come to believe that someone else is in pain, after seeing them wince as they grasp their broken leg.<sup>29</sup>

Secondly, if ‘immediacy’ is understood in the sense that Moran describes—when he says that introspection does not involve an inference from anything more basic—then there will also be difficulties. For instance, some philosophers maintain that we *should* characterise the knowledge of the minds of others as immediate too. Dan Zahavi (2011), for example, notes that while it is common to think of other people’s minds as ‘invisible’ to us, and that our knowledge of them involves complex inferences, this view is not universally accepted. Zahavi describes a view held by some philosophers which

---

<sup>29</sup> Someone might object that the comparison should be between introspective versus non-introspective knowledge of the *same* type of mental state, as opposed to the comparison I’ve made here. Given that I am attempting to articulate features that are common to all forms of introspection, I do not take this to be a serious concern

acknowledges a ‘more immediate experiential access to the minds of others which is prior to and more fundamental than any imaginative projection or theoretical inference’ (2011, p. 546). If this view is the right way to think about the process by which one achieves knowledge of the minds of others, then introspection is not unique in this sense of ‘immediacy’.

There are, of course, other ways of interpreting what Moran might mean by ‘immediacy’ here. One is to say that introspection is *immediate* in the sense that it can only give one knowledge about one’s own mind. If this is right, then we can say that even if the view that Zahavi describes is correct—that there is a certain sense in which our knowledge of the minds of others is ‘immediate’—it will not be literally true that one can *immediately* experience another’s mind. The sense of immediacy that Zahavi describes may be analogous to the kind that an experienced scientist has, when she ‘sees’ the presence of a charged particle by observing a trail in a cloud chamber. The scientist may describe her experience of the charged particle as ‘immediate’, and in some sense it may be, but we realise that she is making an inference, however quick and seamless the inference is.<sup>30</sup> If this is what Moran means by immediacy, then I do not disagree with him. However, it would be superfluous to include the immediacy condition into the theory-neutral account, as this sense of immediacy has already been captured in the first-person condition.

The second of Schwitzgebel’s conditions, that I wish to exclude, is what may be referred to as the *effort condition*. This condition pertains to the level of effort that is required, on the part of the subject, to gain introspective knowledge. One problem with attempting to include this condition in the theory-neutral

---

<sup>30</sup> I borrow this example from Bilgrami (2006, p. 11).

account is that various philosophers have offered contradictory accounts of what could be meant by ‘effort’ in the literature. Bar-On, for instance, notes of the ‘seemingly “effortless”’ (2004, p. 27) way in which we self-ascribe our own psychology; whereas Schwitzgebel describes an alternative interpretation of effort as *not* being ‘*constant, effortless, and automatic*’ (2014).<sup>31</sup> What Schwitzgebel means by this description is that introspection is a deliberate task, one that is undertaken in our more reflective moments. It is not sometimes that we are constantly doing throughout the day.

Given these two contrary ways of construing the effort condition, it seems more appropriate to think of the effort condition as a variable *feature* of a theory of introspection, rather than a condition required at the theory-neutral level. Since the effort required to have introspective self-knowledge of a mental state may differ from theory to theory, the effort condition should not be a condition that is specified at the theory-neutral level.

The third condition I wish to exclude is best referred to as the *detection condition*. This is the requirement that introspection necessarily involves the detection, by some sort of inward glance, of a pre-existing mental state. This is a condition suggested by Brie Gertler who says: ‘inner sense [inner perception] theories construe the...process that results in self-knowledge as an *introspective* process’ (2011a, p. 255). When Gertler contrasts the inner sense view with so-called ‘rival views’, such as the transparency method, she calls them ‘nonintrospective’ (2011a, p. 255). Gertler thinks that the transparency method is a nonintrospective approach to self-knowledge, because it does not involve such

---

<sup>31</sup> As Schwitzgebel points out, an inner sense theorist, such as Lycan (1996), would disagree with this condition because he thinks that introspective monitoring *is* a constant feature of our everyday lives.

detection. Recall that on the transparency method, one's attention needs to be focused upon the content of the mental state rather than the mental state itself.

What reasons are there for thinking that inner perception is essential to introspection? In §1.3, I considered some terminological reasons. There I noted that some philosophers worry that by broadening the concept of introspection to include other views of self-knowledge, such as the transparency method, the etymology of the word 'introspection' would be strained. As I said above, I do not think this is a strong reason to restrict our conception of introspection. I have argued here that what is essential to introspection is the uniquely first-person feature—that is, it is different in kind from the way in which one comes to have other types of knowledge, e.g., the knowledge one has of one's environment, or the knowledge one has of the mind of another. I think we should think of detection as belonging at the theory level of description, not at the theory-neutral level. Internal detection marks out one way which introspection could be, but it is not the only way.

### **1.5.5 Applications**

According to the theory-neutral account of introspection, introspection is: (1) the process by which one gains knowledge of one's own, and only one's own, psychological mental states; (2) a process different in kind from alternative ways of acquiring knowledge about the world; and (3) can only give one knowledge of one's current mental states. Given this set of criteria, we can now ask which of the theories we've considered so far count as introspective.

The inner perception view would certainly count as introspective on this set of criteria. Inner perception would only give one knowledge of one's own

mental states. One would not be able to learn about the mental states of another by *looking inside*. Thus, the process by which one learns about one's own mind would differ from the way in which one learns about the mind of another. Inner perception would also only give one knowledge of one's current mental states.

The transparency method would also count as introspective on this set of criteria. Recall on this view, I can know whether I believe that *P* by judging whether *P* is true. If I gain knowledge of my belief that snow is white by judging that snow is white, for example, I will only be able to learn about my own belief in this way. I cannot learn about another person's belief that snow is white by *judging* that snow is white. Thus, the difference condition is also affirmed. The transparency method will also only give me knowledge about my current beliefs. I will not be able to acquire knowledge of a belief about *P* that I held several months ago, by consciously judging that *P* is true.

The self-shaping view would also count as introspective on this set of criteria. Recall that on this view, one's awareness of one's belief, for example, can be gained by cultivating certain dispositions or commitments that pertain to certain propositions. On this view, I can learn that I believe that Central Park is located in Manhattan, for example, by committing myself to the proposition 'Central Park is located in Manhattan'. This is different from the way in which I would come to have knowledge of another's belief that Central Park is located in Manhattan. I cannot know that another person believes this by committing myself to a proposition. For that, I will need to observe another's behaviour or listen to them speak. Committing myself to a proposition will also only give me knowledge of my current belief.

Neo-expressivism would also count as introspective according to my theory-neutral criteria. Suppose I ascribe the following feeling to myself: ‘I am feeling tired’. On Bar-On’s (2009) view, I would do so by giving articulate vent, in speech or in thought, to my tiredness; as opposed to simply detecting an inner feeling of tiredness by a kind of internal scanner. If this view were true, such a process would only give one knowledge of one’s own mind. One couldn’t gain knowledge of another’s mind by expressing anything. Thus, the difference condition is affirmed. And one’s expressions would only give one knowledge of one’s current mental states.

Ryle’s and Carruthers’ views will not count as introspective on my set of criteria because they deny the difference condition. On Carruthers’ view, in order to know that I have an intention, for instance, I would have to observe my own behaviour, listen to my own speech, or make a theoretical inference—the same sorts of ways I would use to learn about the mental states of others. Similarly, Ryle (1949) denies the claim that there is any significant asymmetry between the way that one knows about one’s own mental states, and the way that one knows about the mental states of another.

All of this can be summarised in the accompanying table (see table 1 below). Although the views listed in this table are by no means exhaustive—I have only included a small sample of views that have been discussed in the recent literature—it does illustrate how one would go about classifying a process as introspective on my view.

**Table 1:**

| <b>Theory of Self-Knowledge</b> | <b>The First-Person Condition</b> | <b>The Difference Condition</b> | <b>The Occurrent Condition</b> |
|---------------------------------|-----------------------------------|---------------------------------|--------------------------------|
| The inner sense view            | ✓                                 | ✓                               | ✓                              |
| The transparency method         | ✓                                 | ✓                               | ✓                              |
| The self-shaping view           | ✓                                 | ✓                               | ✓                              |
| Neo-expressivism                | ✓                                 | ✓                               | ✓                              |
| Inferentialism <sup>32</sup>    | ✗                                 | ✗                               | ✓                              |

According to the theory-neutral account of introspection, a process needs to meet all three conditions in order to count as introspective. Theories which posit processes that do not meet all three conditions will not be considered introspective.

### **1.6 The Targets of Introspection**

Although the main purpose of this chapter is to offer a theory-neutral account of introspection, it will be useful to say a few things about what types of self-knowledge can be introspected. It is uncontroversial that one cannot introspect one's height or age, but what about one's character traits or moods? Here things are less clear. The criteria that I will employ to judge whether a type of self-knowledge can be known by introspection, is whether that type of self-knowledge is capable of being known in the ways described by the theory-

---

<sup>32</sup> Inferentialism, as I understand the view, involves the commitment to the thesis that there is no significant asymmetry between the way in which we know our own minds, and the way in which we know the minds of others. There are, of course, various ways in which one could be an inferentialist, and different ways to interpret the scope of the view. Peter Carruthers (2011), for instance, is an inferentialist when it comes to most of one's propositional attitudes, but not when it comes to one's sensations. Ryle (1949), on the other hands, goes further than Carruthers with respect to the question of scope.

neutral account of introspection. If it can, then it is possible that one can introspect it.

I will now examine seven different types of self-knowledge, and consider how plausible it is that each can be known in the ways described by the theory-neutral account of introspection. Although this list is not exhaustive, it does feature the main types of self-knowledge that have featured in recent philosophical discussions about self-knowledge.

(1) Propositional attitudes: attitudes that subjects have towards propositions.

For example: one's belief that Perth is a city in Australia; one's wish that it will not rain during one's stay in Canberra. Other propositional attitudes include hoping, desiring, intending, fearing, and so on. These mental states are also known as 'intentional states'.

(2) Sensory experiences: the awareness of visual, auditory, touch, smell, and taste experience.

(3) Rationalisations: the reasons why a subject performs a certain action—in the sense that the action is rationally intelligible. For example: Michael may turn on his television *because* he desires to watch his favourite soccer team play. Rationalisations, sometimes called *primary reasons*, are typically understood as a combination of a belief and desire.<sup>33</sup>

(4) Character traits: including courage, laziness, determination, and so on.

(5) Causes of propositional attitudes: the causal explanation for why one desires, fears, or prefers something. For example: Jane likes reading Russian novels *because* they remind her of her childhood spent in Russia.

---

<sup>33</sup> I follow Donald Davidson ([2006] 1963) with this terminology. See also Jaegwon Kim (2010, p. 106).

(6) Moods: including depression, optimism, and so on.

(7) Emotions: including resentment, despair, shock, guilt, empathy, and so on.<sup>34</sup>

The seven candidates for introspective self-knowledge listed here, while not exhausting the different kinds of self-knowledge one can have, provide us with a good starting point for our discussion. I will now briefly examine each type using the criteria for introspection I have given above, in order to see how likely it is that each type of self-knowledge can be introspected.

To my mind, we can begin by dismissing (4)—character traits—from our discussion. This type of self-knowledge seems to be the least likely candidate for knowledge by introspection. Firstly, character traits are not paradigmatic examples of mental states. They tend to be longstanding, and are constituted by different dispositions and preferences. To be considered courageous, for example, one needs to do more than simply assert ‘I am courageous’ or have a momentary feeling of braveness. One would have to, for example, have stood up in the past for what they believed to be right; not have turned away from fearful circumstances, and so on. There does not seem to be any uniquely first-personal way of coming to have knowledge of such facts.

All of this is not to suggest that one is never in a good position to have knowledge of one’s character traits. People who have the right training, or are especially perceptive, might be well-placed to correctly self-ascribe their character traits. And, moreover, one’s introspective self-knowledge that one is feeling lethargic, may count as *partial* evidence that one is a lazy person. If we

---

<sup>34</sup> For a discussion about the difficulties of attempting to classify the different varieties of emotions, see Antonio Damasio (2010, pp. 122–123).

are interested in only the source of such self-knowledge, however, it is hard to see how it could be introspective. A single episode of feeling lethargic, for instance, is not a character trait. There just does not seem to be any uniquely first-person way of knowing what our character traits are.<sup>35</sup>

We can also rule out (5)—the causes of our propositional attitudes—from our discussion. This is because causes are not mental states. There is no uniquely first-person way in which one can come to know the cause of why one believes, intends, or desires what it is one does. Again, I am not claiming that one cannot know what causes one's own desires or beliefs to arise, rather that the source of such self-knowledge should not be thought of as introspective.<sup>36</sup>

The types of self-knowledge that I do think are the most plausible candidates to qualify as self-knowledge by introspection are (1) propositional attitudes and (2) sensory states. These types of self-knowledge appear, at first pass, to meet all the conditions for introspection I gave above. It seems that I can know that I am currently in pain in a uniquely first-person way—in a way that is different from the way others know that I am in pain. The same seems true for some of my beliefs, intentions, and desires. It seems that I can sometimes know by introspection what I believe or what I desire.

Given that the knowledge of (3) rationalisations, (6) moods, and (7) emotions, appears often to be comprised, or in some cases constituted by, certain propositional attitudes and sensations, the question of whether one can introspect (3), (6), and (7), will turn on the question of whether one can introspect (1) or (2). To say that emotions and moods are comprised of certain sensations and

---

<sup>35</sup> For scepticism about the existence of character traits, see Gilbert Harman (2009).

<sup>36</sup> See Timothy D. Wilson (2002, pp. 103–104) for a discussion about how recent data from experimental psychology shows that we are often ignorant about the causes of our various attitudes. See David Hume ([1739–40] 2000) for a classic discussion on the difficulties of being able to observe causation in general.

attitudes, is simply to recognise that some emotions and moods are, partially, or wholly, constituted by them. As Peter Hacker has pointed out, ‘[m]any occurrent emotions, such as anger, fear and excitement are bound up with distinct sensations’ (2013, p. 264). Consider the emotion anxiety, for example. If I am currently feeling anxious, my heart may beat faster, my body temperature may rise, and I may begin to feel tense. These distinct experiences are paradigmatic examples of sensations. Similarly, if I am currently feeling nervous, I may experience ‘butterflies in the stomach’—which is also a paradigmatic example of a sensory experience.<sup>37</sup> The same is true with respect to my moods. If I am feeling depressed, I may experience a feeling of lethargy, or have various occurrent beliefs about my lack of self-worth.

If some emotions and moods are bound up in, or constituted by, various sensations and propositional attitudes, should we simply classify them as sensations or some other attitude? As Wittgenstein asks,

[i]s it hair-splitting to say:—joy, enjoyment, delight, are not sensations?—Let us at least ask ourselves: How much analogy is there between delight and what we call e.g. “sensation”? ([1967] 1981, §484, p. 86e)

Wittgenstein is here asking about classification. If some emotions bear such a striking resemblance to various sensations, should we simply classify them as sensations? I will not attempt to answer this question here. But what the question draws out, in addition to our current discussion, is that emotions, moods, and rationalisations, seem to be comprised of—at least to some extent—various

---

<sup>37</sup> As Ronald de Sousa has pointed out, most emotions also involve the propositional attitude *belief*. He says ‘[i]nsofar as most emotions involve belief, they inherit the susceptibility of the latter to self-deception’ (2013).

sensations and propositional attitudes.<sup>38</sup> Thus, the question ‘Can we introspect such mental states?’ will turn on the question ‘Can we introspect our sensations or propositional attitudes?’

Instead of trying to answer the difficult question of which categories of mental phenomena can or cannot be introspected, and given that we cannot examine all types of self-knowledge in one work, I will be concerned primarily with how we have self-knowledge of our sensory experiences and propositional attitudes in what follows. These are the least controversial categories—that is, most, but by no means all, philosophers would accept that we can introspect them.

Since the focus of this work will be primarily on these two types of self-knowledge, I will end this chapter with a brief discussion of each. Sensory mental states are experiences that are generally categorised as there being ‘something that it is like’ to have them, as Thomas Nagel (1974, p. 436) famously put it. This is to say that sensory mental states are *characterised* by how they feel to a subject. If I have a headache, or taste a mango, for example, then there is something that it is like to have those experiences. A sensation is constituted by a distinctive qualitative experience.

Propositional attitudes such as belief, intention, desire, hope and so on—conversely—pertain to cognition and thinking, rather than any distinct sensation.<sup>39</sup> My belief that water is wet, for example, is not characterised by any distinctive phenomenology. It arises out of my ability to think about the question ‘Is water wet?’ and make a judgement about whether I think that it is. To say that

---

<sup>38</sup> In the case of a rationalisation (or primary reason) it seems to be wholly constituted. As Kim points out, a primary reason for doing something is a ‘coordinated pair of a desire and a belief’ (2010, p. 106).

<sup>39</sup> Though there will be some places where sensations and propositional attitudes intercept. Such as when one has a belief about a sensation—for example, when I believe that I am in pain, or when I believe that I am currently observing a mountain range.

propositional attitudes are *not* characterised by *what it is like* to have them, does not mean that I am claiming they lack *any distinct phenomenology*. In the last few years, a debate has emerged over this point, between what Tim Bayne and Michael Montague call ‘conservatives’ and ‘liberals’ (2011, pp. 2–3). According to Bayne and Montague, conservatives (see, e.g., Tye 1995; Braddon-Mitchell and Jackson 2007) are philosophers who think conscious thoughts, such as my belief that water is wet, lack a distinct ‘phenomenal character’ or ‘what it is likeness’. Whereas liberals (see, e.g., Strawson 1994; Pitt 2004) are philosophers who think that intentional mental states do have a distinct ‘phenomenal character’ or ‘what it is likeness’. A third position, which Bayne and Montague do not mention, can be thought of as a hybrid-view (see, e.g., Gallois 1996, p. 129). This is the view that while some propositional attitudes, such as belief and intention, lack a distinct phenomenology; other propositional attitudes such as some desires do have a distinct phenomenology—for example, when one has a strong desire for a drink.

In what follows, I will not take a stance on this issue. What is important about propositional attitudes, for the purposes of this work, is the fact that they are mental states with intentional content that can be introspected. By intentional content, I mean that such mental states are ‘directed at, or about, or of, objects and states of affairs in the world’ (Searle 1983, p. 1). My occurrent belief that the Beatles were formed in Liverpool has intentional content because it is directed towards the 1960s rock group: ‘the Beatles’. Although such mental states are often referred to in the literature as both ‘propositional attitudes’ or ‘intentional states’ we should distinguish between these terms, because although it’s true that some intentional mental states have propositional content, they need not

necessarily. For example, while my *desire* to travel to the Giza pyramids is *about* the Giza pyramids—and thus has intentional content—it is not, strictly speaking, about a proposition. Similarly, there is intentional content in Bill’s disliking of Thomas—the disliking is directed at Thomas—but it is not about a proposition.<sup>40</sup>

I should also state that this terminology is not without controversy. John Searle, for example, has argued that the term ‘propositional attitude’ is almost ‘invariably disastrous’ (2015, p. 39). He says that because most beliefs (to take an example of an attitude) are not about propositions—they are about people, objects, states of affairs, and so on—it gives the wrong account of intentionality. Since the aim here is not to give an account of intentionality, I will put these concerns to the side. As the term ‘propositional attitudes’ is commonly used in the literature on self-knowledge, I will adopt its use here.

## 1.7 Conclusion

In this chapter, I have offered a theory-neutral account of introspection. According to this account: introspection is the process by which one gains knowledge of one’s own, and only one’s own, psychological mental states. It is a process *different* from other ways of knowing about the world, and can only give one knowledge of one’s *current* mental states. In the last part of the chapter, I turned to the question of what types of self-knowledge could be known by introspection. I said that sensory mental states and propositional attitudes are the most plausible candidates from the list provided.

In the next chapter, I discuss how my claim about introspection is relevant to first-person authority. There I will argue that introspection can

---

<sup>40</sup> On John Searle’s (1983) account of intentionality: *M* is an intentional state—such as a belief or intention—just in case there is an answer to the question ‘Is there something that *M* is about?’ Undirected anxiety, or depression would not count as *intentional* on Searle’s account.

explain first-person authority. If I am right that we can introspect our sensory mental states and propositional attitudes, then this means that we have first-person authority to these types of self-knowledge. This discussion will set the stage for my discussion of scepticism about our ability to introspect our propositional attitudes, which will occur in chapters 3 and 4.



## Chapter 2

### First-Person Authority

You can't have infallibility about your own consciousness. Period. But you can get close — close enough to explain why it seems so powerfully as if you do. Daniel Dennett (2002, p. 13)

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. Tyler Burge (1988, p. 649).

When we attend to our own beliefs, intentions, desires, and thoughts, we typically feel like we are in an authoritative position to correctly self-ascribe them, compared to someone from the third-person point of view. One way of accounting for the asymmetry between first- and third-person psychological ascriptions is to say that we each have a special way of accessing such self-knowledge: in the first-person case, we can use *introspection* to achieve knowledge of these mental states, and in the third-person case, such knowledge must be discerned *interpretatively*—typically by listening to others articulate their own thoughts, or by observing their behaviour.<sup>1</sup>

Even if we accept that this account is cogent, why should it be that subjects who use introspection are in an authoritative position, or more likely to achieve knowledge of their own mental states, compared to subjects who do not use introspection? In this chapter, I argue that first-person authority arises as a result of our propensity to justify our beliefs about what mental state we are in, by introspecting that we are in that mental state. If what I said about introspection in the last chapter is right, then this is a process that can only occur

---

<sup>1</sup> In this chapter, the term 'introspection' will be employed in the theory-neutral sense that was outlined in the last chapter. Where possible, I will avoid reference to any specific theory of introspection.

from the first-person perspective.<sup>2</sup> This account of first-person authority can be expressed, quite generally, as follows:

Introspective justification: *S* has introspective justification for believing that she is in mental state *M*, if and only if *S* has justified her belief that she is in mental state *M*, by introspecting that she is in mental state *M*.

I will argue that introspective justification can explain why some first-person ascriptions are authoritative, in the sense that they are more likely to amount to knowledge, compared to the ascriptions that others make about our own minds. First-person authority, I will argue, does not entail the strong psychological thesis that one's self-ascriptions can never be mistaken—a position known as *infallibility*. Nor does first-person authority entail the epistemic thesis that one's self-ascriptions can never be corrected by anyone else—a position known as *incorrigibility*. I argue that these two views, which continue to be discussed in the self-knowledge literature, are implausible as general accounts of first-person authority. A subject can be said to have first-person authority, in general, and still sometimes be wrong about what she believes about her own mind.

I will proceed as follows. I begin, in §2.1, by considering some terminological issues pertaining to the phrases 'first-person authority' and 'privileged access'. In §2.2, I provide a brief discussion about the different varieties of first-person authority that exist in the literature. In §2.3, I criticise three of these accounts: infallibility, self-intimation, and incorrigibility. In §2.4, I

---

<sup>2</sup> Here I draw upon the work of several recent philosophers including André Gallois (1996), Ram Neta (2011) and Declan Smithies (2012). I agree with these philosophers that justification is the best way to construe first-person authority. The extent to which our views converge and diverge will be addressed in §2.4. I do not claim that introspection guarantees self-knowledge

explicate and defend the introspective justification account of first-person authority. In §2.5, I consider an objection to my strategy of explaining first-person authority with introspection.

## **2.1 First-Person Authority—Terminological Issues**

First-person authority, sometimes also referred to as privileged access, has been construed in various ways in the self-knowledge literature.<sup>3</sup> Some authors, for instance, have construed it in a very demanding sense. Michael Tye, for example, characterises privileged access as the thesis that error, with respect to first-person phenomenal avowals, is ‘impossible’ (2009, p. 184). Tye goes on to reject such a thesis when he says, ‘[if] the position I am advancing requires a rejection of privileged access, so much the worse for that position’ (2009, p. 188).

Not all philosophers have construed privileged access in such a way, however. William Alston (1971) and Ram Neta (2011), have also used the term ‘privileged access’ to describe the epistemic advantage that subjects have to their own psychology, without committing themselves to a thesis as strong as the one that Tye describes.

A similar treatment is given with respect to the term ‘first-person authority’—which is used by philosophers such as Donald Davidson (1984) and A. Minh Nguyen (2004). This term is used much like privileged access to describe the epistemic advantage that subjects have to their own psychology.<sup>4</sup> And like privileged access, first-person authority has been construed in various ways, and has also been said not to exist. Nguyen, for example, claims ‘[f]irst-

---

<sup>3</sup> These terms are typically used interchangeably in the literature.

<sup>4</sup> Davidson (1984) argues that first-person authority arises because we know, in our own case, what we mean by our thoughts or avowals. Davidson argues that such a presumption cannot be made with respect to another’s avowal. In other words, one cannot guarantee that one has correctly interpreted another speaker’s avowal.

person authority does not exist. The problem of explaining it should be dissolved' (2004, p. 472). Like Tye, however, Nguyen offers a controversial definition of what is meant by first-person authority.

First-person authority raises a philosophical issue. In general, the mere fact that a property is capable of being instantiated by a subject does not confer a presumption of truth on his sincere claim that he instantiates it. In general, the mere fact in question does not entitle us to assume in advance that the subject's claim is true. It is not necessary that if a person sincerely claims that his weight, holiday address, or retirement account number is thus and so, then there is a legitimate presumption in favor of the claim. Why is the situation any different when the properties are mental? (2005, p. 458).

What Nguyen is suggesting here is that: if a subject has first-person authority about their own self-ascriptions, then we should be able to assume, *prima facie*, that what that subject says about her own psychology is true. Like Tye's account, however, it is worth asking whether this is the best way to construe first-person authority. It may be that Nguyen's account sets the bar implausibly high. One might agree with Tye or Nguyen that their characterisations of first-person authority do not exist, but disagree that they are the best ways to characterise first-person authority.<sup>5</sup>

Despite the differences in meaning, the terms 'privileged' access' and 'first-person authority' are often used interchangeably to describe the putative

---

<sup>5</sup> In the end, such disputes may simply be verbal disputes. A verbal dispute can be said to occur when, as David Chalmers says, 'two parties agree on the relevant facts about a domain of concern and just disagree about the language used to describe that domain. In such a case, one has the sense that the two parties are "not really disagreeing": that is, they are not really disagreeing about the domain of concern and are only disagreeing over linguistic matters' (2011, p. 515). With respect to the specific domain of concern that we are discussing—first-person authority and privileged access—we can imagine *S* arguing with *Q* over whether first-person authority exists. We can also imagine that by 'first-person authority' *S* and *Q* mean different things. For example, *S* may claim that first-person authority entails infallibility; whereas *Q* does not. *S* and *Q* may both agree, however, that infallibility does not exist. Here the debate between *S* and *Q* would be verbal, as they agree about the facts—namely, that infallibility does not exist—but differ in what words they are using to describe the debate. I will show, as we proceed, claims such as 'first-person authority does not exist', will need to be fully qualified since there are multiple ways that one may construe first-person authority and privileged access.

advantage that subjects have to their own mental states.<sup>6</sup> Sven Bernecker is typical of this approach when he says that he is seeking to account for the epistemic advantage subjects have to their own psychology in terms of ‘first-person authority *or* privileged access’ (2011, p. 33, my emphasis).<sup>7</sup>

Not all philosophers follow this approach, however. Peter Carruthers, for example, uses both terms distinctively. He uses the term ‘privileged access’ to describe the thesis that ‘[o]ne knows...[one’s] mental states in a way others can’t (2011, p. 14). And he uses the term ‘authoritative knowledge’ (instead of first-person authority) to describe the thesis that ‘[t]his knowledge is much more reliable than knowledge of the mental states of others, and cannot normally be challenged from a third-person perspective’ (2011, p. 14). While I think that Carruthers is right to characterise these two terms distinctly, for reasons I will go into below, I am not in complete agreement with the way that he does so. Although his description of privileged access captures something that I think is important about authoritative self-knowledge—namely, the fact that there is a certain asymmetry between the way we know our own minds and the minds of others—his authority thesis is too closely aligned with incorrigibility which, as I will show in §2.4, is problematic.

Given that there are these various accounts in the literature, it is important that we are careful to clarify what we mean by our terms. In what follows, I will use the term ‘first-person authority’ to describe the nature of the epistemic advantage that subjects have to their *own psychology*. This is similar to how the above authors proceed. But rather than treat ‘privileged access’ as synonymous

---

<sup>6</sup> A third term, ‘first-person privilege’, is used by Dorit Bar-On (2004, p. 123) to describe the epistemic advantage that each of us have to our own psychology.

<sup>7</sup> Unlike Tye, Bernecker does not think that such self-knowledge is infallible. He uses the terms ‘privileged’, ‘a priori’, ‘non-empirical’ (2011, p. 33) all synonymously.

with ‘first-person authority’, I will assign a distinct sense to it. Following André Gallois (1996), I think that the term ‘privileged access’ is best used to describe the situation that a subject is in when that subject is most justified, or has better reasons than anyone else, in ascribing a mental state to someone—whether that be with respect to her own mind, or the mind of another. Notice that this thesis makes no mention of introspection, or even the first-person perspective. In order to elaborate upon this thesis, and show why it is important, let us look at an example.

Consider what typically occurs in psychoanalysis. During a session of psychotherapy, a psychoanalyst may discover a fact about her patient’s psychology that the patient herself is completely unaware of. Moran (2001, p. 85) gives an example of an analysand who is unaware that she feels resentment towards a parent for abandoning her. In this example, I think we should say that the psychoanalyst has *privileged access* to her patient’s psychology. The psychoanalyst’s expertise puts her in a privileged position, with respect to achieving knowledge about her patient’s psychology. It does not follow from this admission, however, that the psychoanalyst has *first-person authority* with respect to her patient’s psychology. As I said above, first-person authority is a thesis about the nature of the advantage that subjects have about their own minds. Since the psychoanalyst cannot justify her beliefs about her patient’s psychology by introspection, she does not have first-person authority about her patient’s mind. She must justify her belief that her patient feels resentment towards a parent on the basis of her patient’s behaviour, and what her patient says or does not say. Nevertheless, she has privileged access about *this particular ascription* about her patient’s mind. She has better justification—in

the sense that she has better reasons—for thinking that her patient feels resentment towards a parent.

It is, therefore, always a contingent matter whether someone has privileged access. In the case above, the psychoanalyst has privileged access to her patient's psychology. But with respect to another fact about the patient's psychology, the situation will be different. Suppose the psychoanalyst asks her patient 'How are you feeling today?' when she greets her patient. We can imagine the patient replying 'Awful, I have a terrible headache'. If we suppose that the patient can justify her belief that she has a headache, because she can introspect her headache, then I think the patient has first-person authority with respect to this self-ascription. This is because she can justify her belief that she has a headache in a way that no one else can. We may also suppose that she has privileged access, with respect to her headache. No one else has a justification as strong as the one she does to believe that she has a headache

This distinction allows us to keep separate two key features of the ascription process that are important not to conflate. First-person authority is a thesis about what makes one's own self-ascriptions authoritative—it is restrictive, in the sense that it applies only to the first-person perspective. Privileged access is more encompassing. While one can be said to have privileged access to one's own mind, one can also be said to have privileged access to the mind of another, from the third-person perspective—as our example with the psychoanalyst showed. Since I think that incorrigibility—the thesis that we cannot be shown to be wrong about our own psychology from the third-person point of view—is false, it is important to have a way to talk about such a situation. It is important, as Gallois points out, because '[h]ostility to first-person

authority occasionally arises from conflating first-person authority with non-contingent privileged access' (1996, p. 31). Gallois' point is that some philosophers have been sceptical that we have any special access to our own minds, because we are sometimes not in best position to correctly self-ascribe our own mental states. This, as I will show, is a mistake.

## **2.2 Varieties of First-Person Authority**

As our current discussion has attested, there is more than one way that first-person authority has been construed in the literature. To get a sense of the number of different accounts that exist, it is worth mentioning William Alston's seminal paper, 'Varieties of Privileged Access' (1971). In this paper, Alston lists 34 different versions of how self-knowledge has been characterised by various philosophers as 'privileged'—Alston's term for what we are attempting to explain as first-person authority.<sup>8</sup> Alston's list includes accounts from Descartes, Locke, and Hume, as well accounts from more recent philosophers such as A.J. Ayer and Sydney Shoemaker.<sup>9</sup> Alston notes that with 'sufficient ingenuity' (1971, p. 240) his list of 34 could be further expanded.

While Alston provides us with a valuable taxonomy of the various accounts of first-person authority that have been explored by various philosophers, his most important point, in my view, is one that he makes about scepticism. Alston concludes his paper by saying that since attacks on first-person authority often fail to 'take account of the full range of possibilities' (1971, p. 240), they often miss the mark, because they are directed at one, or some, of the 'possible versions' (1971, p. 240) of first-person authority that exist.

---

<sup>8</sup> I hasten to add that Alston does not think that all of these accounts are all true.

<sup>9</sup> See Descartes ([1641] 1984); Locke ([1690] 1975); Hume ([1739–40] 2000); Ayer (1956); and Shoemaker (1963)

This is an important point, and one that remains underappreciated. Recall that in the previous section, I examined Nguyen's and Tye's claim that first-person authority (in Tye's case, privileged access) may not exist. Nguyen went as far to say that we should stop attempting to explain it. While Nguyen may be right that the version of first-person authority that he targets might not exist—namely, the thesis which maintains that self-avowals are always true—it does not follow from this that other ways of characterising first-person authority are false. Thus, even if Nguyen's claim is true, it does not have the far-reaching implications that he thinks it does, since there are other ways of characterising first-person authority. He has failed to take into account the fact that there are multiple ways in which first-person authority may be construed.

Given that there are these multiple ways of characterising first-person authority, I do not claim that the account of first-person authority that I will defend is the only one that can be given. I grant that there may be various ways in which one may be said to have first-person authority with respect to one's self-ascriptions. What I will argue, however, is that the version of first-person authority that I will defend—what I have referred to so far as introspective justification—is an important account. I will argue that: (i) introspective justification does not fall prey to the various counterexamples that beset other accounts of first-person authority; (ii) introspective justification is explanatorily powerful in the sense that it explains why some self-avowals are authoritative in a way that third-person ascriptions are not; and (iii) introspective justification helps to explain first-person authority quite generally. It helps to explain why we have first-person authority about our beliefs and sensations, but not our age or height. Before getting to this account, I will first criticise three accounts from

Alston's list of 34. The three views are known as infallibility, self-intimation, and incorrigibility. These are three of the most well-known accounts of first-person authority. I will examine just these three because they are the ones that are generally dominant, and targeted, in debates about first-person authority. I will ignore the other accounts from Alston's this list of 34 because they are either not as widely discussed, or are very similar to the three views I discuss here.

### **2.3 Infallibility, Self-Intimation, and Incorrigibility**

I will begin with the strongest version of first-person authority: infallibility. This is the thesis that one cannot be mistaken about what one's own psychology is like. Although infallibility is dismissed by Paul Churchland as 'extraordinary' (1988, p. 75), it is worth examining for at least three reasons. Firstly, the thesis has historical associations with Descartes and the early empiricists Locke and Hume; secondly, there are contemporary proponents of the thesis (see, e.g., Jackson 1973; Chalmers 2003; Horgan and Kriegel 2007); and thirdly, as we have seen above, the thesis continues to be targeted in debates about first-person authority.<sup>10</sup> The infallibility thesis can be formulated as follows:

Infallibility: necessarily, if *S* currently believes she is in psychological mental state *M*, then *S* is in mental state *M*.

For example, if *S* currently believes that she has an intention to eat breakfast, or if *S* currently believes she has a toothache, then it logically follows that *S* does intend to eat breakfast or *S* does have a toothache.

---

<sup>10</sup> The infallibility thesis is also sometimes discussed in debates about free will. Greg Caruso (2012, ch. 5), for example, attempts to motivate scepticism about free action because of the fact that we are prone to make mistakes in our thinking about how our own minds work.

The second account that I will examine is the self-intimation thesis. This is the thesis that it is not possible for *M* to obtain—e.g., that *S* has an intention to eat breakfast, or *S* has a toothache—without *S* believing, or being aware of, *M*.<sup>11</sup>

This thesis can be formulated as follows.

Self-intimation: necessarily, if *M* obtains, then *S* believes, or is aware that, *M* obtains. It is impossible for *M* to be the true, without *S* believing, or being aware of, *M*. It would be impossible for *S* to be in pain, and for *S* not to believe, or be aware, that she was.

The self-intimation thesis is about being *aware* of mental facts. Restrictions about how a subject becomes aware of such facts is not typically included in formulations of the thesis (see Schwitzgebel 2014). If it true that I intend to visit Washington D.C in the summer, then the self-intimation thesis says that I that cannot fail to be aware of this intention.

The third account that I will examine, the incorrigibility thesis, is weaker than the above theses.<sup>12</sup> It is the thesis that while it is possible that *S* could falsely believe that she has an intention to eat breakfast, or have a toothache, it is not possible that someone else, from the third-person perspective, could point this error out. We can formulate this thesis as follows:

---

<sup>11</sup> Other philosophers have considered different ways in which to construe the relation that subjects have to *P* in the self-intimation thesis. David Armstrong (1968, p. 101), for example, thinks one just has to *believe* that *P*. Schwitzgebel (2014) says that one has to *judge* that *P*. I prefer to use the locution *being aware of P* here because I think it contrasts well with the claim that a subject cannot be ignorant about *P*. See Shoemaker (1996, p. 51) for the defence of a ‘weakly self-intimation’ view.

<sup>12</sup> As I mentioned in §2.1, the incorrigibility thesis has featured in recent attempts to define what first-person authority is (see, e.g., Nguyen 2004; Carruthers 2011). Some philosophers (see, e.g., Chalmers 2003) consider incorrigibility and infallibility to be equivalent. However, as our current discussion has shown, there are key differences that are important not to conflate.

Incorrigibility: necessarily, if *S* believes that she is in psychological mental state *M*, it cannot be shown, from the third-person point of view, that she is not in mental state *M*.

I will begin my assessment of these three views by examining a case that I think provides a counterexample to all three. Let us return to the example about the psychoanalyst that was described in section §2.1. Here, we see it is possible for a situation to arise where a subject has falsely attributed a mental state to herself, or is ignorant of a mental state that she has. The subject has resentment towards a parent for abandoning her, but is not aware of having the attitude. In this case, we can imagine that the subject believes that she has no resentment towards her parent. Such an example shows that it is possible for a subject to be wrong about her own psychology—since in this case, the subject falsely believes she has no resentment towards her parent. This shows us that the infallibility thesis cannot be true. Since it is also the case that the subject who feels betrayal or resentment towards a parent is unaware, or does not believe this fact, the self-intimation thesis is also falsified. The incorrigibility thesis is also falsified because the subject has been shown to be mistaken about her own psychology, in this case by her psychoanalyst.

Given that such examples appear ubiquitous and uncontroversial, how might a defender of any of the above formulations respond? One strategy is to restrict the scope of each thesis, so that only beliefs about phenomenal states—such as one’s belief that one is in pain, or one’s belief that one is having a sensation of redness—are considered. This is an approach proposed by Terry

Horgan and Uriah Kriegel, who formulate the following infallibility thesis: ‘[n]ecessarily, for any phenomenal experience E of a subject S and phenomenal property P, if S believes that E is P, then E is P’ (2007, p. 125). Given this adjustment to the original infallibility thesis—and we can imagine similar adjustments being made with respect to the self-intimation and incorrigibility theses—what response can be offered? While it is certainly true that the psychoanalysis counterexample I gave above will not be applicable—since it pertains to a subject’s attitudes and emotions—I still think that such reformulations are susceptible to counterexamples.

These counterexamples include trivial cases such as when a self-pitying person mistakes an itch for a pain<sup>13</sup> to more complex cases such as described in ‘fraternity initiation’ type cases.<sup>14</sup> These are cases where we are to imagine that a subject is tricked into believing that he is in physical pain, when he is actually not. According to a typical fraternity initiation case, we are to imagine that a subject is blindfolded and told, as part of a fraternity initiation, that a razor will cut his neck. While blindfolded, and anticipating the razor’s cut, an ice cube is pressed against the subject’s neck. It is plausible to presume that once the ice cube is pressed against the subject’s neck, he will believe that he is in physical pain. However, since an ice cube being pressed against one’s neck is unlikely to be enough to cause one to be in physical pain, the subject presumably cannot be in pain, and so he will *falsely* believe that he is in physical pain. If this interpretation is correct, then it provides us with a counterexample to the reformulated infallibility, self-intimation, and incorrigibility theses: the subject is

---

<sup>13</sup> This example is from Timothy Williamson (2000, p. 24).

<sup>14</sup> I borrow the term ‘fraternity initiation case’ from Horgan and Kriegel (2007). Variations of this example have also been discussed by several other philosophers (see, e.g., Shoemaker 1996; Chalmers 2003; Schwitzgebel 2012b; Smithies 2013).

mistaken about being in physical pain; he is ignorant about having a cold sensation that he does have; and others, from the third-person point of view, have pointed this out.<sup>15</sup>

Given the existence of such cases, how might a proponent of the phenomenal infallibility thesis (and also the self-intimation thesis and the incorrigibility thesis) respond? One response, given by Horgan and Kriegel—referring specifically to the fraternity initiation case—is to say ‘[e]ven if one is mistaken in how one initially classifies the experience ...one is not mistaken in judging that *it feels like this*’ (2007, p. 130). I find this response problematic. By granting that misclassification is possible, as Horgan and Kriegel do, one must surely also admit that one has formed a false belief about one’s current phenomenal experience. Once the subject realises that it was only an ice cube on his neck, we can suppose that he stops believing that he is in physical pain. If we acknowledge this, I think we should also acknowledge that the subject’s belief about what about mental state he is in has changed even though the sensory feeling that gave rise to his original belief has remained roughly the same. This would seem to suggest that the same type of sensation (or one very similar), only a few seconds apart, can give rise to a belief that one is in physical pain and also to a contrary belief that one is not in pain (when it was experienced).<sup>16</sup> This is enough to show that one can form false beliefs about what sensation one has, even if it is true that one cannot be wrong that one is having *some sensation*. For similar reasons, I think the self-intimation thesis is false—since the subject is having a cold experience and does not believe he is; and it also shows that incorrigibility cannot be true—since the people who placed the ice cube on his

---

<sup>15</sup> Horgan and Kriegel do not think that this is the best way characterise this example (2007, p. 141 note 17). I will assume for the sake of this discussion that it is.

<sup>16</sup> I say roughly the same, because some change has clearly occurred.

neck know that he is not in physical pain. Such strong notions of first-person-authority, therefore, cannot provide us with a general explanation of why our own self-ascriptions are more likely to amount to knowledge, compared to ascriptions others make about our own minds.

I have argued that infallibility, self-intimation, and incorrigibly are not tenable explanations for why one's own self-ascriptions are generally more likely to result in knowledge, compared to the ascriptions another person, from the third-person point of view, would make about another's mind. In claiming this, I am not denying that there is some self-knowledge that may be infallible, self-intimating, or incorrigible. I think that Descartes' ([1641] 1984, p. 17) claim that one cannot be mistaken in thinking that one is a *res cogitans* (a thinking thing) is hard to deny; and I also think that a case can be made for thinking that one cannot be wrong about thinking that one is in *some* mental state—that is, being mistaken about being in a mental state is itself a mental state.<sup>17</sup> My point is that such types of epistemic security are rare.

Infallibility, self-intimation, and incorrigibly do not provide us with good general accounts of why our own self-ascriptions are more likely to amount to knowledge, compared to the ascriptions that others make about our own minds. In the next section, I seek to account for what does.

#### **2.4 First-person Authority: Introspective Justification**

Having argued that certain attempts to characterise first-person authority in terms of infallibility, incorrigibly, or self-intimation are misplaced, we are now finally ready to defend a positive account of first-person authority: introspective

---

<sup>17</sup> Another plausible candidate for a belief that is infallible is given by Gallois, who says '[m]y belief that I have some beliefs is infallible' (1996, p. 22).

justification. This account draws upon recent work from Ram Neta (2011) and Declan Smithies (2012), who themselves draw upon an account of first-person authority that Alston (1971) calls truth-sufficiency—a view that is centrally focused on justification.<sup>18</sup> Although I am largely in agreement with Neta and Smithies that justification is the right way to think about first-person authority, there is some ambiguity in their accounts which, in my view, is the result of an imprecise characterisation of what it is that authoritatively justifies certain self-ascriptions. I think that by appealing to the theory-neutral account of introspection, which was outlined in the last chapter, this ambiguity is abated, and an explanatorily powerful account of first-person-authority emerges.

Let me begin by first describing Neta's view. Neta uses the term 'P-accessible facts' (2011, p. 9) to describe facts that subjects have *privileged access* to (the term Neta uses to describe what I am attempting to explain as first-person authority). What are P-accessible facts? He offers the following two formulations of this view. The first is: '[t]he fact that [*P*] is *P*-accessible to *S* just if and just when: the fact that [*P*] is itself a justification for *S* to believe that [*P*]' (2011, p. 20, my emphasis). An example will help to clarify this conception. Let 'the fact that *P*' refer to the fact that Tom has a headache; and let '*S*' refer to Tom. If Tom's headache is what justifies Tom's belief that he has headache, then Neta thinks that such a fact (that Tom has a headache) is a *P*-accessible fact—and is one that Tom has first-person authority about.

Neta's second formulation of this view draws upon what Alston (1971) calls 'truth sufficiency'. This is described by Neta as follows:

---

<sup>18</sup> As Neta (2011) points out, the truth-sufficiency account has been defended previously, in various ways, by A.J. Ayer (1963) and Roderick Chisholm (1957).

*S* knows, in a privileged way, that [*P*] just if and just when: *S* knows that [*P*], and *S* knows this because *S* believes that [*P*] on the basis of the following justification that she has for believing it: [*P*] (2011, p. 20, my emphasis).

This second formulation says essentially the same thing as the first. If Tom believes that he has headache *on the basis of the fact* that he does have a headache, then he knows that he has a headache in a *privileged way* (again, the term Neta uses to describe what I am calling first-person authority). Neta's view is a combination of these two formulations.

How does Neta's view purport to explain first-person authority? It does so because it is only from the first-person point of view, Neta thinks, that one can justify one's belief that one is in mental state with the mental state itself. To further clarify this response, let us consider how someone from the third-person point of view would come to have knowledge of someone else's mind. Suppose that I also believe that Tom has a headache (and I am not Tom). What justification can I have for believing that Tom has a headache? According to Neta, it cannot be the fact that Tom has a headache, because only Tom can justify his belief that he has a headache with this fact. My justification for believing that Tom has a headache will, typically, come from facts such as: Tom has told me so, Tom is wincing, and so on. Such facts are not Tom's headache. Thus, according to Neta's account, my justification for believing that Tom has a headache is not as authoritative as Tom's justification. Tom can justify his belief that he has a headache with the fact that he has a headache, whereas I must observe his behaviour or listen to his testimony. These things are not Tom's headache.

Is Neta right that this account can provide us with a cogent explanation of first-person authority? I think he is right, but only if we are more explicit about the nature of the justification in question. I think we need to be more explicit about why it is that being in mental state *M*, for example, would give a subject first-person authoritative justification for believing that she is in *M*—in the sense that her belief about being in *M* is more likely to amount to knowledge, compared to the beliefs others have about her being in *M*.<sup>19</sup> Recall that according to Neta’s formulation of truth-sufficiency, one has first-person authority when ‘*S* believes that [*P*] on the basis of the following justification that she has for believing it: [*P*]’ (2011, p. 20, my emphasis). Now while it is true that being in *M* may be *different* from the way that others achieve knowledge of one’s mental state, there is still a question about what makes it authoritative. I think the way to answer to this question is to appeal to introspection, in the theory-neutral sense I offered in Chapter 1. The problem with Neta’s account, however, is that introspection is not mentioned in his formulation of truth-sufficiency; and neither does Neta mention the word ‘introspection’ in his article defending it. Without appealing to introspection, I think there is an ambiguity present in Neta’s account. Let me explain this in more detail.

Recall that according to the theory-neutral account of introspection: if I have the capacity to introspectively know my own mental state *M*, then this process, if introspective, would *not* give me knowledge of any other minds. Introspection can only give me knowledge about my own mind. This helps to explain *why* Tom’s justification for believing that he has a headache is authoritative only if it is introspective. If Tom can have introspective access to

---

<sup>19</sup> I refer to mental states as *M*, as per convention. I refer to propositions as *P*, also as per convention.

the sensation of having a headache, then only he can. No one else can learn about his headache is this way. Since I cannot introspect Tom's feeling of having a headache, I must justify my belief that he has a headache with other phenomena. Since the theory-neutral account of introspection clearly states what is required for a process to count as introspection, we can appeal to introspection as an explanation for first-person authority without begging any questions. On Neta's account, it remains possible that *S* could have first-person authority about being in mental state *M*, just because *S* is in mental state *M*, despite the fact that *S* cannot introspect that she is in *M*. This is something that we should not want to say.

Smithies (2012), unlike Neta, does mention introspection in his account; and, importantly, he mentions the concept of *introspective justification*. He defines this as the 'justification that one has to believe that one is in a certain mental state, which one has just by virtue of being in that mental state' (2012, p. 261).<sup>20</sup> This account is very similar to Neta's conception of truth-sufficiency—in the sense that it focuses on first-person justification. Smithies thinks that his account can explain why self-knowledge is authoritative.

As I said with respect to Neta's view, I think that Smithies' view is basically correct. However, since Smithies does not offer a theory-neutral account of introspection, questions remain about what features a process must possess in order to count as introspective on his view. For example, if Brie Gertler (2011a) is right that only theories that describe the process of achieving self-knowledge in terms of *inner perception* should count as introspection, then some views—e.g., the transparency method—will not count as introspective.

---

<sup>20</sup> Smithies (2012, p. 333) himself cites Neta's view as an inspiration.

This raises the question of whether, on Smithies account, we should still think of these views as being associated with first-person authority. And if not, why? I think that they should if they meet the conditions for introspection on my account, but it is not clear on Smithies' account how or why this is so.

We also do not want all theories of self-knowledge to count as first-person authoritative. Smithies explains his view in terms of the 'justification that one has to believe that one is in a certain mental state, which one has just by virtue of being in that mental state' (2012, p. 261). However, this cannot be sufficient for first-person authority. Suppose that Tom intends to leave the cinema, and suppose that his reason for believing so is that he sees himself getting out of his chair. We should not say that he has introspective justification for believing that he has this intention. This is because the way in which Tom has gained knowledge of his intention, is the same way that another, from the third-person point of view, would gain knowledge of his intention.

We can avoid this concern by appealing to the theory-neutral account of introspection. With respect to the question of what a process must possess to count as introspective, recall what was said in the last chapter—namely, that introspection is the process by which: (1) one gains knowledge of one's own, and only one's own, psychological mental states; (2) it is a process *different* from other ways of knowing about the world; and (3) it can only give one knowledge of one's *current* mental states. Views that are not characterised in terms of inner perception, and yet still meet all these conditions, will still count as introspective on my view. Moreover, any view that describes the process of self-knowledge

without meeting the conditions for introspection that I provided above will not be able to offer introspective justification.<sup>21</sup>

This helps to explain first-person authority because introspective justification can only occur from the first-person point of view. Introspection can only give one knowledge about one's own mind, and *not* anyone else's. So, if Tom is able to learn that he has a headache by using introspection, then no one else can learn about his mind in this same way. This exclusivity, which is stated explicitly in the theory-neutral account of introspection, is the missing element from the above accounts. It is what entitles us to explain first-person authority by appealing to introspection. What better justification could there be? We can now offer a formulation of introspective justification as follows:

Introspective Justification: *S* has introspective justification for believing that she is in mental state *M* if and only if *S* has justified her belief that she is in mental state *M*, by introspecting that she is in mental state *M*.

This account of first-person authority is a general one—general in the sense that it applies to any view of self-knowledge that meets all the conditions for introspection that I gave in the last chapter. This is advantageous because, as I mentioned in chapter one, there is much controversy about what view of self-knowledge is the correct one.

The introspective justification account of first-person authority also helps to explain why not all of one's self-ascriptions will be first-person authoritative. This is explained in the accompanying table (see table 2 below). If one justifies

---

<sup>21</sup> Carruthers (2011), for example, defends the thesis that we cannot introspect most of our propositional attitudes.

one's belief that one has a certain mental state by a non-introspective process, it will not be authoritative in the sense I have been describing.

**Table 2:**

| Psychological self-ascription            | Justification                              | First-person authority? |
|--|--|-------------------------|
| 1. I believe that I intend to see a film | I introspect my intention to see a film    | Yes                     |
| 2. I believe that I intend to see a film | My friend's testimony                      | No                      |
| 3. I believe that I am angry             | I introspect my feeling of anger           | Yes                     |
| 4. I believe that I am angry             | The sight of my reflection in the mirror   | No                      |
| 5. I believe that I intend to see a film | I introspect my desire to eat Indian food. | No                      |

The psychological self-ascriptions (1) and (3) are first-person authoritative, because they would be justified, introspectively, by the mental state that the ascription is about.<sup>22</sup> Self-ascriptions (2) and (4) are not first-person authoritative, because they are justified non-introspectively. This does not mean that (2) and (4) can never be true. What it does mean is that a subject will not have first-person authority with respect to such ascriptions—even though they may have first-person authority in general.<sup>23</sup> If one self-attributes, for example, the intention to leave the cinema on the basis of the fact that one is seeing oneself

<sup>22</sup> Assuming, of course, that I can introspect such mental states.

<sup>23</sup> By having first-person authority in general, I mean that one has the potential to justify a belief about their own mind by introspecting. This does not mean, as the above discussion has attested, that one couldn't find out about their own mind non-introspectively.

get one's coat, then one would have the same justification as another person would have, from the third-person perspective. Finally, even though the psychological self-ascription (5) is justified by an introspective mental state, it is not first-person authoritative, because it is not justified by the mental state that the ascription is about. Not just any introspective justification will do.

Another advantage of the introspective justification account is that it is compatible with errors of the kind that were described above. Recall the fraternity case. In this case, we said it was plausible to suggest that the subject forms the false belief that he is in physical pain after the ice block touches his neck. How can such a case be explained on the introspective justification view? We can say that the subject's belief that he is in pain is not justified by his introspective experience of pain—after all there is no experience of pain to justify the belief. It is more plausible to suppose that he believes that he is in pain because he is both anxious about his current situation and also feels a novel and striking sensation. His belief that he is in pain is not justified by his pain experience, and so is not introspectively justified—even though he may have first-person authority in general.

A second example, provided by Neta, further illustrates this point. Like the fraternity case, it involves a situation where a subject falsely self-ascribes a mental state. Unlike the fraternity case, however, the mistake arises not because the subject has misclassified an experience, but because the subject has justified his belief upon the mistaken testimony of another. In this case, Neta asks us to imagine a person, Herman, who has come to believe that he has a headache.<sup>24</sup> We are to suppose that Herman has formed the belief that he has a headache not

---

<sup>24</sup> See Neta (2011, p. 20).

on the basis of introspection, but rather, by speaking to a noted anaesthesiologist, who is inspecting Herman's brain through a scanning device. Based upon the anaesthesiologist's mistaken diagnosis, Herman forms the belief that he has a headache. This mistake can be explained, on the introspective justification account, by the fact that Herman's justification for believing that he has a headache is the testimony of his anaesthesiologist. This is non-introspective justification. As such, Herman's self-ascription will not be first-person authoritative. To be clear, I am not suggesting that the anaesthesiologist is incapable of forming true beliefs about Herman's mental states. What I am suggesting is that first-person authority only holds with introspection. A subject will have first-person authority, in the sense described, if and only they have justified their belief that they are in a certain mental state by introspecting that they are in that mental state.

In addition to explaining why some of one's own self-ascriptions are first-person authoritative, the introspective justification account of first-person authority can provide us with an answer to the question 'What types of self-knowledge do subjects have first-person authority about?' This question can be answered by attending to another—namely, 'What types of self-knowledge can one introspect?' If one can introspect one's own sensory states, then one can have first-person authority with respect to that type of self-knowledge. If one cannot introspect one's character traits, then one cannot have first-person authority with respect to that type of self-knowledge. This helps to situate the scepticism that I will look at in the next chapter—namely, the position that we cannot introspect most of our propositional attitudes. Finally, the introspective justification account also explains why we don't have any first-person authority

when it comes to our age, height, or weight. This is because we cannot gain knowledge of our age, height or weight by introspection. For in such cases, we will have to take a third-person angle towards ourselves. A subject is only ever contingently in the best position to achieve true beliefs about such facts.<sup>25</sup>

The introspective justification view of first-person authority provides us with an explanation for why it is that we have authoritative access to our mental states; it does not fall prey to various counterexamples; and it can help to inform us about the limits of first-person authority. For these reasons, I think it is a plausible account of first-person authority.

### **2.5 Objection: Introspection as Explanatorily Inadequate?**

The central thesis that I have defended in this chapter is the thesis that introspection, broadly construed, can help to explain why subjects have first-person authority, with respect to some types of self-knowledge. With this account now given, I will consider an objection. This is the objection that citing introspection as an explanation for why some self-ascriptions are authoritative is explanatorily weak, and thus cannot explain first-person authority. This is a point made by Donald Davidson who says the following:

[c]ontemporary philosophers who have discussed first person authority have made little attempt to answer the question why self-ascriptions are privileged. It is long out of fashion to explain self-knowledge on the basis of introspection. And it is easy to see why, since this explanation leads only to the question why we should see any better when we inspect our own minds than when we inspect the minds of others (1984, p. 103).

---

<sup>25</sup> I grant that most of us have what I have called privileged access to our own birthdate, weight, and height. That is, most of us are better placed than anyone else to attain such self-knowledge. But clearly this is not the case universally. Young children who have not learned a language will be ignorant of their birthdate; their age; and their weight. Their parents or relatives will have privileged access to such facts. The point I am trying to make is that there is no uniquely first-person way of coming to know such facts.

Since I have attempted to explain first-person authority on the basis of introspection, Davidson's comments are pertinent to the view I have defended here. Davidson's criticism, as I understand him, is that introspection is not a *mistaken* explanation, but rather, it is an *inadequate* explanation—inadequate in the sense that it does not explain why introspective self-ascriptions are authoritative, or more likely to amount to self-knowledge, compared to the ascriptions that others make about our own minds.

Given what I have said so far about introspection in Chapter 1, I do not think that Davidson's objection is as forceful as he thinks. Here is why. According to Davidson's criticism, citing introspection as an explanation for first-person authority is inadequate because it only raises the further question of *why* we should ascribe any *better* in our own case than when we attempt to ascribe mental states to the minds of others. The thought here is that although introspection may be different—or 'peculiar' as Byrne (2005a, p. 81) puts it—from the way that others gain knowledge of our own psychology, why does this difference help to explain first-person authority? In response to this question, I think that if we appeal to the theory-neutral account of introspection—and in particular the first-person condition—reasons can be offered for thinking that introspection can explain first-person authority.

Recall that according to the first-person condition, a process is introspective only if it reveals knowledge about one's own mind, and *not* anyone else's. Drawing upon our discussion above about introspective justification: if my belief that I desire to travel to Cairo is justified by the fact that I can introspect my desire to travel to Cairo, then I will have unique justification for this self-ascription. If someone else comes to believe that I also desire to travel to Cairo,

they will not be able to justify their belief in the same way that I can. They must justify their belief that I desire to travel to Cairo by observing my behaviour or listening to me speak. As I have argued in this chapter, it is this unique justification that makes introspective justification authoritative. To say that introspection cannot explain first-person authority is to underappreciate this point.

One may respond to what has been said here by insisting that while my explanation may be cogent, it is still inadequate, because I have yet to say anything about *how* reliable introspection is. Since I have been discussing introspection so broadly—in the theory-neutral sense—I have yet to address the question ‘How reliable is introspection? Recall that I commenced this chapter with a quote from Daniel Dennett, who suggests that while we cannot have infallibility about our own minds we can get close—‘close enough to explain why it seems so’ (2002, p. 13). One may, fairly, object that I have not even attempted to answer the question of how close we can get. My response to this concern is to say that such a question is not an appropriate one to ask at the current level of analysis. This is because I am only attempting to explain first-person authority in a general sense, in a sense common to all forms of introspection. I think that such a question should be reserved for accounts of self-knowledge at the theory level. For example, a higher-order theorist of consciousness (see, e.g., Lycan 2004) may wish to account for the reliability of introspection differently from the way that a same-order theorist (see, e.g., Block 1990) would. Since these accounts will yield different answers, to talk of the reliability of introspection at the theory-neutral level is inappropriate, in my

view. This is a positive feature, rather than a weakness, of the theory-neutral account of introspection.

## **2.6 Conclusion**

I have, in this chapter, offered an account of first-person authority. I argued that introspection can explain why we have it. Although I grant that other accounts of first-person authority may be devised, the introspective justification account of first-person authority is advantageous because unlike other accounts—such as infallibility, self-intimation, and incorrigibility—it does not fall prey to counterexamples. It also informs us of the limits of first-person authority. If one cannot introspect an object of self-knowledge (e.g., a belief or desire), then one cannot have first-person authority with respect to that object of self-knowledge. The introspective justification account of first-person authority also captures something important about our common-sense intuitions that give rise to the concept of first-person authority in the first place.

In the next two chapters, I address scepticism, with respect to our ability to introspect our propositional attitudes. Like many, if not most, philosophers I think that we can introspect our propositional attitudes, and thus I think we can have first-person authority with respect to our propositional attitudes. According to some philosophers, however, this is not the case. Gopnik (1993) and Carruthers (2011), for instance, argue that there is no unique first-person way in which we come to have self-knowledge of our propositional attitudes. If they are correct, then we do not have first-person authority—in the sense argued for here—to our propositional attitudes.

## Chapter 3

### Challenging Introspection

Our access to our own thoughts is just as indirect and fallible as our access to the thoughts of other people. We have no privileged access to our own minds.  
Alex Rosenberg (2016)

In this chapter, and the next, I will consider the possibility that we cannot introspect our own propositional attitudes (e.g., our beliefs, desires, wishes, hopes, intentions, fears, and so on).<sup>1</sup> Such a position is a radical because it involves the rejection of the thesis that there is any significant asymmetry between the way in which we acquire knowledge of our own minds, and the way in which we acquire knowledge of the minds of others. If such a view were true, we would not, on my account, have first-person authority with respect to our own propositional attitude self-ascriptions.<sup>2</sup> We can see this by recalling the introspective justification view of first-person authority that was defended in chapter 2:

*Introspective Justification:* *S* has introspective justification for believing that she is in mental state *M*, if and only if *S* has justified her belief that she is in mental state *M*, by introspecting that she is in mental state *M*.

As we can see from this above formulation, without introspection there can be no introspective justification, and thus, there can be no first-person authority with respect to our propositional attitudes self-ascriptions.

---

<sup>1</sup> To claim that one does not have introspective access to a certain type of self-knowledge is to reject the claim that one can *sometimes* introspect that type of self-knowledge. By the term ‘introspection’, I continue to use it in the theory-neutral sense that was outlined in chapter 1.

<sup>2</sup> I am only concerned with the question of whether we can introspect our propositional attitudes in this chapter. The question: ‘Can we introspect other types of self-knowledge, such as our emotions and moods?’ is not one that I will address here.

In what follows, I will examine two different ways in which introspective justification may be brought into doubt. The first comes from a view called *content externalism*. This is the view that the knowledge of one's thoughts cannot wholly be justified by introspection alone. On this position, the knowledge of one's thoughts will also need to be justified, or determined, in part by one's environment. This view is relevant to our current discussion of first-person authority because if our self-ascriptions about our own propositional attitudes cannot be fully justified by introspection, then, according to the account of first-person authority that I have defended, we would not have complete first-person authority, with respect to our propositional attitudes.<sup>3</sup>

The second challenge to introspection that I will examine stems from a view that is best referred to as *inferentialism*. This view is far more radical than content externalism because it violates, in a far more explicit way, the difference condition and the first-person condition—two conditions that I claimed were required for introspection. On this view, there is no difference in kind between the way in which one would acquire knowledge of one's own mind, and the way in which one would have knowledge of the mind of another. According to Rosenberg, a philosopher who endorses this view, 'our access to our own thoughts is just as indirect and fallible as our access to the thoughts of other people' (2016).

---

<sup>3</sup> In this chapter, I understand first-person authority in binary terms—that is, either you have it or you don't. If introspection alone is incapable of giving one knowledge of a mental state then they will not have first-person authority in the sense I have described.

In response to these two different views, I will argue (i) there is no conflict between content externalism and first-person authority; and (ii) the arguments in favour of inferentialism are untenable.<sup>4</sup> The chapter will be structured in the following way. In §3.1, I explicate content externalism. First, I offer a brief explanation of the view, and then show why there is no conflict between it and introspection. In §3.2, I give a brief overview of inferentialism. In §3.3, I discuss a version of inferentialism—namely, Peter Carruthers’ Interpretive Sensory-Access (ISA) theory of self-knowledge.<sup>5</sup> In §3.4, I present two objections to Carruthers’ ISA theory. In §3.5, I introduce what I will call the confabulation arguments for the ISA theory. In §3.6, I criticise one of the confabulation arguments: the ontological parsimony argument. This will set the stage for my analysis of the second confabulation argument, the patterning argument, which will be the topic of chapter 4.

### 3.1 The Content Externalism Challenge

The first challenge to introspection that I shall discuss comes from the philosophical doctrine of *content externalism*.<sup>6</sup> According to this theory, the content of one’s thoughts is not wholly determined ‘in the head’, but rather is constituted, in part, from facts in one’s environment—facts which lie ‘outside’ the head. The locution ‘in the head’ is from a quote by Hilary Putnam, which, in

---

<sup>4</sup> A third view, *eliminative materialism* (see, e.g., P.M. Churchland 1981; Stich 1983; P.S. Churchland 1986), is also a challenge to introspection—with respect to propositional attitudes. This is the view that propositional attitudes such as beliefs, desires, and intentions do not actually exist. On this view, a subject’s self-ascription that she has an intention to go to the cinema or a belief that snow is white would be mistaken, since such mental states do not exist. If subjects do not have propositional attitudes, then they do not have introspective access to them, and thus they cannot have first-person authority with respect to them. As this view has few supporters, and has proven unpopular in the last few years, I will not address it here.

<sup>5</sup> Rosenberg (2016) endorses Carruthers’ view, and Cassam (2014, p. 139) notes that his own view is similar to Carruthers’.

<sup>6</sup> Proponents of this view include Putnam (1975) and Burge (1979).

its entirety, is: '[c]ut the pie any way you like, "meanings" just ain't in the *head*!' (1975, p. 227).<sup>7</sup>

Content externalism relates to the account of first-person authority I am developing here in the following way. If the content of one's thoughts is not wholly determined 'in the head', and is partly determined by one's environment, then it appears that introspection *alone* is incapable of giving a subject knowledge of what she believes, intends, desires and so on. If I need to have knowledge of my environment in order to determine what I believe, then two of the conditions that I claimed are required for introspective knowledge—the first-person condition, and the difference condition—cannot be met.

In the introduction, I said that content externalism is less radical inferentialism. How can this be so if they both deny two of the conditions—namely, the first-person condition and the difference condition—that I said were required for introspection? Wouldn't they both be just as radical? In my view they are not, because I do not think that content externalism does violate these two conditions. I will argue there is no conflict between content externalism and introspection.

Before getting to the conflict between content externalism and introspection, it is important that I say a bit more about mental content. This can be done by considering a typical example. Suppose that I am thinking about Australian lakes and form the belief that there is water in Lake Eyre. This belief,

---

<sup>7</sup> This slogan is somewhat misleading, in my view. It should perhaps say that not all meaning is in the head; as it seems to suggest that all content is external, which would not accurately describe content externalism. What is it for something to lie inside versus outside the head? Recently, Brie Gertler (2012b), has argued that a coherent conception of 'in the head' versus 'out of the head' (internal versus external properties) cannot be given. Gertler is therefore sceptical of whether a coherent version of content externalism can be devised. Although I am sympathetic with Gertler's concerns, I think the current theory-neutral account of introspection (as outlined in chapter 1) can provide guidance here. When I refer to a source of self-knowledge as being 'in the head', I mean that it is knowledge gained by introspection (in the theory-neutral sense); and when I refer to a source of self-knowledge as being 'out of the head' I mean that it is knowledge gained non-introspectively (in the theory-neutral sense).

like other thoughts I am capable of having, has content—which in this case would be ‘There is water in Lake Eyre’. Other mental states, such as my hope that there is water in Lake Eyre, would share the same content as my belief that there is, but would differ in psychological type (see Lau and Deutch 2014).

In recent years, controversy has arisen surrounding the question of whether a subject can have knowledge of the content of their propositional attitudes (e.g., beliefs, desires, intentions, and so on) without needing to have further knowledge of the facts in their environment. So, in our example above, the question would be: ‘Can I come to have knowledge of my belief there is water in Lake Eyre without investigating my environment?’ This debate has been characterised as one between *internalists* (see, e.g., Mellor 1977; Searle 1983)—those that think that mental content is *narrow*, meaning that it is fully determined by what goes on in the head; and *externalists* (see, e.g., Putnam 1975; Burge 1979; Davison 1980)—those that think mental content is *wide*, meaning that such content is determined by not just what goes on in the head, but also from facts which lie within one’s environment. To say that one must rely on the environment in order to have knowledge of content is to say that one must have knowledge of the causal route that one came to learn the words one uses; that one must know how one’s words are employed in specific linguistic communities; and so on (see Kriegel 2008). On the internalist view, I can know that I believe that there is water in Lake Eyre without investigating my environment; on the externalist view, I cannot.

This debate is relevant to our current discussion because, as I mentioned above, there is controversy over whether content externalism is compatible with introspective self-knowledge, and first-person authority. As I am interested in

what implications content externalism has for first-person authority, I will be primarily focused on the compatibility question, and will not assess any arguments for or against content externalism.<sup>8</sup> According to a recent poll, taken in the year 2009 (see Lau and Deutch 2014), a slight majority (52%) of professional philosophers accept externalism. So, although the internalism versus externalism debate remains a matter of some controversy amongst contemporary philosophers—a debate which I will not attempt to enter into here—the compatibility question is one that is not only relevant to our current concerns, but is also relevant to the slight majority of contemporary philosophers who accept externalism.

Let us call anyone who thinks that externalism precludes first-person authoritative knowledge of one's own propositional attitudes, as per custom, an *incompatibilist* (see, e.g., Boghossian 1989; McKinsey 1991). An incompatibilist can be understood as someone who thinks the following two claims are incompatible with each other:

(CE) Content externalism: the content of our own thoughts is determined in part by the environment.

(IKC) Introspective knowledge of content: one can have knowledge of the content of one's thoughts by introspection alone.

Let us call the incompatibility of CE and IKC the *incompatibility thesis*. In what follows, I will examine an argument for the incompatibility thesis that is given by one incompatibilist, Paul Boghossian (1989), who contends that a subject

---

<sup>8</sup> Burge (1988) and Davidson (1987) also argue that content externalism is compatible with first-person authority. Their accounts of first-person authority differ from mine, however.

cannot have knowledge of the content of their thoughts by introspection alone, given the truth of externalism. Before looking at Boghossian's argument for the incompatibility thesis, I first address some terminological issues.

It will be noted that I have set up this conflict as one between CE and introspective self-knowledge. Although this is often done in the literature (see, e.g., Boghossian 1989; Warfield 1992), it is not uncommon to see the debate set up as one between CE and '*a priori* self-knowledge' (see, e.g., Gallois 1996; Lau and Deutsch 2014), or even 'armchair self-knowledge' (see, e.g., Parent 2017). Although I think that there is merit in these alternative approaches, I will set the dialectic up between CE and introspection—in theory-neutral sense laid in out chapter 1—because doing so makes it clear just how it is that content externalism relates to our concerns about first-person authority.

We can begin our analysis of the incompatibility thesis by considering the following from Boghossian, who says:

[n]ow, doesn't it follow from such anti-individualistic views [here Boghossian is referring to content externalism] that we cannot know our thoughts in a direct, purely observational manner? The following line of reasoning might seem to lead rather swiftly to that conclusion. To know my water thoughts, I would have to know that they involve the concept water and not the concept twater without investigating my environment...I could hardly know such facts by mere introspection (1989, p. 12).

Although Boghossian is not explicit in this passage about what is meant by introspection, or how introspection relates to first-person authority, he does make clear his contention that an *investigation* of one's environment will be required in order to acquire knowledge of the facts he mentions about water. As such, he thinks that introspection alone will not be sufficient to have knowledge of his thoughts.

In order to clarify this challenge to introspection, I will now examine a thought experiment (we can call this the *switching situation*) that Boghossian (1989, p. 13) offers as support for the incompatibility thesis. In this thought experiment, we are to imagine that a subject is transported between two different linguistic communities, on two separate planets: Earth and a virtually identical planet to Earth called Twin Earth (a planet that is identical to Earth except that instead of water being composed of H<sub>2</sub>O, it is composed of a chemical compound called XYZ or ‘twater’). According to this thought experiment, we are to imagine that a subject is switched back and forth between planets, in her sleep, so that she cannot tell whether she is on Earth or Twin Earth (Boghossian stipulates that the subject is unaware of the switch, and nothing in her environment can inform her about what planet she is on). Boghossian thinks that after some time of switching back and forth, and having spent a sufficient amount of time in each community, it is reasonable to think that the subject will not be able to tell whether her thoughts are about water or twater. For example, she will not be able to determine whether she believes:

(1) There is water in lake Eyre; or

(2) There is twater in lake Eyre

The implication of this thought experiment, according to Boghossian (1989, p. 14), is that since someone in the switching situation cannot tell by introspection alone whether they are entertaining thought (1) or (2), they cannot tell, by introspection alone, whether they are entertaining (1)—after all it may possible

that they are entertaining (2). Therefore, a subject in the switching situation would not be able to establish, without investigating her environment, whether the following was true:

(3) I believe that there is water in Lake Eyre and do not have the thought  
I believe that there is twater in Lake Eyre.

Now that we have examined the switching situation thought experiment, and noted the implications that it has for introspective self-knowledge, I will now attempt to challenge Boghossian's claim that IKC is incompatible with CE.

First, let us consider a practical response. Here one might say that although the switching situation would indeed preclude a subject from being able to establish (3), without investigating the world, none of us are actually in the switching case, so the implications of this result are irrelevant, practically speaking, to us. This is a reply given by Ted Warfield (1992), who argues that, at best, all that the thought experiment establishes is that a subject in the switching case may not, *necessarily*, know, by introspection alone, (3). But since none of us are in the switching case, we should not think that this thought experiment has any relevance to practical matters. Warfield claims that the thought experiment is only relevant to the following question: 'Given externalism, is it *necessary* that the contents of a thinker's thoughts are knowable to the thinker on the basis of introspection?' (1992, p. 235). Warfield thinks that while the answer to this question is 'no', there are no far reaching consequences, because no one on Earth is in the switching situation.

Although Warfield's point here is a good one, it does not get to the heart of Boghossian's concern. The fact that we (inhabitants of Earth) are not in the switching situation is something that can only be determined by investigating the world. And thus, without investigating the world, a subject will not be able to answer (3). Since this answer doesn't seem to address a subject's inability to differentiate their water thoughts from their twater thoughts, further explanation is required.

One reason why Warfield's response is inadequate—as is pointed out by Peter Ludlow (1995)—is that although in actuality we may not get taken away in our sleep to Twin Earth, we do, often unknowingly, move from one linguistic community to another, meaning that the contents of our thoughts may sometimes shift without our awareness.<sup>9</sup> This means that someone who switches between different linguistic communities may well be in the same situation as the subject in the switching situation, and may fail to be able to distinguish their *P* thoughts from their *P\** thoughts. Given that such occurrences are not only conceivable, but most likely actual, I do not think Warfield's reply is adequate; and thus, we must pursue an alternative response.

A more plausible reply, one which I will draw upon here, has been suggested by Kevin Falvey and Joseph Owens (1994), who distinguish between introspective knowledge of *content*, and introspective knowledge of *comparative content*. We are familiar with mental content, but what is introspective comparative content? According to Falvey and Joseph Owens, an individual has introspective knowledge of comparative content of any two of one's mental states when 'an individual can know authoritatively and directly (that is, without

---

<sup>9</sup> See Ludlow (1995, p. 47) for a real-world scenario involving someone who uses the word 'chicory' while traveling between England and America—counties which use this word in different ways.

relying on inferences from his observed environment) whether they have the same content (1994, pp.109–110). To know, by introspection, the comparative content in the switching case, one would have to know that one believed that there is water in Lake Eyre and did not believe that twater is in Lake Eyre without first observing one's environment.

Falvey and Owens concede that a subject in the switching case could not differentiate between a water thought and twater thought, as is stated (3). What they disagree with, however, is that this concession helps to support the incompatibility thesis. They claim that the comparative content of two thoughts is not required in order to know, by introspection, that one believes something. The result of the switching case, then, does not support the incompatibility thesis. All that this thought experiment shows is that introspection cannot give us knowledge of the comparative content of any two distinct thoughts. So, while I can know that I believe that water is wet by introspection, I cannot know by introspection alone, whether my thought refers to water or twater. For this, I will need to investigate my environment.

I agree with Falvey and Owens that even if someone in the switching situation could not tell whether she has a water or twater thought, without first investigating her environment, this does not mean that a subject cannot have knowledge of her own thoughts by introspection alone. All this shows is that we lack introspective knowledge of comparative content (as opposed to content), which is not just a problem that arises when externalism is assumed. In order provide to provide further support for this claim, I want to now shift the current discussion to the topic of *context*. Doing so, I claim, will help show why accepting that fact that a subject in the switching situation could not tell whether

she has a water or twater thought, does not preclude introspective knowledge of our thoughts given externalism.

To see how context is relevant to the current discussion, let us consider an example provided by André Gallois (1996, p. 181). Gallois claims that if he was in a room and asked the question ‘Is there an object in front of you?’ he would be unable to answer it until the specific context was given. This is because the question is ambiguous. If the specific context was given, and he was asked ‘Is the object in front of you a table or a chair’ he would have no trouble answering it. However, if the question was ‘Is the table from the next room, or some other room?’ he would not be able to answer it until he investigated the situation further. Gallois’ point here can be extended back to our above discussion by looking at the different forms that the question ‘Do you believe that there is water in Lake Eyre?’ make take. If the specific context of the question is ‘Do you believe that there is a watery substance, rather than green goo, in Lake Eyre?’ I will have no problem answering the question, without first having to investigate my environment. However, as with the table example, if the specific context of the question is such that it pertains to the causal origins of my belief—for example, ‘Is your thought a water or twater thought?’—I will *not* be able to answer the question, until I have first investigated my environment. As stated above, one cannot not know, by introspection alone, such comparative content. Like the case involving the table, however, while such comparative content may be relevant to my belief, it is not pertinent to the self-knowledge of the belief itself. If someone just wants to know whether they have a certain belief—namely, that they believe that there is water in lake Eyre—then the switching situation will not be enough to preclude the self-knowledge one can

have of one's belief. What someone in the switching situation will *not be able* to do is have introspective knowledge of some of the features of one's thought, such as the causal history of the belief. I do not think, however, that such an acknowledgement is enough to show that one cannot have knowledge of one's own propositional attitudes, by introspection alone.<sup>10</sup>

### 3.2 The Inferentialism Challenge

The second challenge to introspection that I shall discuss comes from a view of self-knowledge that we may refer to as *inferentialism*. Proponents of this view (see, e.g., Ryle 1949; Gopnik 1993; Carruthers 2011; Cassam 2014) claim that the self-knowledge of our propositional attitudes (e.g., our beliefs, intentions, desires, and so on) is derived *solely* by an inferential process.<sup>11</sup> Cassam, for example, characterises inferentialism as the view that the 'knowledge of our own beliefs, desires, hopes, and other "intentional" states is first and foremost a form of inferential knowledge' (2014, p. 137).<sup>12</sup>

Inferentialism is antithetical to introspection because knowledge acquired by an inferential process is typically construed as being less direct than knowledge acquired by an introspective process. This is because inferential processes are typically reliant on some third-person verifiable evidence. For

---

<sup>10</sup> In recent years, some philosophers (see, e.g., Stalnaker 1981; Chalmers 2006) have developed detailed accounts of meaning that can be used to assign truth values to the narrow and wide content of a thought. Such views have been called *two-dimensional (2D) semantics*. (The view gets its name because theorists construct a two-dimensional framework to represent truth values.) As David Chalmers says, '[t]he core idea of two-dimensional semantics is that there are two different ways in which the extension of an expression depends on possible states of the world. First, the actual extension of an expression depends on the character of the actual world in which an expression is uttered. Second, the counterfactual extension of an expression depends on the character of the counterfactual world in which the expression is evaluated' (2004, p. 59)

Such an account, then, allows us to capture the two different dimensions of meaning that thoughts have. For example, such an account allows us to assign truth-values to the content of the subject's thoughts in the switching case, in order to see under what circumstances the subject's thoughts are true or false. With respect to the internalist versus externalist debate, 2D semantics is neutral. However, internalists and externalists would disagree about what constitutes *the meaning* of a sentence (see Laura Schroeter 2017).

<sup>11</sup> As I go on to show, inferentialists differ on the scope of this claim.

<sup>12</sup> Recall that the term 'intentional state' is often used interchangeably with the term 'propositional attitude.' As I stated in the introduction, in this chapter I am concerned only about propositional attitudes.

example, if I were to see someone grimacing after being struck by a wayward tennis ball, I would infer from their behaviour that they are in pain. With respect to my own pain, however, it doesn't seem like I have to infer anything to know that I am in pain. It seems like I can know that I am in pain by introspecting my own pain sensations. Inference, then, is not something that philosophers typically associate with introspection. Dan Zahavi, for example, says 'first-person ascriptions of psychological states are immediate in the sense of not being based on observational or inferential evidence' (2005, p. 13). Inferentialists such as Cassam, by contrast, describe the process by which we achieve self-knowledge of our propositional attitudes in exactly these terms—as based on observational or inferential evidence.<sup>13</sup> Inferentialists thus deny that there is any significant asymmetry between the way that we acquire knowledge of our own propositional attitudes and the way in which we achieve knowledge of the propositional attitudes of others. Cassam thinks of such an asymmetry as 'another philosophical myth' (2014, p. 153). Since this appears to violate one of the conditions I said was required for a process to count as introspective—namely, the difference condition—inferentialism is antithetical to what I will call the *introspection thesis*: the epistemic thesis that we can introspect our own propositional attitudes.

To further demonstrate why inferentialism precludes introspective justification of our propositional attitude self-ascriptions, let me introduce a distinction that Declan Smithies (2014) makes between inferentially and non-inferentially justified beliefs. Smithies claims that a belief is inferentially justified if and only if it depends upon the justification of other beliefs; and a

---

<sup>13</sup> As Cassam points out, there is no 'magic bullet' (2014, p. 140) that inferentialists claim will count as evidence here. Cassam thinks that such evidence is not just limited to behaviour. He holds that inner speech, dreams, and feelings could also be used as evidence (2014, p. 138).

belief is non-inferentially justified if and only if it depends upon the justification of the mental state itself. An example will help to clarify this concept. Suppose I believe that I have the intention to see the movie *Star Wars*. If such a belief is non-inferentially justified, then it will be justified by the intention itself to see *Star Wars*. Now, suppose I were to ascribe the same intention to Jane—that is, I were to believe that Jane intends to see the movie *Star Wars*. To do so, I will need to observe the evidence of Jane’s behaviour or listen to Jane’s avowals of interest to see *Star Wars*. These facts are not the mental state itself. Thus, inferential knowledge differs from non-inferential knowledge because it requires evidence gathering.

According to inferentialism, my own beliefs about what propositional attitude mental state I am currently in will be justified in the same way that I justify my beliefs about what Jane intends to do. My beliefs about my own propositional attitude events cannot be justified by the mental state itself. As Cassam says, the ‘concept of first-person awareness, or self-knowledge, which inferentialism is trying to capture is that of awareness or knowledge that *is* based on evidence, behavioural or otherwise’ (2014, p. 138). Another inferentialist, Peter Carruthers, offer a similar analysis:

[w]hat we take to be non-inferential access to our purely propositional thought is, in reality, the result of a swift bit of self-interpretation, but one that takes place so smoothly and quickly that we do not know what we are doing (2005, p. 130).

Notice here that both authors are denying that there is any unique first-person way that we can achieve knowledge of our propositional attitudes. Inferentialism’s positive claim, therefore, involves a negative claim. By saying that one must rely solely on inferential evidence to know what one believes or

intends, the inferentialist is denying that one can *sometimes* have introspective access to one's own propositional attitudes. Subsequently, if inferentialism is true, we would not have first-person authority, in the way that I construe it. This is because one cannot justify by introspection one's beliefs about what current propositional attitudes one has.

I have given a brief overview of inferentialism, and shown why the view precludes the introspection thesis. What grounds are there for accepting this view, however? Some contemporary inferentialists, such as Cassam, argue that one reason to accept inferentialism is that the rival views of self-knowledge proposed in the literature are implausible. He says, '[i]t had better be the case that we can know our attitudes by inference because there is no viable alternative to inferentialism' (2014, p. 141). Carruthers (2011) puts forward a similar argument for what he calls the Interpretive Sensory-Access (ISA) theory of self-knowledge, but offers a far more comprehensive defence of the view, compared to the one Cassam offers. For this reason, I will focus on Carruthers' view.<sup>14</sup> While I will attempt, wherever possible, to note to what extent the objections I make to Carruthers' view can be generalised to these other views I have mentioned, I will, for the most part, focus specifically on Carruthers' ISA theory.

### **3.3 Carruthers' Interpretive Sensory-Access (ISA) Theory**

According to Carruthers, 'the ISA theory maintains that the mind contains a single mental faculty charged with attributing mental states (whether to oneself

---

<sup>14</sup> Carruthers (2011, p. xxi) says that his own view is closest to Gazzaniga's (1998). He situates the ISA theory between Gopnik's (1993) view—which he claims goes too far; and Wegner's (2002) and Wilson's (2002) views—which he claims do not go far enough. Cassam (2014) compares his own view to Carruthers' (2011), though he acknowledges there are some key differences between the two views. Another reason to focus on Carruthers' account is that he draws upon *recent* empirical data. As Georges Rey (2013, p. 26), points out, a major piece of evidence for Gopnik's (1993) view—evidence from the false belief studies—is now seen as highly controversial by many (see Onishi and Baillargeon 2005).

or others), where the inputs to this faculty are sensory in character' (2011, p. 1).<sup>15</sup> In other words, all psychological ascriptions that a subject is capable of making, whether to themselves or to another, will ultimately be grounded in some sensory experience of the person who is making the ascription.<sup>16</sup> Any self-knowledge that a person achieves that is *not* of a sensation—such as that of a propositional attitude or a character trait—will be acquired *indirectly*, or *inferentially*, in the way described in the previous section. The ISA theory, as Georges Rey succinctly puts it, 'suggests that our introspectable lives consist *merely* of sensations' (2013, p. 264). By the terms 'sensory' and 'sensations' Carruthers is referring to 'all forms of perception' (2011, p. 1) one is capable of having, from the first-person point of view.<sup>17</sup> This encompasses the five traditional senses: sight, hearing, taste, smell, and touch; as well as others such as interoception and proprioception.<sup>18</sup>

It is important to point out that the ISA theory does not entail the thesis that propositional attitudes do not exist. Neither does the ISA theory entail the thesis that we cannot have self-knowledge of our propositional attitudes. What the ISA theory does claim is that our knowledge of such mental states will always be based upon our own sensory input—just as our knowledge of the propositional attitudes of others will always be based upon some sensory input.

---

<sup>15</sup> Carruthers's main aim in his book length defence of the ISA theory, *The Opacity of Mind*, is to challenge the commonly held view that 'knowledge of our own mental states is somehow special, and radically different from other-knowledge' (2011, p. xii).

<sup>16</sup> These sensory experiences themselves are introspected.

<sup>17</sup> This suggests that the terms 'sensation' and 'perception' should be understood as analogous. Although this is commonly understood to be the case, some philosophers, such as P.M.S Hacker (2013), think it is important not to conflate the two. For example, Hacker (2013, pp. 273–275) thinks that illusions are *possible* with respect to perceptions (e.g., one can hallucinate a perception of a tomato); whereas sensations are not susceptible to hallucination (one cannot hallucinate a sensation of a tomato). As I interpret Carruthers, he uses the terms 'sensation' and 'perception' to mean what Hackers calls a 'sensation'.

<sup>18</sup> According to Frédérique de Vignemont, *proprioception* 'provides information about the position and movement of the body. The mechanisms of proprioception include muscle spindles, which are sensitive to muscle stretch, Golgi tendon organs, which are sensitive to tendon tension, and joint receptors, which are sensitive to joint position...*Interoception* provides information about the physiological condition of the body in order to maintain optimal homeostasis, namely, cardiovascular, respiratory, energy (feeding and glucose), and fluid (electrolyte and water) balances' (2015).

For example, in order to know what another person desires, I must *perceive* their actions, or *hear* them articulate their own thoughts. According to the ISA theory, the process by which I acquire knowledge of my propositional attitudes—e.g., my belief that *P*, or my intention to do  $\Phi$ —is the same *in kind* as the process that I acquire knowledge of the propositional attitudes of another.<sup>19</sup> (Note that we can already see that this appears to violate the difference condition for introspection that was given in chapter 1.5.) The basic idea of this division is outlined by Carruthers as follows:

introspection is here divided into two categories: introspection of propositional attitude events, on the one hand, and introspection of broadly perceptual events, on the other. I shall assume that the latter exists while arguing that the former doesn't (2010, p. 76).

One issue that immediately arises with this division—between broadly perceptual events (e.g., sight, hearing, touch, and so on) and propositional attitude events (e.g., beliefs, desires, and so on)—is that some propositional attitudes appear to be embedded in perceptions. As Rey (2013, p. 265) has pointed out, many sensory mental states appear to involve the attitudes that the ISA theory suggests that we cannot have introspective access to. Rey notes that ‘pain involves *aversion*; hunger a *desire* to eat; thirst a *desire* to drink; fear a *belief* that danger is imminent’ (2013, p. 254). Some propositional attitudes, therefore, appear to ‘come free’ with one’s perceptions. If it is true that one can introspect one’s sensory states, then it seems that one will be able to introspect a certain class of propositional attitudes too. This is not what we should expect from a theory that says that we cannot introspect our propositional attitudes.

---

<sup>19</sup> It may be objected that this view is easily refuted because one will still be able to self-attribute one’s own propositional attitudes even if one is in a dark room, unable to observe any of one’s own behaviour (see Rey 2013). Carruthers (2011, p. 158) own response to this objection is that such a person will still have visual imagery, inner speech, and affective feelings to draw upon.

Such exceptions require Carruthers to make a distinction—namely, between *perceptual* and *non-perceptual* propositional attitudes. He claims that the ‘mindreading faculty lacks direct access to the subject’s own non-perceptual judgements, decisions and other propositional attitudes’ (2011, p. 69).<sup>20</sup> This qualification allows Carruthers to acknowledge what has been said above—namely, that some propositional attitudes are embedded in broadly perceptual events, while at the same time still affirming the ISA theory. Since many, if not arguably most, propositional attitudes will fall into the *non-perceptual* category, the view is still revisionary. According to the ISA theory, then, perceptual beliefs such as ‘I believe that there is a computer screen in front of me’ can be grounded by introspection, without any self-interpretation, or inference. Whereas non-perceptual beliefs, such as ‘I believe that the British ruled India until 1947’, will require some self-interpretation.

According to Carruthers, the ISA theory can be seen as a conjunction of the following three theses:

- (a) there is a single mental faculty underlying our attributions of propositional attitudes, whether to ourselves or others;
- (b) this faculty has only sensory access to its domain;
- (c) its access to our attitudes (or rather to *most* kinds of attitude) is interpretive rather than transparent (2011, pp. 1–2).

Although these three theses are stated independently, they are importantly related: given (a), that there is only a *single* mental faculty for attributing both our own propositional attitudes and the propositional attitudes of others, and given that we do not have introspective access to the propositional attitudes of

---

<sup>20</sup> Carruthers leaves open the question of what constitutes a perceptually-embedded attitude (2011, p. 83). For example, he points out that a grandmaster in chess may be able to ‘see literally...that white is doomed to lose’ (2011, p. 84) after looking at a particular chess configuration.

others, it must follow that we do not have introspective access to own propositional attitudes. How we *do have knowledge* of our propositional attitudes is stated in (b); and further articulated in (c).

What reasons are there for accepting the ISA theory? One of the main arguments that Carruthers gives for the ISA theory comes via the evidence from the confabulation data. He regards this as ‘the central, key, prediction made by the ISA theory’ (2011, p. 6). The confabulation data that Carruthers refers to here comes from cases where subjects in experimental settings have falsely, and *sincerely*, self-attributed mental states.<sup>21</sup> While subjects are capable of confabulating a variety of types of mental states and processes—such as beliefs, rationalisations, judgements, intentions, perceptions, and memories—the main focus in this chapter and the next will be mostly with how the confabulation data affect propositional attitude self-attributions; and the subsequent implications the data have for introspection, as a way to know one’s own propositional attitudes.

The sorts of examples that Carruthers cites come from cases where subjects have confabulated a propositional attitude—typically in experiments that have been carefully set up by psychologists and neuroscientists. On Carruthers’ view, such data give us a compelling reason to accept the ISA theory, as the ISA theory predicts that subjects will often confabulate (2011, p. 1). Moreover, one of Carruthers’ main criticisms against most other theories of introspection is that although they may have the resources to accommodate the *existence* of these data, few can accurately predict or explain the *pattern* of error

---

<sup>21</sup> Carruthers appears to use the term ‘confabulation’ to mean any kind of self-attribution error that a subject sincerely avows. For example, suppose that Jill sincerely claims that she is anxious about an upcoming physics examination, when in fact she isn’t. This would count as an example of confabulation, on such a construal, since Jill sincerely self-attributes a mental state to herself, which she does not have. Some care needs to be taken with the term, however. As William Hirstein points out, ‘[a]nyone broaching the topic of confabulation is faced immediately with a huge problem: there is no orthodox, problem-free definition of “confabulation”’ (2009, p.2). This issue will be discussed at greater length in chapter 4.

present in these data (2011, p. 6). The fact that the ISA theory can do this is considered by Carruthers a major virtue of the theory.<sup>22</sup> I think, therefore, that the ISA theory is important to consider in detail, because most other theories of self-knowledge make no mention of such data.

In my view, while it is indeed a virtue of the ISA theory that it can explain the presence of such data, this is not enough to render it plausible. To claim that introspective access to our propositional attitudes is an illusion, I think, is overstated. It will be my aim in what remains in this chapter, and also the next, to show why I claim this. My strategy will be the following. In §3.4, I offer two objections to the ISA theory. As I do not claim that these objections are decisive, I will then, in section §3.5, turn to the role that the confabulation data play in the argument for the ISA theory. Following a discussion of one of the arguments for the ISA theory, in §3.5, I then, in next chapter, examine some actual examples of confabulation data that Carruthers cites as evidence for this theory. I reject Carruthers' claim that such data do indeed support his theory.

### **3.4 Two Objections to the Interpretative Sensory-Access (ISA) Theory**

With the main details of the ISA theory laid out, I will now raise two objections to it. I do not wish to suggest that these two objections are insurmountable. All I will argue here is that by accepting the ISA theory, one would be forced to accept two highly controversial theses. Since Carruthers does not offer us with much of an explanation for how such controversial theses could be true, apart from the fact that they are implied by the ISA theory, I think that his acceptance of the ISA theory is premature.

---

<sup>22</sup> There is some truth to be found in this claim. In a recent volume on the topic of self-knowledge, *Consciousness and Introspection*, edited by Declan Smithies and Daniel Stoljar (2012), the confabulation data are only briefly mentioned.

### 3.4.1 The Accuracy Objection

The first objection that I will raise pertains to the accuracy of our propositional attitude self-ascriptions. The accuracy in question here pertains to the sense in which our propositional attitude self-ascriptions are more likely to amount to knowledge compared to the ascriptions that others make about our own minds. I will call this the *accuracy objection*.<sup>23</sup> It runs as follows: if the process by which we self-attribute our propositional attitudes really is inferential, or interpretative, as the ISA theory states—meaning it is no different in kind from the way in which others attribute mental states to us—then why are we generally so much better at self-ascribing our own propositional attitudes than others? If, as the ISA theory maintains, the process by which we achieve knowledge of our propositional attitude is always predicated upon our own sensory experience, then shouldn't we expect that the accuracy of our own self-ascriptions to only ever be as accurate as the ascriptions we make about another's psychological states?<sup>24</sup>

There are at least two ways for the proponent of the ISA theory to respond here. First, one could simply deny that our self-ascriptions really are so reliable. One could argue that the accuracy objection begs the question—that is, one could argue that the common-sense belief that our self-ascriptions are accurate is an illusion. Carruthers considers such a response, and says the

---

<sup>23</sup> Rey (2013, p. 274) raises a related objection, which he calls the reliability objection. He claims that the ISA theory appears to be at odds with our ability to reliably self-attribute—that is to say, our self-ascriptions are usually true. This is a legitimate objection—and one that I share. The accuracy objection that I am raising here pertains to our ability to be *more accurate* in our self-ascriptions compared to ascriptions that others make about our own mental states.

<sup>24</sup> This is because the (inferential) ascriptions we make about other peoples' minds are always based upon some perception. In order for me to know that another person believes that snow is white, for example, I must perceive his actions, or hear him speak. Carruthers thinks that the same is true with respect to a person's own propositional attitudes. According to the ISA theory, in order to know that one believes that snow is white, for example, one must perceive one's own actions, or interpret one's own feelings, and make an inference based upon those data.

‘question whether any such certainty and special reliability really exists is precisely what is at stake in these debates, however, and can’t be stipulated at the outset’ (2011, p. 157).

Although this point is a legitimate one, and is one that would avoid the objection altogether, I think that it is problematic to deny that our self-ascriptions *are* generally successful, and no more accurate than the ascriptions one makes about another person’s psychological states. Even philosophers who are sceptical of the accuracy of introspection, qualify their scepticism in comparative terms—that is, in comparison to the accuracy of our visual perceptions. Not in terms of introspective ascriptions being less accurate than the ascriptions we make about another person’s mental states.

One recent sceptic of the reliability of introspection, Eric Schwitzgebel (2011), who makes much of the errors made in self-ascriptions, characterises his critique in terms of introspection’s relation to perceptual knowledge. As Smithies has pointed out, such scepticism is best stated in comparative terms (2013, p. 1179), meaning that Schwitzgebel’s scepticism is best seen as one that says that knowledge by introspection is *no more reliable* than perceptual knowledge. It is, however, one thing to claim that introspection is no more reliable than perception, and quite another to say that it is no more reliable than the ascriptions we make of other people’s mental states. Carruthers says nothing about how such accuracy could be an illusion.

The second—and, in my view, more plausible—way for the ISA theorist to respond would be to acknowledge the fact that we generally *are* better in ascribing mental states in our own case, and then proceed to explain how this can occur without introspection. Carruthers suggests one way in which this might be

done: he thinks that ‘people are, probably, excellent interpreters of themselves’ (2009, p. 127). Although this answer would be one way in which a proponent of the ISA theory could respond to the accuracy objection, it lacks explanatory power. Carruthers does not offer us much of an explanation for how we could be such excellent interpreters. After all, we do not seem to be excellent interpreters with respect to other types of mental phenomena that we can only have inferential access to, such as our character traits or the causes of our desires. Why should the case be different with our propositional attitudes? Carruthers does not have an explanation for how this could be the case. While I do not think that such an objection is insurmountable, I do think it raises a serious problem for the proponent of the ISA theory.

### **3.4.2 The Sensory Data Objection**

The second objection to the ISA theory that I shall raise is one which I will call the *sensory data objection*. This objection challenges one of the main theses of the ISA theory, namely thesis (b), which maintains that the introspective faculty humans possess has only sensory access to its domain.<sup>25</sup> The sensory data objection runs as follows: if we consider the various kinds of propositional attitudes that we are capable of self-attributing—from beliefs to desires to intentions and so on—it is difficult to imagine that in *all* instances of self-attribution, there exists some sensory data which a subject is drawing upon to make a self-attribution.

Consider, for example, some of my occurrent thoughts: my belief that the Northern White Rhino is endangered, my intention to leave work at 5pm to see a

---

<sup>25</sup> This just means that all of one’s own self-ascriptions will be grounded upon sensations.

film, or my hope that the Australian cricket team will win the Boxing Day Test match. It does not seem true to say that in order for me to self-ascribe each mental state, I need to draw upon some perceptual data. Moreover, if we think about the vast range of propositional attitudes that exist, it seems very unlikely that there also exists perceptual data upon which to base *every* propositional attitude self-ascription.<sup>26</sup> As Rey has pointed out, ‘[d]esire, wonder, doubt, pretence, curiosity, for example, don’t seem to be linked to any specific sensations’ (2013, p. 274).

I am not claiming that thoughts or propositional attitudes are never associated with some sort of sensory experience. Some desires have a distinct cognitive phenomenology—and arguably so does hoping and fearing. Perhaps less so with belief, but a case can be made for some beliefs. What I am claiming here is that it is very difficult to maintain that in every case of propositional attitude self-attribution, there is some sensory input. Since I think such a position is very difficult to maintain, the sensory data objection provides a considerable difficulty for the ISA theory.

Carruthers may reply that, contrary to what has been said here, we do draw from perceptual cues when we self-attribute our propositional attitudes (as is described by the ISA theory). This is, obviously, a fair response. However, it is a response that simply cannot be claimed without further evidence. Given our considerations above—namely, that there doesn’t seem to be perceptual input

---

<sup>26</sup> This may seem to commit Carruthers to a very strong version of phenomenal intentionality, which very few philosophers would be willing to accept—namely the view that David Bourget and Angela Mendelovici (2016) have called the strong phenomenal intentionality theory (hereafter, ‘strong PIT’). According to strong PIT, *all* propositional attitudes are phenomenal states. That is, all propositional attitudes are grounded in phenomenal consciousness. On this view, even one’s belief that grass is green would have some sort of phenomenology associated with it. If Carruthers were committed to this view, I think his position would be even more difficult to accept. In response to this concern, I think Carruthers would deny that he is committed to strong PIT, because he argues against there being phenomenal consciousness associated with propositional attitudes in the first place (see Carruthers and Veillet 2011).

associated with every propositional attitude self-attribution—the burden of proof is squarely on Carruthers, to show how we could be so mistaken in our common-sense thinking. Because no explanation is forthcoming, the sensory data objection provides a significant problem for the ISA theory.

To recap, I've offered two objections to Carruthers' ISA theory. The first objection was the accuracy objection, which stated that the ISA theory appears to be at odds with our ability to accurately attribute mental states to ourselves. The second was the sensory data objection, which challenged one of main components of the ISA theory—namely, that every propositional self-ascription must draw upon sensory data.

While I do not take these criticisms to be decisive, I do think that when considered together, they provide a major obstacle for the proponent of the ISA theory to overcome. As we have not yet examined any of the confabulation data—the key prediction made by the ISA theory—it is premature to rule out the theory at this stage. I will now show how Carruthers uses the results from the confabulation data in his argument for the ISA theory.

### **3.5 Confabulation and Interpretation**

So far, I have explained the main details of the ISA theory, and have raised two preliminary objections. I have not claimed that my objections are insurmountable, but rather that they should be thought of as counterintuitive implications of ISA theory. I will now consider how Carruthers implements the confabulation data (the experimental evidence from cognitive science that show subjects falsely self-ascribing mental states) into a positive argument for the ISA theory. The existence of the confabulation data is, in Carruthers' view, the

‘central, key, prediction’ (2011, p. 6) that the ISA theory makes. As such, we will need to carefully examine the data, as well as the arguments for the ISA theory based upon the data, in order to complete our critique of the theory. I will call such appeals to the confabulation data: *confabulation arguments* for the ISA theory.

In what follows, I will examine two distinct confabulation arguments. I will call the first of these arguments the *ontological parsimony argument*. This is the argument that, given the existence of the confabulation data, the ISA theory is the simplest theory; and since, all things being considered, simplicity is a theoretical virtue, the ISA theory should be preferred. This argument will be examined in §3.6. I will call the second of these arguments the *patterning argument*, which says that the patterns of error present in the confabulation data are best explained by positing the ISA theory. This argument will be examined in chapter 4.

Before getting to these arguments, I will present an example of confabulation. Although I will examine a set of the data in more detail, in the next chapter, it will be useful to have an idea about what confabulation is, while discussing the confabulation arguments. The example I will present is one that comes from the much discussed commissurotomy (hereafter, ‘split-brain’) patients. Split-brain patients are subjects who have had their corpus callosum (the connection between the left and right hemisphere of the human brain) cut. With such patients, it is possible to expose information to only *one side* of the brain by *only* giving stimulus to the corresponding visual field. If, for example, information was to be flashed to the subject’s right visual field, then such information would be available to the left side of the brain. In one specific case,

though these cases have been replicated (see Gazzaniga 1995, 2000), a ‘walk!’ sign was flashed to the right hemisphere of the subject by the experimenters (the hypothesis being that the right hemisphere controls action). Immediately after this, the subject got up and started walking. When asked why he was walking, the subject answered, ‘I’m going to get a coke’. According to the experimenters, the subject had no conscious awareness of the flashed sign, which, the experimenters suggested was the real reason the subject got up and starting walking.

There are two main features of this example that I wish to draw attention to here that, purportedly, occur in all the confabulation examples I will examine: (i) that a false self-attribution error has occurred, and (ii) that everything seems normal from the subject’s point of view. These two features of the confabulation data are what Carruthers draws upon to support the ISA theory. To get a general idea of *how* he does this, consider the following passage from Carruthers who says:

[t]he argument is not, “We sometimes make errors about our own attitudes, so we might always do so,” or anything of the sort. It corresponds, rather, to the role that errors and illusions play in the cognitive science of vision. That is, the ways in which a system tends to break down can reveal to us something about the manner in which it normally works. And in the present case, what the confabulation data will reveal is that our access to our own thoughts and thought processes is always interpretative (2011, p. 68).

It is important to note Carruthers’ clarification about the form of this argument. As he states, the argument is not just that we make errors, so we might always do. What Carruthers is arguing here is that such errors provide insight into how

the self-attribution process normally works. Such errors are revealing, he thinks, because of the patterns they reveal. Carruthers claims that the ISA theory:

predicts that it should be possible to induce subjects to *confabulate* attributions of mental states to themselves by manipulating perceptual and behavioral cues in such a way as to provide misleading input to the self-interpretation process (just as subjects can be misled in their interpretation of others). Likewise, the account predicts that there should be no such thing as awareness of one's own propositional attitudes independently of any perceptually accessible cues that could provide a basis for self-interpretation. The accuracy of these predictions will be discussed and evaluated in due course (2009, p. 123).

As is made clear in the above passage by Carruthers, it is not just the *existence* of the confabulation data that supports the ISA theory, but it is the *pattern* of the confabulation data—that is, the way in which the subjects have made errors. Carruthers claims that such errors, and the pattern in which they appear, support the ISA theory because such errors are just what should be expected if we only had sensory access to our propositional attitudes. Before getting to this predictive argument, the focus of chapter 4, I will first consider a more fundamental argument—namely, the ontological parsimony argument. This argument suggests that rival theories of self-knowledge will struggle to explain the presence of the confabulation data.

### **3.6 The Ontological Parsimony Argument**

The ontological parsimony argument is the argument that, given the presence of the confabulation data, the ISA theory is simpler—with respect to its ontological commitments—than other accounts of self-knowledge.<sup>27</sup> The basic thought here is that: why posit an introspective method *in addition to* an inferential method (hereafter, the 'dual-method theory') for discerning knowledge of one's own

---

<sup>27</sup> This term 'ontological parsimony argument' is my own, and not Carruthers.

propositional attitudes, when an inferential method alone (hereafter, the ‘single-method theory’) will suffice?<sup>28</sup> In other words, positing a dual-method theory would be superfluous, given that a single-method appears to be sufficient.

This argument can be expanded upon further by breaking down a typical case of confabulation. Let us recall the example mentioned above—namely, the split-brain patient who confabulated an intention to get up and get a drink. Such an example—in addition to the many other examples of confabulation (see chapter 4)—yields the following premise:

- (1) Subjects in experimental settings confabulated a propositional attitude.

Although we can contest (1) in certain circumstances—that is, whether a subject *S* really did confabulate a certain mental state *M* in experiment *E*—the fact that we are capable of confabulating is not controversial. What is controversial, however, is *why* subjects confabulate in the first place; and what, if any, implications the confabulation data have for our current discussion about introspection. One possibility, according to the ISA theory, why subjects confabulate is given in the following:

- (2) The reason why subjects in experimental settings confabulated is because they grounded their beliefs about what current mental state they are in by interpreting misleading sensory cues—and not by any introspective process.

---

<sup>28</sup> This is because we use inference to interpret other people’s mental states

Recall from the confabulation example above, the subject self-ascribed an intention to get a drink that he did not actually have. One explanation, according to (2), is that he did this because he grounded his belief that he had an intention to get a drink by observing his own behaviour. The idea here is that subject *saw* himself getting up from the chair and this led him to believe that he intended to get up and get a drink. For the sake of argument, let us assume that this is the correct way to explain what has occurred in the case.<sup>29</sup>

Even if we accept that this is the right thing to say here, how does such an explanation lend support to the ISA theory? After all, wasn't it conceded in the previous chapter that one can self-ascribe a mental state in this way? Recall I said in the last chapter that one could come to believe that one is angry by seeing one's reflection in the mirror. Carruthers' response to this claim is that: if it is possible to self-ascribe inferentially—on the basis of sensory cues—why should we postulate an additional introspective faculty? He thinks that all we need is the following:

- (3) Single-Method Theory: '[t]he mind contains a single mental faculty charged with attributing mental states (whether to oneself or to others), where the inputs to this faculty are all sensory in character' (Carruthers 2011, p. 1).

On this view, we have an explanation for why subjects confabulate: subjects only have sensory-based access to their propositional attitudes. When those sensory inputs are misleading—e.g., when it *looks* like one is in a certain mental state—such as in the case of the split-brain example, a person will confabulate. The

---

<sup>29</sup> As I will show in the next chapter, it is not clear to me that this right thing to say about this case. I will assume here, for the purposes of this discussion, that it is sound.

question is ‘Since it is possible to self-ascribe by using an inferential method, why posit the addition of an introspective method also?’ Wouldn’t this just add unnecessary complexity?

One way to respond to such a question is to show that there is nothing superfluous about positing a dual-method system. That is exactly what I will argue here. In my view, although there are such occasions where subjects really have confabulated—and did not use introspection—I do not think that this precludes the fact that they *could have*, or even that they possess a faculty that is *sometimes* able to introspect propositional attitudes. Confabulation, or self-interpretation, after all may be rare.

Such a strategy has been recently pursued by Georges Rey (2013) and Alvin Goldman (2006), who advocate the dual-method theory. This is the theory that subjects can acquire knowledge of their propositional attitudes by introspection; and can also acquire knowledge of their propositional attitudes inferentially (non-introspectively). Such a position would entail the following thesis.

- (4) Dual-Method Theory: humans possess two distinct faculties for acquiring self-knowledge of their own propositional attitudes: an introspective and inferential (non-introspective) faculty.

Now clearly (3) and (4) cannot both be true. By accepting (4), that a dual-method theory is true, we deny (3), that the single-method is true, and would thus undermine one of the main theses of the ISA theory. The decision to accept (3) or (4), therefore, plays a critical role in the acceptance or rejection of the ISA

theory. I will now argue that there is nothing superfluous about accepting the dual-method theory, and thus attempt to challenge (3). While the ISA theory—with its commitment to a single-method theory—is less ontologically demanding than the dual-method theory, I will argue that this is not in itself a virtue of the theory.

Apart from simply insisting that the dual-method theory is true, what arguments can be given for thinking that it is not superfluous to posit two distinct processes of achieving self-knowledge of our propositional attitudes? One suggestion is offered by Goldman (2006), who thinks that a non-introspective faculty might act like a backup to an introspective faculty. The idea here is that when introspection is not working properly a second, *inferential* process, can be used by subjects to acquire self-knowledge of their mental states. Although I think that Goldman's suggestion is an interesting one, it raises further questions, rather than explaining the plausibility of the dual-method theory. By suggesting that the inferential process of achieving self-knowledge works like a backup, Goldman implies that the inferential process would be an alternative way for a subject to have self-knowledge about her own mind. This might not seem problematic at first glance, but I think that it is, because it implies that the backup system would be able to perform the same task as an introspective process.<sup>30</sup> If this were the case, then I think Carruthers is right to question why a dual-method would be needed. Thus, I do not think Goldman's suggestion works.

A second suggestion for accepting the dual-method theory is offered by Rey, who says: 'why wouldn't it be a perfectly good idea to have both multiple and overlapping systems? Redundancy of information from different

---

<sup>30</sup> Even if the backup did not work quite as well as the main system, the fact that it could sometimes do the same job as introspection would still require some kind of explanation.

perspectives can fortify confidence' (2013, p. 267). Rey's point here is that a subject could benefit from having two distinct ways of acquiring knowledge of their own mental states. Positing the existence of an introspective way of achieving knowledge of one's own mental states, in addition to an inferential one, could provide a subject with greater amounts of data, which would ultimately help facilitate more accurate self-ascriptions. While I think that this is an interesting response, I think that, like Goldman's response, it raises further questions. For instance, why would the existence of an inferential process help one achieve self-knowledge? Wouldn't introspection alone be enough?

In my view, the best reason for thinking that the dual-method theory is true, is for the simple reason that we are possessors of a perceptual system. As possessors of a perceptual system we are able to feel, hear and smell objects in our environment. Such a propensity allows us to attribute psychological states to other people by seeing their actions, or hearing them speak. When I perceive a person jumping for joy, for example, I infer that they are happy; just as when I perceive a person wincing, I infer that they are in pain. As my own actions, or my own utterances, fall under the purview of things that I can perceive, they too can be interpreted in such a way. That I can interpret my own behaviour is, to my mind, merely an unremarkable consequence of having a perceptual system. That it is possible for me to *sometimes* achieve self-knowledge of my own mind by inference, doesn't show that introspection is superfluous.

To further expand upon this idea, and show why there is nothing superfluous about the dual-method theory, let us consider what can occur with sensory mental states. Consider my belief that I am in pain, for example. Typically, I know that I am in pain by introspecting a pain sensation. But, as we

acknowledged in the last chapter, I can also form a true belief that I am in pain by looking at a neuroimaging machine, or by seeing my own grimacing face in a mirror. Now while I can come to know that I am in pain in such an inferential way, this is not the usual way that I come to have knowledge of the fact that I am in pain. Moreover, such a method is not as likely to result in knowledge of the fact I am in pain, compared to introspecting that I am in pain. So, while it is true that I can come to believe that I am in pain by observing a neuroimaging machine (an inferential process), I do not think it is right to say that such an inferential process should be thought of as a backup system (as Goldman's idea suggests). Neither do I think that looking at a neuroimaging machine provides me with any additional information that I can use to form the belief that I am in pain (as Rey's idea suggests). If anything, observing a neuroimaging machine, for the purposes of learning that I am in pain, would only provide an obstacle to self-knowledge, because it would add unnecessary complexity to the self-ascriptions process.

Our current discussion has shown that there is nothing superfluous about positing the dual-method theory with respect to one's own *sensory mental states*. I think that almost everyone would agree with this. If this is the case, then I think the same should be true with respect to our own propositional attitude self-ascriptions. To my mind, since Carruthers would have no problem with accepting a dual-method theory for sensory mental states, he should have no problem with accepting a dual-method theory for propositional attitudes.

To see why this is so, consider the fact that I can learn that I have an intention to leave the cinema by looking at myself pick up my coat, and get up from my chair. As I said above, in our discussion about sensory mental states, the

reason I can self-ascribe a mental state on the basis of my behaviour (in this case by looking at myself grab my coat), is simply a consequence of the fact that I possess a perceptual system. Now to go further, and say that this perceptual system can do the same job as introspection, is to beg the question. It assumes that an inferential process would be able perform the same task as an introspective one. This is clearly a further claim, and cannot be assumed at the outset. The ontological parsimony argument only has force with this assumption—which is obviously highly controversial. Since I think that Carruthers would accept the dual-method theory is true, with respect to sensory mental states, I think he ought to accept that the dual-method theory is true, with respect to propositional attitudes.

It is true that simplicity may be a theoretical virtue (all things being equal), as many philosophers have pointed out (see, e.g., Quine 1966; Lewis 1973), but it isn't, itself, the only desideratum worth considering.<sup>31</sup> Parsimony is, after all, not merely simplicity. Only if two theories can adequately explain the same data should the simple one be preferred. What I have argued for here is that the single-method theory cannot explain the data equally as well as the dual-method theory. One main reason to think this is because it seems hard to deny that we can *sometimes* introspect our non-perceptual propositional attitudes. This would require us to posit two distinct methods—an introspective one, in addition to an inferential one.

While the dual-method theory is indeed more complex than a single-method theory, I have argued that the dual-method theory is not needlessly complex. There are good reasons why we should accept it. Indeed, Carruthers

---

<sup>31</sup> As Albert Einstein ([1950]1954, p. 349) famously pointed out: the term 'simple' is not just vague, but there is also nothing to warrant the assumption that a theory which is simple is also true.

himself would accept that the dual-method theory is true for sensory self-knowledge, so the question is ‘Why doesn’t he accept it for propositional attitudes?’ I’ve argued that he ought to. The ontological parsimony argument against the dual-method theory is, therefore, unconvincing.

I have now dealt with the first confabulation argument—the ontological parsimony argument. I am yet, however, to address the second confabulation argument—namely, the pattering argument. According to this argument, the patterning of the various confabulation data is best explained by positing the ISA theory. In fairness to this argument, we will need to examine the confabulation data, before our treatment of the ISA theory is complete.

### **3.7 Conclusion**

This chapter has examined two challenges to introspection: content externalism and inferentialism. First, I argued that content externalism is compatible with introspective self-knowledge of propositional attitudes, and thus first-person authority. Next I examined a particular version of inferentialism: Peter Carruthers’ ISA theory. After explaining the view, I went on to raise two objections to it—namely, the accuracy objection and the sensory data objection. I claimed that while these two objections are themselves not decisive, they do show that if one is to accept the ISA theory, one would have to accept two highly counterintuitive theses. I then turned to the ontological parsimony argument for the ISA theory. I argued that it fails.

We are not, however, done with the ISA theory. We still have the second confabulation argument, the pattering argument—considered, by Carruthers, to be the main argument for the ISA theory—to discuss. If *this* argument is

successful, then there may be strong grounds for accepting the ISA theory, even if it does involve accepting counterintuitive theses. It is necessary, then, that we address this argument before leaving the ISA theory behind. It is this task that I shall focus upon in the next chapter.

## Chapter 4

### Evidence and Error

If philosophy indeed is autonomous to the point that it can be practiced in indifference to the actual world, then it plainly follows that if an analysis, presented as philosophical, in fact goes shipwreck against empirical truth, it was not philosophy to begin with.  
C.L. Hardin (1988, p. x)

The upshot of all these discoveries is deeply significant, not just for philosophy, but for us as human beings: There is no first-person point of view.  
Alex Rosenberg (2016)

At the end of the chapter 3, I said there were good grounds for accepting the dual-method theory. Recall that this is the view that we possess two distinct ways of achieving knowledge of our own propositional attitudes: an introspective process and an inferential (or interpretative) process. This view is a challenge to the Interpretive Sensory-Access (ISA) theory, which is committed to the single-method theory: the view that we *only* have inferential (or interpretative) access to most of our propositional attitudes.

In this chapter, I will continue my critique of the ISA theory by considering the second of the confabulation arguments that I introduced in the previous chapter: the patterning argument. According to this argument, even if one accepts the dual-method theory—and can explain the *existence* of the confabulation data—one will still face the further challenge of explaining the *patterning* of the confabulation data. In order to respond to this challenge, it will be necessary to examine the experimental data that Carruthers (2009, 2010, 2011) cites to support the patterning argument.<sup>1</sup> Such experimental data come from cases where subjects have confabulated—which to say that they have

---

<sup>1</sup> Scaife (2014) and Rosenberg (2016) cite similar empirical data to also motivate scepticism with respect to introspection.

mistakenly self-attributed a psychological state to themselves. The fact that the ISA theory predicts such data empirically distinguishes it, in Carruthers view, from ‘almost all other theories of self-knowledge’ (2011, p. 6). Because I agree with Carruthers on this specific point, I will pay close attention to these experimental results.<sup>2</sup>

In what follows, I will argue that even though the ISA theory is indeed unique in making such a prediction about the confabulation data, dual-method theorists still have the resources to account for such data. Furthermore, and more importantly, I will argue that the specific patterning of errors that Carruthers thinks supports the ISA theory, is predicated upon questionable interpretations of the experimental results. I will support this contention by challenging *some* of the interpretations that Carruthers gives of these experimental results, and, where such disagreement occurs, offer a competing interpretation. The central aim in this chapter is, as Robin Scaife puts it, to take the confabulation data ‘seriously’ (2014, p. 470).

This chapter will proceed as follows. I begin, in §4.1, by clarifying the concept of confabulation. I then offer three different explanations for why subjects may have confabulated a propositional attitude. These include: errors of self-deception, brute errors, and basing errors. Such explanations, which I claim are compatible with the dual-method theory, will provide us with the framework needed for understanding how one can explain the confabulation data, without positing the ISA theory. In §4.2, I examine several notable examples of the confabulation data, which Carruthers cites as evidence for the ISA theory. I argue that some of the interpretations of these data that Carruthers provides are

---

<sup>2</sup> As I mentioned in the last chapter, Carruthers does not hold the view that we are always confabulating.

questionable. In §4.3, I look at what patterns emerge from the confabulation data, and conclude, pessimistically, that no noteworthy pattern emerges—at least none that would support the ISA theory.<sup>3</sup>

#### **4.1 Introspection, Confabulation, and the Varieties of Error**

The claim that we sometimes make mistakes in our propositional attitude self-attributions is not, in itself, controversial. No contemporary philosopher holds the view that human beings possess the ability to achieve infallible self-knowledge of his or her beliefs, intentions, desires, and so on. What *is* a matter of some contention is *why* such errors occur, and what, if any, consequence this phenomenon has for our interest in developing an account of introspection and first-person authority.

This chapter will focus on a specific set of circumstances in which subjects make errors in their psychological self-attributions—namely, the experimental results that I have been referring to as *the confabulation data*. The aim here will be to review some of the more notable examples of these data, and then contest Carruthers' claim that the set of data is best explained by positing the ISA theory. Before we examine some examples of these data, in §4.2, it is necessary that we first look at several alternative explanations for why subjects may have confabulated—explanations which differ from the one offered by the ISA theory. Once we have examined a series of these different explanations, we will *then* be in a position to assess the claim that the ISA theory offers the most plausible explanation for why it is that subjects confabulate in experimental settings. Three such explanations will be examined here. In §4.1.1, I look at

---

<sup>3</sup> Pessimistically for the ISA theorist.

errors of self-deception; in §4.1.2, I will look at brute errors; and in §4.1.3, I look at basing errors. These three different explanations should not be understood as rival explanations, rather they should be understood as different possible explanations for why a subject has made a mistake in her self-attribution.<sup>4</sup> Neither should this triad of explanations be seen as exhaustive: I concede that other explanations could be discerned.

Before we get to these individual explanations, some additional stage setting is first required. Specifically, I need to say a bit more about what is meant by the term ‘confabulation’. Recall that in the last chapter, I gave quite a broad description of it. I said that a subject who has confabulated is one who has sincerely, yet falsely, avowed that she is in a certain mental state—e.g., a perception, a belief, an intention, and so on. Given that this description is quite brief, and given that there is some controversy in the literature about how the term ‘confabulation’ should be defined (see, e.g., Hirstein 2005, 2009; Deluca 2009), I will now go into some more detail.

There are two questions that warrant further attention here. One is ‘What types of false self-avowals count as confabulation?’, and the second is ‘What is the scope of confabulation?’ I will address the former question first. The reason that this question is important is because we do not want to classify all *false avowals* as confabulations. For example, the teenager who lies to his parents when he says that he believes he has done his homework, when he actually believes he has not, speaks falsely about what he currently believes about the status of his homework. This is not an example of confabulation, because the teenager is not ignorant of his belief, rather he is trying to mislead his parents.

---

<sup>4</sup> The three different possible explanations are not all rival explanations, because some turn out to be compatible with each other. Errors of self-deception, for example, can be thought of a very specific set of basing errors.

Similarly, slips of the tongue may be false avowals, but they are not examples of confabulation. For example, if Jill says ‘I believe Osama is the former president of the United States’, when she really meant to say ‘Obama’, we do not take Jill to have *mistakenly* self-ascribed the belief that ‘Osama was the former president of the USA’. She has simply misspoken.

What we need, then, is a more detailed account of what it means to say that someone has confabulated. To fulfil this task, I will draw upon a set of criteria that William Hirstein has recently developed, which he calls the ‘epistemic definition’ (2009, p. 5) of confabulation. I will adopt and expand upon this definition in what follows. Although I am largely in agreement with Hirstein about the following set of criteria, there is an important epistemic thesis that he does not consider, which I will suggest should be integrated into our conception of confabulation. Before stating what this entails, let us first consider the conditions that Hirstein lists. According to Hirstein, *S* would confabulate that *P* if and only if:

- (1) *S* claims that *P* (e.g., Sam claims that he believes that Canberra is the capital city of Australia).
- (2) *S* believes that *P* (e.g., it is true that Sam believes that he believes that Canberra is the capital city of Australia).
- (3) *S*’s thought that *P* is ill-grounded (e.g., Sam’s belief that he believes that Canberra is the Capital city of Australia is ill-grounded).
- (4) *S* does not know that his thought that *P* is ill-grounded (e.g., Sam’s is unaware that his belief that he believes Canberra is the Capital city of Australia is ill-grounded).

(5) *S* should know that his thought that *P* is ill-grounded (e.g., Sam's is unaware that his belief that he believes that Canberra is the Capital city of Australia is ill-grounded.).

(6) *S* is confident that *P* (e.g., Sam is confident in his belief that he believes that he believes that Canberra is the capital city of Australia).<sup>5</sup>

In addition to these six conditions, there is a seventh criterion that I think should also be part of our conception of confabulation. This is the requirement that the subject's avowal is false. We can summarise this in the following:

(7) *S*'s belief that *P* is false (e.g., Sam's belief that he believes that Canberra is the capital city of Australia is false. As a matter of fact, it is not true that Sam believes that Canberra is the capital of Australia.)

Before giving reasons for this amendment, I will first elaborate upon each criterion. (1) captures the idea that confabulation needs to be discernible from the third-person point of view. Since a subject must be unaware that he is confabulating, the subject must make his thought publically known, as in the form of an avowal. Such avowals, it is worth noting, need not necessarily be vocalised. As Hirstein (2009, p. 5) points out, a subject could simply point to a sign or write something down. (2) requires that the subject believe what he avows. In cases of lies, or slips of the tongue, as discussed above, subjects do not actually believe their avowals to be true—hence, they do not count as confabulating. (3) indicates the reason why the subject confabulated in the first

---

<sup>5</sup> This list is adapted from Hirstein (2009, p.5).

place. The subject has come to form their belief in a way that is ill-grounded. And (4) describes the subject's ignorance of the fact that (3). This is an important point, because the subject must believe that there is nothing out of the ordinary about his self-ascription. (5) is a normative claim. The subject *should*, under normal circumstances, have been able to correctly self-ascribe. (6) refers to the subject's confidence about his own self-ascription. Because the subject is unaware that his self-ascription is ill-grounded, he will not doubt that what he says is true.<sup>6</sup>

While I think that Hirstein's list of six criteria provides us with a plausible construal of confabulation, his list is incomplete. We must also add (7)—the requirement that the subject's belief that *P* is false. This is because without a subject making a false self-attribution, it will not be possible to identify that self-avowal as a confabulation from the third-person point of view. On Hirstein's view, one could be confabulating and have a true belief about his own psychology. This is something that we should not say.

To see why this is a problem, let us consider the example we used above. Suppose that Sam avows that Canberra is the capital city of Australia, and let us imagine that all items from Hirstein's list obtain. We can suppose that Sam believes that he believes that Canberra is the capital city of Australia; we can suppose that Sam's belief is ill-grounded—imagine that Sam believes that he believes that Canberra is the capital because of the fact that it's raining outside; we can suppose that Sam does not know that his belief is ill-grounded, yet should know that it is; and finally, let us also add the assumption that Sam's avowal is true: Sam really does believe that Canberra is the capital city of Australia. Such a

---

<sup>6</sup> One might object that such a condition cannot be distinguished from (2), since to be confident about *P* is just to believe *P*. The sense in which I think confidence should be understood here is the sense in which one has a high measure of confidence in the belief.

scenario is possible on Hirstein's construal, and yet it seems implausible to suppose that his case counts as confabulation.

The main reason that I claim this is because I think it is hard to see why anyone would be tempted to say that Sam is confabulating in such a case. In such a case, someone paying attention to Sam's behaviour would have no reason to suspect that what he avows is false, and would, thus, have no reason to suspect that he is confabulating in the first place. The addition of (7), which says that the subject's avowal that *P* is false, fixes this lacuna, which ultimately gives us a more precise characterisation of confabulation.

The second question about confabulation that requires further attention pertains to scope—that is, 'What types of mental states, or facts about one's own psychology, are we capable of confabulating?' According to Hirstein (2009), there are two schools of thought on this matter. One is that the term 'confabulation' should be confined to cases of memory, as the term was originally implemented. The second school of thought suggests that the term should be applied to other types of psychological self-ascriptions, such as beliefs, emotions, motives, intentions, and so on. In this chapter, I will adopt the second school of thought on this matter. If a certain self-avowal meets the criteria offered above, then I think it should be considered as an example of confabulation.

With these terminological, and conceptual, concerns addressed, we can now look at the different ways in which subjects might confabulate a propositional attitude. Although our interest here is with propositional attitudes, it will, for purposes of clarity, be helpful to mention other types of mental states as we proceed.

### 4.1.1 Self-Deception

Why would someone who has the ability to introspect her own propositional attitudes sometimes confabulate? This section considers one explanation: because of self-deception (or in other words, in cases where a person has been self-deceived). Before showing how self-deception is related to confabulation, it is important that some preliminary clarifications are made, as the term ‘self-deception’, within philosophical parlance, is not without its controversy. For example, in his entry on self-deception in *the Stanford Encyclopedia of Philosophy*, Ian Deweese-Boyd writes ‘[v]irtually every aspect of self-deception, including its definition and paradigmatic cases, is a matter of controversy among philosophers’ (2016). Although this section will not enter this controversy in any great detail, it is necessary, given this concern, that we are careful to clarify what is meant by self-deception before we look at its connection to confabulation.

A good place to commence this discussion will be to look at where some general agreement lies. Deweese-Boyd notes that most philosophers would accept that:

self-deception involves a person who seems to acquire and maintain some false belief in the teeth of evidence to the contrary as a consequence of some motivation, and who may display behavior suggesting some awareness of the truth (2016).

Some examples will help elucidate this. Consider, first, a survey of university professors which found that 94% thought they were better at their jobs than their average colleague.<sup>7</sup> Here it is plausible to suppose that the professors who participated in the survey let their own emotions and desires unduly influence their own beliefs about how talented they were. A second example can be given

---

<sup>7</sup> This example is cited in Mele (2001, p. 1).

by considering a scenario where a husband continues to believe that his wife is being faithful to him, even though his wife behaves in ways that would raise suspicion in others (e.g., she disappears for hours at a time, and has drastically changed her behaviour). Because the husband does not want it to be the case that his wife is being unfaithful, he may fail to adequately give weight to such considerations when he forms and maintains his belief.<sup>8</sup>

Now while such examples are ubiquitous, and rather unremarkable—though in my view good candidates for paradigmatic cases of self-deception—the actual beliefs and desires expressed by the subjects in such cases need not necessarily be cases of confabulation because they need not involve occurrences where subjects avow a *false belief*.<sup>9</sup> While such beliefs may lack strong justification, or be ill-founded, because they are unduly influenced by the subject’s own desires, they are not, at least some of the time, speaking falsely. Such mental states, then, fail to meet one of the conditions for confabulation that was given above: namely criterion (7), that subjects are mistaken in their self-attribution of belief.<sup>10</sup> The same is true with some examples that are found in the confabulation literature. One notable case that is discussed in the literature is a condition called Capgras Delusion (see, e.g., Hirstein 2009; Mele 2009a)—a condition where patients believe that their family members are in fact imposters, who look identical. Although there is some debate involving whether self-deception is involved here (see Mele 2009a), or what exactly the underlying causes of this condition are

---

<sup>8</sup> Such a case is discussed by Mele (2001) and Hirstein (2009). The fact that Hirstein considers such a case to be an example of confabulation shows how our conceptions of confabulation differ. I do not consider this to be a case of confabulation.

<sup>9</sup> The point here is that not all cases of self-deception will be episodes of confabulation. Recall what was said above—namely, that confabulation requires that a subject mistakenly self-attributes a particular psychological state.

<sup>10</sup> It is important not to confuse the truth or falsity of an attribution of the belief that *P*, with the truth or falsity of the belief that *P* itself. Someone may be right in thinking that they believe that *P*, and yet it could turn out that their belief that *P* is *false*.

(see Hirstein 2005), I do not think that such a case is best described as an example of confabulation, as there is no reason to doubt what the subjects believe about their own psychology is false. In other words—I do not think there is any reason to doubt that Capras Delusion patients actually do believe that a member of their family has been replaced by imposter. If I am right here, then this conflicts with (7)—the requirement that what the person says false. A typical Capras Delusion patient may believe that his family members are imposters, but he is not wrong about his own psychology. While it is true that his family members are not imposters, and thus, it is true that the patient's belief is false, this is not what is pertinent to confabulation. There are, after all, many false beliefs that we have. To call them examples of confabulation would be to push the concept of confabulation further than it ought to be pushed.

Another notable case discussed in the literature (see, e.g., Hirstein 2009; Mele 2009a; Smithies 2013) that does, in my view, better fit the description of confabulation—and seems to involve some self-deception—is Anton's syndrome: a condition where patients who are blind, report being able to see. These patients report, confidently, that they are perceiving objects in front of them even though they are blind, and so could not be. For example, they walk into objects, all the while maintaining that they can see. Self-deception might be one explanation for why subjects in such cases make such an error. The subject sincerely believes she is having a certain visual perception (e.g. of a table) even though such a belief is false and ill-grounded. We may imagine a sufferer of this condition coming to believe that she is experiencing a sensation of a red chair, because she fails to accept that she is blind. In this case, the subject confabulates. She is not actually have this experience, but she believes she is.

Consider another example involving the (real-life) case of Neil Harbisson, a man who can only see black and white. In order to help overcome this shortcoming, Harbisson wears an external electronic eye, which can pick up colour frequencies through a camera that is able to transform them into sound vibrations. This allows him to identify colour, by identifying sound vibrations. Let us imagine that because of his desire to experience colour, that after a while he comes to believe that he is actually experiencing the real colours. He might say, for example, ‘I am experiencing red like everyone else can’. Since Harbisson cannot experience red (he can only experience sound vibrations) this would strictly speaking be false.<sup>11</sup> Here I think it would be appropriate to say that Harbisson has confabulated a perceptual experience.

We have looked at two examples where it is plausible to say that a subject, because of self-deception, has confabulated a perception. Let us now consider non-perceptual propositional attitudes—as these are the mental states that, according to the ISA theory, we cannot introspect. Let us imagine a case where a juror has confabulated a decision that *P*. Let us suppose that the juror sincerely says, ‘I have not decided the guilt of the defendant yet’, when in fact the juror has decided that the defendant is guilty.<sup>12</sup> One way in which this scenario may be said to occur is because the juror falsely believes that he has not decided the verdict yet, because he thinks of himself as a fair person, and he also thinks that no fair person would decide the guilt of a defendant without first hearing the case. He has been self-deceived. This would count as confabulation, according to

---

<sup>11</sup> This depends on what is meant by ‘seeing red’, of course. It may be argued that Harbisson is simply labelling his experience as ‘red’, all the while acknowledging that he is not perceiving any colour. In stating that it is possible that Harbisson does confabulate a colour experience of red, I imagine something like the following. Let us suppose that Harbisson, after matching certain black and white perceptions with various vibrations over time begins to form beliefs that he is seeing colour—in the same way that people with colour vision can. Since Harbisson can only see black and white, such perceptual beliefs would be confabulated.

<sup>12</sup> This juror example has been drawn from Bilgrami (2006, p. 152).

my construal of confabulation above, because the juror sincerely avows ‘I have not decided the guilt of the defendant yet’, when in actual fact he has.

Self-deception provides us with one explanation for why someone who has the ability to introspect her own propositional attitudes may have confabulated.

#### **4.1.2 Brute Errors**

The second explanation I consider for why someone who has the ability to introspect her own propositional attitudes might have confabulated is because a *brute error* has occurred.<sup>13</sup> The concept of a brute error is most commonly associated with the act of perception, so it will be useful to introduce this piece of terminology by first explaining its use with reference to perception.

Consider, for example, the following from Anthony Brueckner who says that brute errors occur when ‘the perceptual system is functioning properly but the world fails to cooperate’ (2011, p. 174). One example he gives to illustrate this idea is the classic example of an infant being fooled by a bent-looking stick in the water. In this example, we are to imagine that a child forms a false belief about the actual nature of a stick—namely that it has the property of ‘being bent’. This is because the child is ignorant of the effect that the refraction of light has upon certain objects. Unlike an adult, who is aware of this phenomenon, the child makes an error with respect to the stick’s nature.

This example is useful in illustrating two key ideas about perception: (i) that perceptual awareness is *caused* by independently existing objects in the world; and (ii) that perception is representational. These two features give rise to a third feature—namely, (iii) that misrepresentations are possible. If the visual

---

<sup>13</sup> Here, I follow Smithies (2013) and Burge (1988) with this terminology.

system misrepresents my environment, and I hallucinate that there is a tomato on my desk, when there is none, then we can say that a brute error has occurred because the perceptual system is *not* working properly.

Brueckner's characterisation above makes it seem like a subject would typically be responsible for such an error, but this will not be the case in scenarios where the visual system is malfunctioning. This is a point stressed by Tyler Burge, when he describes brute errors as errors that 'do not result from any sort of carelessness, malfunction, or irrationality on our part' (1988, p. 657). This would be exemplified in cases such as illusions and hallucinations, when a subject's perceptual system misrepresents the subject's environment. What is important in Burge's description, that is absent in Brueckner's, is that the subject is playing a passive role in the error. This idea is further developed by Declan Smithies who characterises brute errors as 'justified false beliefs that are properly based on justifying evidence' (2013, p. 1180). Like Burge, Smithies highlights the passivity of perception. If I form the false belief that there is a tomato in front of me because I am hallucinating, the fault lies with my perceptual system, not with me.<sup>14</sup>

Given the fact that some philosophers (see, e.g., Armstrong 1968; Lycan 2004) have defended perceptually inspired views of introspection—referred to in the literature as 'quasi-perceptual' and 'inner sense' views—some have thought that brute errors are possible with respect to the introspection of propositional attitudes. Just as a subject's perceptual mechanism may misrepresent a visual object in the world, as described above, a subject's introspective mechanism may also misrepresent a propositional attitude. A subject may, as Dorit Bar-On puts it,

---

<sup>14</sup> While it is true that my perceptual system is a part of me, it is not the case that I am responsible for the accuracy of my perception system. This is not to say that one cannot *influence* the accuracy of one's perceptual system, e.g., by taking hallucinogenic drugs.

‘mistake one state for another’ (2004, p. 201). For example, just as Sally may make a brute error when she hallucinates a tomato, when there is none, she may also make a brute error when she believes that she intends to see the film *Star Wars* when she does not. Naturally, as the quasi-perceptual model of introspection is controversial, so too is the notion of brute errors with respect to propositional attitudes. Some philosophers (see, e.g. Smithies 2013), accordingly, reject the idea that brute errors of the sort outlined above are possible.

In order to give a better idea of this approach, let us look at one version of the quasi-perpetual (representational) approach to self-knowledge—namely, the higher order theory of consciousness. Although there are different versions of this theory, they are united, as Ned Block says, by the following biconditional ‘[a] mental state is conscious if and only if the state is the object of a certain kind of representation arrived at non-inferentially’ (2011, p. 421). One proponent of the higher order theory, Gregg Caruso (2012)—who defends David Rosenthal’s (1997) higher-order thought theory of consciousness—thinks that the theory is well-suited to explain instances of confabulation. He says, ‘since we are dealing with a representational relation between two states [first- and second-order states], the possibility of misrepresentation always exists’ (2012, p. 160). Caruso claims that such misrepresentations are actually beneficial to subjects at particular times. An example he gives here is the cocktail party effect: the experience people have when they are able to block out the background noise in a group of large people and focus on their current conversation. According to Caruso, the higher order thoughts misrepresent these other conversations as ‘indistinguishable chatter’ (2012, p. 161).

As well as allowing for the possibility of *misrepresentations* with respect to first-order states, Caruso acknowledges that the higher order theory also allows for the possibility that one can have a belief about being in a mental state without any target state existing at all (e.g., one can believe that one intends to  $\Phi$  even though no intention to  $\Phi$  exists to be misrepresented). Caruso contends that this is what happens when we *confabulate*, as he states in the following.

If, for example, we were to confabulate a want or desire,  $P$ , to explain a particular action  $X$ —when in reality the true cause for  $X$  was  $Q$ —according to the HOT theory, we would subjectively feel like as if we were doing  $X$  because of  $P$ ...I think it is a virtue of the theory that it allows for the possibility of misrepresentation and confabulation (2012, p. 162).

Even though the higher order theory and the notion of brute errors are both controversial in the literature, they provide us with another explanation for why a person might confabulate, even if they possessed an introspective way of accessing their own mental states.<sup>15</sup>

### 4.1.3 Basing Errors

The third explanation I offer for why someone may have confabulated is because they have made a *basing error*. Unlike brute errors, which, as I said above, are controversial amongst those who oppose quasi-perceptual theories of introspection (see, e.g., Burge 1988; Shoemaker 1996; Bar-On 2004; Hacker 2013; Smithies 2013), there is little, if any, controversy to be found with the notion of a basing error.

Following Declan Smithies (2013), I shall use the term ‘basing error’ to denote a situation where a subject forms a false belief about an occurrent mental

---

<sup>15</sup> See Block (2011), for a discussion about why the idea of non-existent target states is problematic.

state (a second-order belief about a first-order target state), because she has based, or grounded, her belief upon a non-introspective source. In order to clarify this concept, let us consider an example. Recall from our earlier discussion about first-person authority, in chapter 2, where I said it was possible for a subject to form the belief that she is in pain when she is in fact not. There, I said that we could conceive of a situation where a subject has formed this false belief by examining a scan of her brain on a neuroimaging machine. In such a case, the subject would have falsely attributed the mental state to herself because she has *based* her judgement on a non-introspective source—namely, the image of her brain on the machine, rather than her conscious experience of being in pain. Such errors, to borrow Smithies’ expression, are ‘unjustified false beliefs [subjects form about their mental states] that are not properly based on justifying evidence’ (2013, p. 1180).<sup>16</sup>

Basing errors, unlike brute errors, do not involve a situation where the introspective faculty has failed or malfunctioned; rather, basing errors involve a situation where the introspective faculty is not being relied upon at all. Although Smithies does not use the term ‘confabulation’ in his discussion of basing errors, the connection between basing errors and confabulation seems to me to be a congruous one. To support this contention, let us consider the fact that one of the examples that Smithies (2013) uses to illustrate the concept of a brute error is Anton’s Syndrome—a condition where blind patients believe that they can see. Anton’s Syndrome cases not only meet the desideratum for confabulation that we outlined above, but also feature in other authors’ discussions of paradigmatic

---

<sup>16</sup> It is only introspective evidence, on Smithies’ view, that could count as justifying evidence for a mental state self-ascription.

cases of confabulation in the literature (see, e.g., Hirstein 2009; Wheatley 2009; Block 2011).

Now, although Smithies is mainly concerned with sensory mental states, like pain and visual perception, I think that the concept of basing errors is equally well suited to explain confabulation with respect to propositional attitudes. Just as a subject's mental illness may lead a blind person to confabulate about their own perceptions, a subject can—for many different reasons—base their own beliefs about their own propositional attitudes on non-introspective sources. A subject who falsely avows that she believes that she intends to visit Washington in the spring because her close friend tells her so, makes a mistake about what she intends to do because she has *based, or grounded*, her belief about what she intends to do upon the testimony of another. The concept of basing errors provides us with another explanation for why one might confabulate, even if one possessed an introspective way of accessing one's own mental states

#### **4.2 The Patterning Argument**

Now that we have looked at several different explanations for why subjects may have confabulated a propositional attitude, we are now in a position to examine Carruthers' claim that rival theories of self-knowledge will fail to be able to explain the patterning of the experimental data—even if they can explain the existence of such data.<sup>17</sup> By 'rival theories of self-knowledge' I am, recall, referring to theories of self-knowledge that maintain that we can have both *introspective* as well as *inferential* (or interpretative) access to our propositional attitudes—views that we have categorised as dual-method theories. As I have not

---

<sup>17</sup> The preceding section supports the claim that rival theories can at least explain the existence of the confabulation data.

defended any specific view of introspection yet, the forthcoming section will only be concerned with introspection in very general terms, which is to say it will be in accordance with the theory-neutral account offered in chapter 1. I will now reintroduce the patterning argument that was given in the last chapter:

*The Patterning Argument:* the patterning of the confabulation data is best explained by positing the ISA theory. While other theories may be able to explain these data, most do not predict it, and neither will they be able to account for the patterning of the errors.

The ‘patterning’ that is being referred to here relates to the sensory component of the ISA theory, which was elaborated upon in the previous chapter. Recall that, according to Carruthers:

[s]ince the [ISA] theory claims that our only access to our thoughts and thought processes is interpretive, relying on sensory, situational, and behavioral cues, there should be frequent instances where the presence of misleading data of these sorts leads us to attribute attitudes to ourselves *mistakenly* (2011, p. 325).

One should *not* get the impression that on such a view it would be the case that we are always confabulating. Carruthers grants that we often do form true beliefs about our own minds. What he is claiming here is that *all of our self-ascriptions* will ultimately be grounded by sensory, situational, or behavioural cues. When those cues are misleading, Carruthers thinks one will confabulate.<sup>18</sup>

My plan for dealing with this argument will be as follows. I will, in the following three subsections, examine a series of recent empirical results that are

---

<sup>18</sup> Such a position is compatible with a scenario where someone always accurately self-ascribes a mental state. As a matter of fact, however, Carruthers thinks that we do often confabulate.

discussed by Carruthers. As it is often difficult to determine the correct interpretation in each case, and whether confabulation has occurred, I will not simply accept the interpretation that Carruthers gives. I will offer my own interpretation of each case. I will do so by keeping, roughly, to the following set of questions.

- (a) Does the subject make an error in the case?
- (b) If so, what explanation can be given for why the subjects made the error?

Once we have looked at a series of these cases, I will, in §4.3, ask:

- (c) Are there any significant patterns that arise from such data? That is, can the ISA theory explain the patterning or error more satisfactorily than rival theories of self-knowledge?

Before getting to the experimental results, I need to first point out some difficulties with respect to answering questions (a) and (b). First, it is not always a trivial task to identify errors in people's self-attributions, from the third-person point of view. Some cases that we will look at appear open to various interpretations, and without any universally agreed upon way of verifying whether a subject really does not believe what they claim, we are presented with the problem of how to identify a mistaken self-attribution. This thought is explained by Jonathan Schooler and Charles Schreiber in the following.

The paradox of introspection stems from the fact that personal experience corresponds to that which we know best subjectively, yet least empirically...Although subjectively incontrovertible, it is impossible to directly assess the contents of experience, and thus no decisive way to empirically determine when reported introspections accurately vs. inaccurately characterize underlying experience (2004, pp. 17–18).

Schooler and Schreiber are not, of course, suggesting that we can never have knowledge of the mental states of others from the third-person point of view—we clearly can and do. What they are saying is that there is controversy over how exactly we can know that another person is wrong in her self-report. Eric Schwitzgebel raises a similar point in the following.

A key challenge in assessing the accuracy of people's beliefs or judgments about their attitudes is the difficulty of accurately measuring attitudes independently of self-report. There is at present no tractable measure of attitude that is generally seen by philosophers as overriding individuals' own reports about their attitudes (2014).

Since identifying confabulation requires that we do attempt to measure the accuracy of a subject's self-reports, this is an issue we need to keep in mind as we proceed.

The second issue to keep in mind pertains to the use of manipulation techniques in certain experiments. Since some experiments are designed to trick or mislead subjects, it is questionable as to what extent the subjects' responses should be seen as representative of what they actually believe. I said above that in order for confabulation to have occurred, a subject must not only avow a false self-report, but they must also believe what they say is true. To see why this might be an issue, let us consider the following example given by Daniel Kahneman (2009) in his book *Thinking Fast and Slow* called the *Moses Illusion*.

Kahneman describes an experiment where participants were asked the question ‘How many animals of each kind did Moses take onto the ark?’ It was observed that most people who answered this question replied with an answer of ‘two’. Now, since in the actual biblical story it was Noah who took the animals onto the ark, the participants should have said something like ‘Moses didn’t take any animals on’. As we are aware of this manipulation, we do not attribute the belief ‘Moses brought two animals of every kind onto the ark’ to these subjects, because we know the experimenters deliberately mislead the participants. We would not classify this case as one of confabulation. I will suggest that with some cases of purported confabulation, a similar concern is present.

I will now examine several examples of the confabulation data. Since space prevents us from looking at all the cases that have been discussed in recent literature, I will focus on some of the more notable cases.<sup>19</sup>

#### **4.2.1 Choice and Confabulation**

The first set of experimental data comes from experiments Johansson et al. (2005) performed, which have since been replicated using different materials by Hall et al. (2010). In these experiments, male subjects were presented with picture pairs of female faces, and were then asked to choose which face in each pair they found most attractive, before being asked to place the pictures down. After a few seconds, the pictures were turned up and the subjects were asked to give reasons for why they had chosen the face that they did. In some trials, the experimenters manipulated the experiment by quickly switching the picture that was turned down with a different picture, which was, in some cases, quite

---

<sup>19</sup> All of the examples that I discuss here are ones that Carruthers cites to support the ISA theory.

dissimilar from the original one they selected. Other times, the experiments switched the originally selected picture with one that had been rejected by the participant. Interestingly, not only did a small percentage—about 30%—of the participants fail to notice the change, but some provided reasons for why they had selected the switched picture. The responses that were given by participants who failed to notice the change, and whose picture were switched, seemed to suggest that subjects had little knowledge of the reasons why they had chosen the picture that they did. For instance, subjects gave answers such as: ‘I chose her because she had dark hair’ (Johansson et al. 2005, p. 118), in a situation where the subject had originally chosen a picture with a woman with blonde hair.

According to some authors (see, e.g., Johansson et al. 2005; Carruthers 2011; Scaife 2014), the answers that these subjects provide are confabulated, because they could not possibly refer to their original selection. In their view, avowals such as ‘I chose picture A because the women in the picture had dark hair’ is confabulated. Before we look at the implications of such a claim, I think it is worth considering the possibility that confabulation is not occurring here. I do not wish to enter into a verbal dispute with such authors here over what is meant by the term ‘confabulation’; rather, I want to consider the possibility that subjects did not really make a mistake in the way these authors describe.

One reason for thinking that the subjects *do not* speak falsely when they give reasons for choosing the picture that they did, is because they are not giving reasons for why they liked the original picture. It is possible that when the subjects are asked questions such as ‘Why did you choose this picture?’ and ‘Tell me what you find most attractive about the picture you chose’ they are forming judgements about the picture that they are currently perceiving. Because it is

unlikely that the subjects have thought carefully about the reasons why they chose the original one that they did, they may reason as follows: ‘If I picked this photo, there must have been something I liked about it’, and then proceed to list features about the current picture.

An opponent of this ‘no confabulation’ interpretation may object at this point that this is an interesting interpretation, but one that is without empirical support.<sup>20</sup> In response to this concern, I will mention some evidence that I think is relevant here that Goldman (2006, p. 234) has recently discussed. According to experiments performed by Simons and Rensink (2005), visual representations of an earlier display can decay or be overridden very quickly. Their article suggests that detailed representations might only last for 0.5 seconds, which is, as Goldman points out, less than the time it takes for the experimenters in this experiment above to switch the picture. When we consider the fact that subjects are under the impression that the picture is the same, in addition to the fact that they may not be able to recall the precise details, I think this ‘no confabulation’ response is one that should be taken seriously. If this response is to be taken seriously, then the picture experiment does not support the patterning argument. The evidence, as I’ve presented it, weighs more in favour of the no confabulation interpretation, although it doesn’t rule out the confabulation reading.

As the confabulation interpretation cannot be ruled out, we still need to consider the possibility that confabulation is occurring; and consider whether the patterning of such confabulation is of the kind predicted by the ISA theory. Before doing so, an important question needs to be addressed. It is this: ‘If subjects are confabulating in this set of data, then what type of psychological

---

<sup>20</sup> The charge could equally apply to the confabulation interpretation, but let us leave that worry to one side.

state are the subjects confabulating?’ Since I agree with Matthew Boyle that we should reject the uniformity assumption, the view ‘that a satisfactory account of our self-knowledge should be fundamentally uniform’ (2009, p .232), we need to be careful to identify what type of psychological state is under investigation here.<sup>21</sup> To clarify this thought: even if the subjects in these experiments are confabulating, and the errors made are best explained by positing the ISA theory, it will not count in favour of the patterning argument unless a false self-ascription is occurring with respect to a propositional attitude. Recall that, it is introspection for propositional attitudes that is supposed to be brought into question by such data. The data should show that a false propositional attitude self-ascription has occurred because of misleading sensory, behavioural, or situation cues.

How does Carruthers make the case for this explanation then? In his view, the results of this study are ‘rather remarkable’ (2011, p. 147). He says, ‘subjects plainly had no awareness of what it was about the original photographs that had induced liking’ (2011, p. 148). Even if we accept this claim, though, how is the pattern of error more coherently explicated by positing the ISA theory, rather than some other explanation for why the subject confabulated? Carruthers offers the following explanation:

while subjects viewing the photographs have perceptual access to the represented face...and while they have introspective access to their own affective reaction of liking or disliking, they have no access to the specific properties of the face that give rise to their reaction (2011, p. 148).

On Carruthers view, because of the misleading perceptual cues presented to the subjects, the subjects came to form false beliefs about their own psychology—

---

<sup>21</sup> I consider this claim in more detail in chapter 7.

such as, ‘I chose picture A because the women in the picture had dark hair.’ Even assuming this interpretation is correct, however, there are still problems with thinking that this counts as evidence for the ISA theory. In my view, the main problem with the data is that it is unclear what propositional attitude Carruthers thinks that the subjects are confabulating. Even supposing that the subjects are wrong when they say, ‘I chose picture A because the women in the picture had dark hair’, it is not obvious that the subjects are confabulating a propositional attitude. That is, it is not at all obvious that such false-ascriptions provide us with examples of a belief, desire, intention, and so on—exemplars of the set of propositional attitudes. It is more probable, in my view, that the subjects are reporting on what properties gave rise to, or caused them to have, the attitude they ended up having. Since the ISA theory entails the claim that the confabulation data support the thesis that we only have inferential access to our propositional attitudes, such data are not directly relevant to the thesis that we cannot introspect our propositional attitudes.

The second study that we will examine is Richard Nisbett and Timothy Wilson’s (1977) clothing experiment, which has been one of most widely discussed cases of supposed confabulation in recent years.<sup>22</sup> In this experiment, participants were told to examine a set of four identical nylon stockings that were placed upon a table, and then select the one that they preferred. The fact that the items were identical to each other was not stated by the experimenters. After the participants had selected their preferred option, they were prompted to give answers for why they had selected the item they did. They gave answers such as ‘I like this one because it is the softest’ and ‘I like this one because of the

---

<sup>22</sup> According to Petitmengin et al. (2013), Nisbett and Wilson’s paper has had almost 7000 citations (as of January 2013).

colour'. Given that the items were all the same, these results were surprising; leading the experimenters to question our ability to have authoritative knowledge of our own decision-making processes from the first-person point of view.

The experimenters originally postulated that '[p]eople have little or no introspective access to higher order cognitive processes' (1977, p. 231). If propositional attitudes are to be included in this set (higher order cognitive processes), then such results may be directly relevant to the patterning argument.<sup>23</sup> Before we investigate this claim, I should state, for purposes of contextualisation, that one of the authors of this experiment, Wilson, has since qualified his original sceptical conclusion. In a recent book, *Strangers to Ourselves*, Wilson distinguishes between mental contents and mental processes and now thinks that the original conclusion was too strong. He now holds 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' (2002, p. 105), but grants that people have 'privileged access to a great deal of information about themselves, such as the content of their current thoughts' (2002, p. 105). The idea here is that all that the experiment shows is that we do not have introspective access to the causes of certain thoughts.

Now, as Goldman (2006, p. 233) points out, this makes the original conclusion far less controversial, and does not bring into doubt a subject's ability to introspect her own propositional attitudes. The interpretation of this case that I favour is one that has gained acceptance in the literature by several recent philosophers (see, e.g., Rey 2008; Levy 2014, p. 8). It is the view that the

---

<sup>23</sup> Hirstein (2009, p. 1) uses this example to show that confabulation can occur in normal everyday situations.

participants in the experiments were unaware of what was, most likely, a right-hand side bias and, thus, have confabulated the *cause* of their own decision.

Although I agree with the causal ignorance interpretation and, thus, grant that self-knowledge of our decision-making processes may very well be limited, I still think we should not be too quick to dispute the validity of the subjects' avowals. Even though the subjects in these experiments may be unaware of the cause of their decisions, I do not think we should immediately take their claims such as 'I judge the right item to be the softest item' to be false. Because of this right-hand side bias, subjects may have given more attention to the right-hand side items and, thus, may have taken the time to notice a specific feature of the item. I will now look at two recent responses to this interpretation, which will be important to address because, if these judgements *are* confabulated, then the experiment may be relevant to the patterning argument.

First, there is Robin Scaife's claim that 'this interpretation of the results seems unlikely when we consider that participants offer a vast array of different explanations for exactly the same bias' (2014, p. 23). The worry here is that if that if all the participants are affected by the same right-hand side bias, then why would they give different answers? My response is that we *should* expect different answers because the right-hand bias may affect people differently. One subject may be predisposed to notice colour, another softness, and so on. A second objection is from Carruthers, who says of the experiment that:

subjects are *also* confabulating and attributing to themselves a *judgement* (albeit one they believe to have caused their action)—at least, if we can assume they didn't *actually* judge that the right hand-item was the softest (or a nicer color, or whatever) (2011, p. 336).

In other words, Carruthers, like Scaife, doubts that the participants speak truthfully when they state their judgements. In Carruthers' view, this means that when the subjects state 'I judge that the right hand-item was the softest', they are confabulating one of their own propositional attitudes—in this case a judgement.<sup>24</sup> Moreover, he thinks that the *reason* such confabulation occurred is because of the misleading perceptual cues presented to the subjects. Carruthers asks, '[h]ow could one claim otherwise?' (2011, p. 336).

One could claim otherwise by taking seriously the effect that the bias is having on the participants. As I mentioned above, by simply paying more attention to an object, which may be brought upon by the bias, one may be more inclined to notice a certain feature of the object. In response to this claim, Carruthers says that 'the causal pathways postulated here are mysterious' (2011, p. 336). In Carruthers view, what is actually occurring in such cases is that subjects are *unconsciously* reasoning as follows: 'I need to make a choice. But there is nothing to choose from between these items. So I might as well select this one [that I am looking at]' (2011, p. 336). While this may be correct and compatible with the ISA theory, I find this explanation to be without any evidential support, and equally mysterious. Moreover, I do not see how such an explanation would support the claim that subjects have confabulated *because* they have been misled by perceptual cues—as one would have to claim, for the patterning argument to be supported. This is not the sort of mental state an onlooker would attribute to a participant, if they were observing them select their preferred item of clothing.<sup>25</sup>

---

<sup>24</sup> Another way to express this would be: I believe that I have judged that *P*.

<sup>25</sup> Recall it is not just error that supports the patterning argument, but the way in which the subject makes the error. The misleading cues are supposed to be the sort that an onlooker would notice if they were attributing a mental state to someone.

I grant that such a debate is unlikely to be settled by such conjecturing, as I think there is not enough evidence either way. However, I do think that the burden of proof is on Carruthers and Scaife, because they are challenging the legitimacy of the first-person reports. Without any evidence for this claim, it is premature to insist that subjects have confabulated here because of misleading perceptual, behavioural, and situational cues.

We have examined two experiments that pertain to choice and confabulation. I concede that these data present a challenge to anyone advancing a first-person authoritative account of self-knowledge with respect to the finer-grained details of our own decision-making processes. However, recall that our focus here is only on propositional attitudes. Therefore, while these results may be worth considering within the greater context of self-knowledge, they do not by themselves support the patterning argument, unless we make, what I have claimed to be, unwarranted assumptions.

#### **4.2.2 Judgement and Confabulation**

In this section, I will examine sets of data that pertain to the self-knowledge of one's own judgements. I will argue that such data do not support the patterning argument for the ISA theory. I do so by showing that the data do not provide us with examples of subjects who have confabulated *because* of the presence of misleading behavioural, perceptual, and situational cues.

First, I will examine data from experiments undertaken by Wells and Petty (1980), which have since been replicated, and further expanded upon, by Briñol and Petty (2003). In the original experiments performed by Wells and Petty (1980), it was shown that by inducing subjects to nod their heads up and

down (in order to resemble saying ‘yes’), the experimenters were able to get the participants to express a greater degree of belief in a particular proposition. In contrast, it was shown that subjects who were induced to shake their heads from side to side (in order to resemble saying ‘no’) expressed reduced agreement in the same proposition. In addition to the ‘up and down’ group and the ‘side to side’ group, there was a control group who were not given any instructions. The participants performed this task under the impression that they were testing out a set of headphones (for activities such as jogging, dancing, and so on).

The experimenters supposed that the subjects’ previous experience with nodding and shaking had fostered a tendency to associate ‘nodding’ with agreement, and ‘shaking’ with disagreement. And essentially argued that head movements were biasing peoples’ thoughts in such a way that head nodding made their thoughts more positive than the head shaking (see Briñol and Petty, 2003, p. 1123).

How do these data support the patterning argument? First, we would say that the subjects confabulated—that is, they claim that they judge a proposition  $P$  to be true, when in reality they do not. Second, we would say that such a mistake arises because of misleading perceptual cues. One would say, as Carruthers does, that subjects are reasoning as follows: ‘[s]ince I have been nodding/shaking my head, this is evidence that I believe/disbelieve the propositions asserted’ (2011, p. 343). This would be exactly what is predicted by the ISA theory, because we have a case where certain behaviour is correlated, *consistently* in each scenario, with the subjects’ self-ascriptions. The movement of the heads in this experiment is what is giving rise to the subject’s false self-ascriptions. This is the same

conclusion an onlooker would come to if they were to ascribe a mental state to such subjects. Let us call this the *ISA explanation* (hereafter, ‘ISA-E’).

An opponent of ISA-E might reply that there is a second, and equally plausible, explanation. One could say that all that is happening in the nodding/shaking experiment is that the subjects are being primed.<sup>26</sup> It wouldn’t follow, then, that the subjects are not introspecting their own mental states. The subjects may not be confabulating at all.<sup>27</sup> Nodding/shaking may simply put the subject in a certain mood, which may affect the way in which they make their judgements. But they may still be able to accurately introspect their judgement, nevertheless. Their psychological self-ascription need not be the result of an inference based upon their behaviour. Let us call this explanation the ‘priming explanation’ (hereafter ‘PR-E’).

Even though the difference between ISA-E and PR-E is small, its significance within the context of this debate is large. In this particular example, a proponent of ISA-E will maintain that the source of the subjects’ self-ascription is their behaviour; while a proponent of PR-E can say that, while the behaviour may influence the subjects’ self-ascription, it need not be the source. Furthermore, if PR-E is the right interpretation here, then we need not be committed to the view that participants in these experiments are confabulating their mental states. Such data would not then count in favour of the patterning argument. Our question thus becomes: ‘What reason do we have to favour either account?’

---

<sup>26</sup> The term ‘prime’ should be understood, following Andrew Coleman, as ‘a cue given to facilitate a particular response’ (2015, p. 600).

<sup>27</sup> Confabulation alone doesn’t count in favour of the ISA theory. The way in which subjects confabulates is what is important.

One issue with attempting to resolve this question in the example above, is that both accounts appear to fit the data. In response to this specific point, Carruthers (2011, p. 343) agrees. He concedes that the data *can be* interpreted both ways, but thinks that the follow up experiment does not suffer from this problem. In the follow up experiment, Briñol and Petty (2003) not only tested for head shaking/nodding, but also modified the persuasiveness of the message itself. It was found, as in the original experiment, that when the message was persuasive, the head nodding made the subjects' judgements more favourable and the shaking made them less favourable. However, when the experimenters made the message *unpersuasive*, they found the opposite took place. Nodding then made judgements less favourable and the head shaking made them more favourable. According to the experimenters, when the message is unpersuasive, the participants interpret their head nodding behaviour as confirming their own negative thoughts; while the head shaking is interpreted as disagreeing with the message.

Carruthers claims that this is not what one would expect if PR-E were the correct explanation here. This is because subjects who were induced to nod their heads didn't always tend to agree with the proposition, as in the first experiment. When the message was unpersuasive, the head nodding had the opposite effect. Carruthers thinks this speaks in favour of the ISA theory.

In my view, it is not at all clear how this result shows ISA-E to be true. Carruthers does not state why it couldn't be that, by modifying the persuasiveness of the messages, the subjects are being primed in a way that was not taken into consideration in the first experiment. So again, we appear to have two explanations that can be made to fit the data. One where a subject has

confabulated, and the other where a subject has not. In an attempt to resolve this issue, I will now examine a second experiment, where Carruthers argues that PR-E will not work (2011, p. 344). To anticipate, I will suggest that a similar problem arises, and that differences in interpretation are unlikely to be settled by looking at further examples.

In a second set of experiments, which sought to test the same concept in a different setting, Briñol and Petty (2003) asked subjects to write either three good or three bad characteristics about themselves that might potentially affect their professional careers. The experimenters manipulated the experiment by getting half of the participants to write their responses down in their right hand (or preferred hand), and the other half to write their answers down in their left hand (or their non-preferred hand). Because writing with a non-dominant hand is infrequent, the writing sometimes appeared ‘shaky’. Following this task, the experimenters asked the subjects to rate the confidence of what they had just written. The experimenters found that the subjects who wrote their response in their non-preferred hand, and had produced ‘shakily’ written sentences, expressed less confidence in the propositions expressed, compared to those who were asked to write down their response in their preferred hand. For example, a subject might have written down ‘I am intelligent’ or ‘I care about attention to detail’. Intuitively, it doesn’t seem likely that the way in which a proposition is written down would make any difference to whether that proposition is assented to or not. The surprising result of the experiment is that this manipulation did make a difference.

In Carruthers’ view, ISA-E provides us with the best explanation of what is going on in this case. Carruthers thinks such data vindicate the ISA theory’s

core claim: by modifying the perceptual input a subject is receiving, that subject's self-ascription will change in accordance with what is being perceptually presented. Carruthers thinks this explains why changing the perceptual cues can alter the subjects' confidence in the propositions. Carruthers thinks this explanation is superior to the one offered in the priming explanation, as described in PR-E. He asks, 'why should writing sentences shakily with one's left hand induce feelings of doubt, except via a mindreading inference?' (2011, p. 344).

Before offering a response to this claim, I will first state why Carruthers thinks that these experimental data are just what the ISA theory would predict. First, he notes that the subjects are being asked to report upon a *current judgement*, which means that subjects are not being asked to report upon a memory. And neither are the subjects being asked to report upon the reasons why they made their judgement (2011, p. 344). This leads Carruthers to say

it is...hard to see why the supposed mechanism of inner sense should have failed to detect such a judgment...[i]f transparent access to judgments really exists, then the subjects in this study should have just accessed and reported on their current judgment' (2011, p. 344).

The first thing to say in response to this claim is that we cannot simply assume that subjects have made errors in these experiments: we need an independent reason to think this. If we return our focus to the handwriting case—though this point is applicable to the other examples—I do not think it is obvious that the subjects speak falsely. If PR-E is the right way to interpret the case, we could say that the subjects' judgements were being influenced by the experimenter's manipulation. By writing down something in one's left hand (or one's non-

preferred hand) one may see messy hand-writing and immediately recall previous criticisms one has received about their writing ability. And similarly, by writing down a well-written sentence one may recall instances of compliments one has received by colleagues about one's penmanship.

Once again, then, we are faced with the choice of accepting either ISA-E or PR-E as a possible explanation, and we have no additional data to differentiate between them. The fact that we have run into this problem multiple times, however, points to a bigger issue that I will call the *identification problem*. It involves the following question: 'How can we identify whether the behaviour in the so-called confabulation cases is *influencing* the subjects' judgements, or whether it is the *source* of their self-ascriptions, as the ISA theory would suggest?' Since one cannot simply assume that confabulation has occurred, one cannot simply assume that such data provide us with examples of subjects grounding their beliefs about their own mental states by perceptual cues, rather than being influenced by them. We are, thus, left without a clear way to overcome the identification problem. Given that the existence of such data is supposed to be one of the main reasons to accept the ISA theory, this is an extremely problematic result for the theory.

Before leaving the attitude judgement behind, I will consider some other empirical data that pertains to moral judgements. One study cited by Carruthers (2011, p. 140), which was undertaken by Schnall et al. (2008) sought to investigate how disgust can influence our moral judgements. The experimenters were interested in addressing the question: '[m]ight some people look to their bodily reactions for guidance more than others do' (Schnall et al. 2008, p. 1099

To answer this question, the experiments asked participants to fill in a questionnaire, which required them to make a moral judgement. The experimenters manipulated the experiment by eliciting feelings of disgust in the participants. In one experiment, a foul smell was present; in another, the participants were asked to work in a disgusting room; in another, they were asked to recall a disgusting experience; in another, they were asked to watch a disgusting video. It was observed that in all four cases disgust *increased* the severity of moral judgements. The experimenters concluded that there is a ‘causal relationship between feelings of physical disgust and moral condemnation’ (Schnall et al. 2008, p. 1105).

The data from these experiments is relevant to the patterning argument because, depending on the nature of this causal relationship, it may be that subjects were using their feelings of disgust, which they can presumably introspect, as the source of their self-ascribed judgement. This is just what would be predicted by the ISA theory. According to ISA-E, the reason that the subjects confabulated their moral judgement, in such experiments, is that they were presented with misleading perceptual cues—in this case, the feeling of disgust.

One problem with such an explanation is that it assumes that the subjects confabulated their moral preferences. This is something that we cannot take as a given. After all, it could be that by eliciting disgust in a subject, a subject is primed to react in a certain way (as I said above in PR-E). Such data, then, might only show that priming subjects in a certain way can affect the severity of their moral judgements. This is, however, not enough to show that the subjects confabulated their moral judgements. It could be that the subjects are speaking truthfully when they state their moral judgements. It just may be that the disgust

they feel has a significant influence on them. Furthermore, it may be that the subjects have also introspected their moral judgements.

Again, we seem to have an example of the identification problem. We have data that show that a sensation may have causally influenced a judgement. But we cannot be sure that the sensation, in this case disgust, was the source of the subject's self-ascription, as the ISA theory would suggest. In my view, the burden of proof is squarely on Carruthers to show why we should think that the sensation was the source. If Carruthers could establish that confabulation had occurred in such experiments—that is, the subjects falsely attributed a moral judgement to themselves—then I think Carruthers may have a point. This would show that the subjects did not introspect their judgement, because they are wrong in their self-attribution. Moreover, it would then be plausible to suppose they had made the mistake *because* of a misleading sensation. The problem for Carruthers, however, is that such a scenario cannot be shown to be the case, with respect to the above data. Thus, I think that Carruthers' claim that the above data show that the source of the subjects' self-ascriptions was a sensation, lacks supporting evidence.

The preceding cases show that under *some* circumstances self-attributions are correlated with various sensations, behaviour, or situational cues. The more controversial thesis, that the sensations, behaviour, and situational cues, are the source of subject's self-knowledge, is not fully supported by the data. Such data do not provide us with evidence for thinking that (i) we often confabulate our propositional attitudes and (ii) this is because we are presented with misleading sensory cues, behaviour, or situational cues.

### 4.2.3 Intention and Confabulation

In this final subsection, I will examine data from experiments conducted by Michael Gazzaniga (1995, 2000), which feature studies of commissurotomy (hereafter, ‘split-brain’) patients who have had their corpus callosum—the connection between the left and right hemisphere of the human brain—cut. With such patients, it is possible to expose information to only one side of the brain by giving stimulus to the corresponding visual field. If, for example, information was to be flashed to the subject’s right side, then such information would be available to the left side of the brain.

In one specific case, though these cases have been replicated (see, e.g., Gazzaniga 1995, 2000), a ‘walk!’ sign was flashed to the right brain of the subject by the experimenters (the hypothesis being that the right hemisphere controls action). Immediately after this, the subject got up and started walking. When asked why he was walking, the subject answered, ‘I’m going to get a coke’. According to the experimenters, the subject had no conscious awareness of the flashed sign, which, the experimenters suggested was the real reason the subject got up and starting walking.

This is, to my mind, the most compelling case of confabulation that we have looked at so far. It is a case where the experimenters appear to have knowledge of why the subject got up, when the subject himself does not. The subject believed that the reason he got up was that he wanted to get a coke. This self-attribution, given what we know about the experiment, seems false. The reason that the subject got up was not because of this. The real reason was that the sign was flashed. With respect to these results, Carruthers says ‘[t]his attribution of a current intention to himself is plainly confabulated, but delivered

with all of the confidence and seeming introspective obviousness as normal' (2010, p. 85). He uses this point to motivate a more general scepticism.

The split-brain data do seem to show decisively that we have no *introspective* warrant for believing that we ever have introspective access to our own judgments and decisions, however. This is because patients report plainly confabulated explanations with all of the same sense of obviousness and immediacy as normal people. It follows that subjects themselves can't tell when they are introspecting and when they are interpreting or confabulating. So for all we know, it may be that our access to our own judgments and decisions is *always* interpretative, and that we *never* have introspective access to them (2010, p. 86).

Although I think that a legitimate case can be made for thinking that such a self-attribution counts as an example of confabulation, I disagree with Carruthers that such a result supports the ISA theory. This is because the type of psychological self-ascription described in such a case is not a paradigmatic example of an intention (a propositional attitude). Although this may appear like a trivial point, it is of significance for the patterning argument that we are currently assessing.

Before expanding upon this classificatory point, I think it is worth pausing for a moment, as we have done throughout, to ask whether it is right to view this case as an example of confabulation. I said above that it seems plausible to say that it is, given that it appears to meet all the desiderata that were laid out in §4.1. I think it is a case where a subject has confidently made a mistaken psychological self-attribution—one that they should have been able to correctly self-ascribe. Not all commentators agree with this, however. Goldman (2006, p. 232), for instance, considers the possibility that the left and right hemispheres may be separate streams of consciousness, and so the attribution that the subject makes is not, strictly speaking, a confabulation, because the

psychological ascription refers to different streams of consciousness.<sup>28</sup>

Others have questioned the confidence of the subjects' self-attributions in these experiments. Brian Fiala and Shaun Nichols (2009), for example, suggest that there are some examples in the split-brain literature where subjects express low confidence in their self-reports. If this is true, it may not be right to identify such a case, as well as other cases like it, as confabulation. Recall from our discussion in §4.1: in order for a self-ascription to count as being confabulated, the subject has to be confident in their avowal. Furthermore, as Rey (2013, p. 265 fn. 9) points out, there has been no sustained attempt to determine whether the split-brain subjects do actually express any hesitancy, when they give their answers, compared to other avowals they make.

I acknowledge that these concerns are legitimate ones for the confabulation interpretation, but they are not ones that I will pursue here any further. If the split-brain data do not provide us with an example of confabulation, then such data will not contribute to the patterning argument, or motivate scepticism of the type I am concerned with addressing here. The more serious issue, in my view, is that the type of self-ascription that is described in the split-brain cases is not best described as an intention. To see why this is so, let us recall that a typical split-brain patient may respond to the question 'Why did you get up?' by saying 'I'm going to get a Coke'. With respect to such an avowal, Carruthers says, this 'current intention to himself is plainly confabulated' (2010, p. 85). In my view, it is not quite right to think of such a self-ascription as an intention. If I am right about this, then the split-brain data are only indirectly related to the patterning argument.

---

<sup>28</sup> Goldman does not press this response, however.

In order to support my claim that such an avowal is not most accurately described as an intention, let us first ask: ‘Why might one be tempted to think of this type of self-ascription as an intention?’ One reason is because one is accepting a notion of intention that is similar to the one articulated by Daniel Wegner, who says: ‘[i]ntention is normally understood as an idea of what one is going to do that appears in consciousness just before one does it’ (2002, p. 18). Understood this way, it seems that the subject’s action could certainly be described as an intention. However, as Alfred Mele (2009b) has pointed out, the above definition is not exhaustive. Mele notes that there are at least three different types of intentions that can be conceived of. First, there are intentions that involve planning, such as intending, early in the year, to attend the New Year’s Eve parade—which Mele calls *distal*. Such intentions do not need to be conscious, just before the action, in the above sense. Then there are intentions that resemble Wegner’s conception, such as the intention one has just before one turns the ignition on in one’s car—which Mele calls *proximal*. A third type, that Mele calls *mixed*, can be said to occur when, for example, a distal intention to attend the New Year’s Eve parade causes one to immediately write the date down in a phone, or book a taxi, which would be a proximal intention.

Although it may seem like the split-brain patients are confabulating a proximal intention, I think this thought should be resisted. In my view, the type of psychological self-ascription that is being confabulated here is best understood as what Donald Davidson has termed a ‘rationali[s]ation’ ([1963] 2006, p. 23). According to Davidson, rationalisations are *causal* explanations for why an agent has performed an action on a certain occasion. They can be thought of as the reason, or reasons, *why* the agent performed a certain action. Davidson suggests

that when we want to know *why* an agent performed an action, we are asking for a rationalisation, or a ‘primary reason’ ([1963] 2006, p. 25). For example, suppose someone were to ask me ‘Why did you complain about the soup?’ or ‘Why did you drive to the hardware store?’ Typical answers to these questions that I would give would be ‘*because* the soup was cold’ and ‘*because* I needed some paint’. We can think of such responses as primary reasons for why I performed the action.

Now, although such rationalisations *are* typically known to us, the thought that we are sometimes ignorant of them should not be seen as controversial. As Jaegwon Kim (2010, p. 106) describes, one may go to the refrigerator late at night, look inside, and be completely unaware of why one has done so. One may think to oneself: ‘Why am I opening the fridge?’ Perhaps the actual answer was to get cheese, perhaps it was to get milk—facts the subject may be unaware of. While this phenomenon is common enough for most of us to be familiar with it, it is not a pervasive feature of our daily lives (see Kim 2010, p. 106). As Kim correctly notes, if such a thing happened multiple times a day, we would feel extremely disconnected from our own actions. I think that the results from the split-brain studies are much like the behaviour of the person standing at the fridge. The subjects are being asked a ‘why’ question by the experimenters, which they are clearly unaware of, and thus confabulate.

I’ve suggested that the type of psychological self-ascription, that is confabulated in the split-brain cases, is more accurately described as a rationalisation and, thus, the experimenters are really asking a causal question. Further support for this conceptual distinction can be given by considering a

series of guidelines offered by Russell Hurlburt and Eric Schwitzgebel (2010) in their recent book *Describing Inner Experience?*—a book about testing the reliability of introspective self-ascriptions. According to one of the fifteen guidelines that the authors propose, for attempting to test the reliability of introspective self-ascriptions, one should *avoid* asking subjects to infer causation. That is, one should avoid asking subjects “‘why’ questions’ (Hurlburt and Schwitzgebel 2010, p. 18). This is because, Hurlburt and Schwitzgebel recognise, inferring causation is not a paradigmatic example of a psychological mental state. As the authors were interested in testing the accuracy of *introspection*, with respect to psychological mental states, they made a point to avoid asking such ‘why’ questions. Since such ‘why’ questions are exactly the sorts of questions that are being asked in the split-brain cases, I do not think Carruthers is right to classify the ascription as an intention.<sup>29</sup>

My claim has been that the split-brain data do not lend support to the patterning argument, because the data do not provide us with examples of subjects who have confabulated an intention. What the data do show, I have argued, is that the subjects in the split-brain cases were unaware of the reasons why they got up. Such a result is undoubtedly related to the concept of an intention, but it does not provide us with an example of a confabulated propositional attitude self-attribution, in the way that the ISA theory would predict.

---

<sup>29</sup> I say more about the conception of ‘intention’ in chapter 8.

### 4.3 Patterns and Classification

We have now examined several sets of the confabulation data. I have argued that no discernible pattern, of the sort predicted by the ISA theory, emerges. No doubt further analysis of these data, as well as future data, will reveal other patterns that may be relevant to questions about why confabulation occurs. What I have argued here is that the data currently available do not support the claim that we *always* infer what propositional attitude mental states we are in from sensory, behavioural, or situational cues. So, the claim that we cannot introspect our propositional attitudes remains unsubstantiated

One of the main points that I have sought to express in this chapter is the importance of classification. Whilst this may seem like a minor point, it is of some significance in our current discussion. I have argued that we need to be careful not to draw conclusions about the reliability of certain types of mental state (or mental process) self-ascriptions by looking at the way in which we come to know another type of mental state or process. For instance, suppose that there were data showing that subjects consistently confabulated their character traits. Suppose experimenters asked lazy subjects to state whether they believed that they were lazy people, and most said they did not. Whatever one would be inclined to say about these data, one should not attempt to draw conclusion about a person's propensity to confabulate their intentions, or sensations, because they confabulate their character traits.

To further illustrate this point, let us consider the following from Scaife, who is sympathetic to Carruthers' inference (interpretation) only account. Although he does not advocate an inference (interpretation) only account for all

propositional attitudes, as Carruthers does, he does think that the confabulation data point towards a single-method account for decision-making. He says:

[i]rrespective of how the debate between dual-method and interpretation-only accounts is resolved, we should be concerned about the reliability of our self-knowledge. This is because cases of confabulation are indistinguishable from cases where we gain correct information about our own decision-making. This leaves open the sceptical possibility that, any time we consider our own motivations, we might not be getting accurate information (2014, p. 471).

Although we can agree with Scaife that we *should* be concerned about the reliability of certain self-ascriptions, we need to be careful to clarify what types of self-ascriptions are being brought into question here. Though less radical than Carruthers' position, Scaife is espousing a sceptical argument against subjects possessing two distinct methods—an introspective method and an inferential method—for acquiring knowledge of their own *decision-making* processes. But notice that he does so with reference to *motivations*. As I have mentioned throughout, it is important to distinguish between decisions and motivations. A subject may be wrong about what has motivated her to pick *A* over *B*, but it doesn't follow from this that she lacks self-knowledge about her decision to pick *A* over *B*. This is not to say that motivations are not an important part of what it is to make a decision. Rather, it is to say that motivations are not identical to decisions. Thus, what is true of the self-ascription process of one, may not necessarily be true of the other.

#### **4.4 Conclusion**

I have, in this chapter, challenged the view that the confabulation data support the position that we cannot introspect our propositional attitudes. I began by

clarifying the concept of confabulation, and then explicated three distinct ways in which dual-method theorists could accommodate such a phenomenon. These were: errors of self-deception, brute errors, and basing errors.

Next, I turned to the patterning argument—the argument that states that the pattern of error that exists within the confabulation data is best explained by positing the ISA theory. After examining multiple sets of these data, I argued that this claim is made plausible only at the cost of accepting what I claimed to be unwarranted assumptions about these data. I concluded that there were no reasons to favour the ISA theory’s explanation of the patterning of these data. This result, combined with the results of the previous chapter, gives us good grounds for rejecting the ISA theory.

This now ends my focus on the sceptical accounts of introspection for propositional attitudes. I will now turn my attention to providing a positive view of introspection for propositional attitudes in the forthcoming chapters.



## Chapter 5

### Doxastic Transparency, Rationality, and Introspection

Belief is profoundly analogous to action. Both are commonly grounded in reasons; both are a basis for praising or blaming the subject; both are sensitive to changes in one's environment; both can appropriately be described as objects of deliberation.

Robert Audi ([2001] 2015, p. 27)

[T]o say "I believe that *p*" itself carries, in general, a claim that *p* is true. To say "I believe that *p*" conveys the message that *p* is the case.

Bernard Williams (1973, p. 137)

In the last two chapters, I addressed scepticism about our ability to introspect our own propositional attitudes. I argued that such accounts fail to show that we cannot introspect our own propositional attitudes. In this chapter, I begin to defend a positive account of introspection—namely, a view that is commonly referred to in the literature as the transparency method (hereafter, 'TM').<sup>1</sup> This is a view of self-knowledge that is opposed to the traditional 'inner perception' model of introspection (alternatively referred to as the 'inward looking' or 'inner observation' model).<sup>2</sup> By contrast, TM construes the process by which one can achieve introspective knowledge of what one believes (as well as other kinds of propositional attitudes, as I will argue in chapters 7 and 8) by attending to outward-directed phenomena: such as states of affairs, intentional objects, and reasons in favour of believing something.<sup>3</sup>

---

<sup>1</sup> I follow Cassam (2014) with this terminology. TM is also sometimes referred to in the literature as the 'assent routine' view (see, e.g., Gordon 2007).

<sup>2</sup> By 'inner perception' views, I am referring, quite broadly, to views that construe the process of introspective self-knowledge in terms of the 'internal detection', or 'internal scanning', of a psychological state—such as the 'inner sense' view (see, e.g., Armstrong 1968; Lycan 1996).

<sup>3</sup> As I will go on to explain, the phrase 'outward-directed phenomena' need not necessarily be understood in terms of third-person observational evidence. I will argue that TM can be understood as an introspective approach to self-knowledge.

Although TM has been widely discussed in the past few years, controversy—with respect to several key interpretive questions—remains. These include: questions pertaining to the scope of the view; questions about how TM should be understood within the context of a recent debate between rationalists and empiricists about the epistemic basis of self-knowledge; and questions about whether TM is a genuine competitor to the inner perception view. Such open questions warrant further analysis of TM, which will occur over the course of the next four chapters.<sup>4</sup>

I begin the task, in this chapter, by showing how TM can provide us with a cogent explanation of how one can come to know what one believes. This task will lay the groundwork for subsequent chapters, where I will attempt to show that TM can also explain how one can come to know what one desires, intends, wishes and so on. This chapter will proceed as follows. In §5.1, I explain TM—noting how the method can yield self-knowledge, as well as showing how TM can be thought of as an account of introspection. In §5.2, I consider two positive arguments that can be given in favour of TM. In §5.3, I look at how TM relates to a recent debate between rationalists and empiricists about the epistemic basis of self-knowledge of our beliefs. In section §5.4, I examine a challenge that Brie Gertler has recently put to rationalists. In §5.5, I address Gertler's challenge and argue that the empiricist approach to TM is inadequate.

## **5.1 Doxastic Transparency**

Perhaps the best place to begin our discussion of TM is by examining a passage by Gareth Evans—a passage which has been the subject of much discussion in

---

<sup>4</sup> Some of these issues are discussed in a recent edited volume by Smithies and Stoljar (2012) titled *Introspection and Consciousness*.

the past few years (see, e.g., Moran 2001; Byrne 2005a; Gordon 2007; Silins 2012, 2013). Before showing how TM counts as a form of introspection, I will first present the details of the view. According to Evans,

in making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. If someone asks me 'Do you think there is going to be a third world war?', I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' 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  $p$ ...If a judging subject applies this procedure, then necessarily he will gain knowledge of one of his mental states: even the most determined sceptic cannot find a gap here in which to insert his knife (1982, p. 225).

In this passage, Evans is claiming that one can acquire knowledge of what one believes by attending to an 'outward-directed' question about the world. In order to clarify what is meant by this locution, let us summarise Evans' thought in the following formulation:

*Looking outward method:* one can gain knowledge of what one currently believes by attending to an outward-directed question. One can know whether one believes that  $P$ , by answering the question 'Is  $P$  true?'

I will now explain this account in further detail. First, I will address the question of how an outward-directed approach to self-knowledge can be thought of as a form of introspection. Recall that at the outset of this chapter I claimed that I would defend *an account of introspection*. Doesn't Evans' injunction to attend to outward phenomena—as is described in the above formulation—mean that one would be required to observe one's own behaviour, or other publically assessable data, in order to know what one believes? Wouldn't this, furthermore, make it a

non-introspective approach to self-knowledge?<sup>5</sup>

To show why the answer to this question is ‘not necessarily’, let us examine an example of someone attending to ‘outward phenomena’.<sup>6</sup> Let us consider the example that Evan gives. Suppose it is true that I believe that there will be a third world war. According to the formulation given above, I can know that I do believe this by attending to the following question about the world (outward phenomena): ‘Will there be a third world war?’

According to the looking outward method, I can answer this question by deciding whether I really do think the proposition ‘there will there be a third world war’ is true. Typically, I may do so by considering whether there is any tension within the UN; or by considering whether any countries have active nuclear programs; or by recalling what the newspapers and television stations have been reporting about recent conflicts between the ‘superpower’ countries. After such deliberation, I may judge that there will indeed be a third world war. According to the looking outward approach, if I judge that there will be a third world war, then I am entitled to ascribe the belief to myself that there will be a third world war.

This example shows that deliberation can be involved—at least to some extent—in the belief-ascription process. However, we need to be careful not to overstate the importance of deliberation, and to generalise from Evans’ example

---

<sup>5</sup> Recall from chapter 1 that some philosophers, such as Gertler (2011a), consider TM as a non-introspective approach to self-knowledge because it is not construed in terms of inner perception. In chapter 1, I argued that we should not construe introspection in such restrictive terms. Wolfgang Barz also appears to share the same view as me, as he uses the locution ‘transparent introspection’ (2015, p. 1993) to describe TM, which indicates that he considers the view to be an account of introspection.

<sup>6</sup> I say *not necessarily* here, because I allow that one could follow TM in an inferential (non-introspective) way. For example, one could answer the question ‘Do I believe there will be a third world war?’ by observing one’s own behavioural responses to the question. This would, strictly speaking, count as following TM, but it would be a non-introspective approach to self-knowledge. We should therefore distinguish between an introspective way of following TM, and a non-introspective way of following it. In this chapter, I will be concerned with the introspective approach to TM.

to every case of belief-ascription. While it is true that in the example above, deliberation was involved, it is not the case that every time one gains knowledge of what one believes, one needs to *deliberate*, in this active, cognitively demanding, sense. This is an important point, and one that is worth stressing because, as I will show below, controversy has arisen surrounding the interpretation of how deliberation should feature in the formulation of TM.<sup>7</sup>

What *can* be generalised from Evans' third world war example, I argue, is that it is a subject's judgement, that the proposition 'there will be a third world war' is true, that can give that subject knowledge of what she believes about there being a third world war. Now while it is true that, in the example above, deliberation was clearly involved, this need not be so in all cases of belief ascription. This can be demonstrated by considering another example. Consider my belief that I live in Australia. According to the looking outward method, I can know that I do believe this by attending to the question 'Do I live in Australia?' Since this is a simple question—cognitively speaking—I can attend to it without deliberating, in any significant sense. What I have still done in this case, however, is that I have attended to the question 'Do I live in Australia?'—just like I did in the third world war case. In both cases I am *judging* whether I accept a particular claim about the world to be true, in order to know what I believe. What is different, with respect to my belief about where I live, is that I do not need to 'deliberate' in the way that I did in the third world war example. The proposition instantly 'seems' true to me.<sup>8</sup>

---

<sup>7</sup> I discuss the issue of whether deliberating can 'corrupt' what one believes in the next chapter.

<sup>8</sup> To say that a proposition seems to feel true to me, does not necessarily mean that I have any deep understanding of the proposition. Jordi Fernández gives an example involving Gödel's theorem (2013, p. 47) which nicely illustrates this point. Fernández says that even if someone were to hand him a proof of Gödel's theorem it would still not 'seem' true, even if he followed all the steps to the proof. Nevertheless, Fernández judges that the proposition 'this is a proof of Gödel's theorem' is true.

This discussion pre-empts a further concern with Evans' third world war example—one that has been pointed out by Nico Silins (2012), as well as several other philosophers. This is the problem that Evans' injunction to deliberate about whether there will be a third world war means that one may be possibly launching into a new inquiry, which means one may 'corrupt' what one previously believed. For instance, it may be true that before I deliberated about whether I think there will be a third world war, I neither believed nor disbelieved that there would be a third world war. It is only after deliberation that I come to believe that there will be. This is problematic if the outward directed method promises to yield self-knowledge of what I currently believe.

Although a complete response to this objection will be given in the next chapter (§6.2), a brief reply can be given here. If it is true that Evans' view is committed to the deliberation requirement, then I agree with Silins that we should 'leave that emphasis behind' (2013, p. 296). However, I do not think Evans' view is committed to the thesis that deliberation is needed in every case of belief self-ascription. As I said above, I think that the lesson to learn from Evans' third world war example is the following: by judging that a proposition *P* is true, one can gain knowledge that one believes that *P* is true. Now while explicit deliberation will sometimes be involved here, it need not necessarily be so. This was the case with respect to the example I gave above, where I said that I can attend to the question 'Do I live in Australia?', in order to know whether I do believe that I live in Australia, without any explicit deliberation. Since I have held this belief for a long time, I do not corrupt my belief by attending to the question.

In my view, the Evans inspired view is best described as the view that one can gain self-knowledge of what one believes by attending to an outward directed question about the world. While this will sometimes involve deliberation, it need not always. This process can be construed as a form of introspection, according to my theory-neutral account, in the following way. When I attend to a question about the world, I am judging whether that proposition is true. This counts as an introspective process because such a method can only reveal knowledge of my own mind—in this case, knowledge of what I believe. My attention to a question about the world will not give me knowledge about the mind of another, as I will explain in more detail below. This makes it a process *different* from other ways of gaining knowledge about the world, and can only give me knowledge of my *current* beliefs—again for reasons I will go into below.

Another way to think about the looking outward approach is to say, as several recent philosophers have (see, e.g., Moran 2001, Byrne 2005a, Boyle 2009, Fernández 2013), that the question of what I believe about *P* is transparent to the question ‘Is *P* true?’<sup>9</sup> I think that by focusing on the ‘transparency’ component, rather than the ‘outward-directed’ feature of the view, it is more easily seen as an introspective approach to self-knowledge. For this reason, I will now use the term ‘transparency method’ (TM) to refer to the view. This process can be described as transparent because of the relationship between two questions. *S* can provide an answer to question *Y* (e.g., ‘Do I currently believe that *P*?’) by attending to question *X* (e.g., ‘Is it true that *P* is the case?’). *Y* is

---

<sup>9</sup> This sense of transparency is often related back to G.E. Moore’s discussion in ‘The Refutation of Idealism’, where he says, ‘[w]hen we try to introspect the sensation of blue, all we can see is the blue: the other element is as if it were diaphanous’ (1903, p. 450). Moore’s point here is that we become aware of our experience of blue, not by attending to the experience itself, but to the blue object.

transparent to  $X$  because  $S$  can gain knowledge of  $Y$  by attending to  $X$ . It is important to note that because one question is transparent to another, this does not mean that one question *reduces* to another—a point that Moran (2001, p. 61) makes. One should not get the impression from this discussion that the transparency relationship is reductive in the ontological sense that is commonly understood in philosophy—that is, one should not get the idea that  $X$  is nothing over and above  $Y$ , such as water is nothing over and above  $H_2O$ .

This notion of transparency should also not be confused with other usages of the term ‘transparency’ that exist in the literature. Sydney Shoemaker, for instance, uses the locution ‘transparency thesis’ in his discussion of the Cartesian inspired view that says that ‘nothing can occur in a mind of which that mind is not conscious’ (1996, p. 50). Such a strong thesis is certainly not entailed by TM, I will argue. A similar usage is given by Timothy Williamson, who discusses the ‘myth of epistemic transparency’ (2000, p. 11). A view which he defines as ‘for every mental state  $S$ , whenever one is suitably alert and conceptually sophisticated, one is in a position to know whether one is in  $S$ ’ (2000, p. 24). TM, as I will go on to show, is compatible with error.<sup>10</sup>

Such clarifications are important to make, because if we recall from above, Evans does say that by following TM, one will ‘necessarily...gain knowledge of one of his mental states: even the most determined sceptic cannot find a gap here in which to insert his knife’ (1982, p. 225). Such a passage, with its usage of the word ‘necessarily’, does appear to imply a thesis which resembles the strong views given above. This is a criticism that has recently been pointed out by Quassim Cassam, who thinks that Evans’ usage of the word

---

<sup>10</sup> A related term, ‘transparent’ is used by Peter Carruthers who says, ‘the term “transparent” will be used...in a semi-technical sense, to mean “access to mental states that isn’t interpretive”’ (2011, p. 8). Such a description does not reflect the view that I have been describing above.

‘necessarily’ does suggest that Evans thinks that TM ‘delivers a kind of infallible knowledge’ (2014, p. 102). Despite the existence of this phrase by Evans, I do not think TM should be interpreted in such a way. Not only is the infallibility view quite implausible—as we saw in chapter 2—but Evans himself does not appear to accept the infallibility view, saying that ‘the procedure I have described does not produce infallible knowledge’ (1982, p. 228). TM, I will argue, is compatible with error.

We are now in a position to formulate TM. This can be done as follows:

(TM) A subject can acquire knowledge of what she believes by attending to a question about the world. The question ‘Do I believe that *P*?’ is *transparent* to the question ‘Is it true that *P*?’

In order to comprehend this formulation of TM, it is worth seeing how TM differs from the inner perception view. On the inner perception view, one would attend to one’s inner psychological states by implementing one’s faculty of inner sense. This is the view that Evans appears to target when he says, ‘I shall quite avoid the idea of this kind of self-knowledge as a form of perception’ (1982, p. 225); and, ‘in order to understand the self-ascription of experience we need to postulate no special faculty of inner sense or internal self-scanning (1982, p. 230). TM, by contrast, requires one to attend to the content of what the belief is about, rather than an inner psychological state. For example, instead of detecting the existence of a certain psychological state, such as one’s belief that London is in England, one’s attention would be directed at the content of the belief—namely, to the question: ‘Is it true that London is in England?’

Although I think such passages clearly show that Evans intends TM to be thought of as a rival view to the inner perception view, not all agree. Silins, for example, opposes the idea that there is any conflict between TM and the inner perception view and says, ‘this tempting thought is not clearly correct...it remains perfectly possible that we obtain self-knowledge through inner observation and [the] transparency [method] at the same time’ (2012, p. 305). While I agree with Silins that it is possible to construe TM in such a way, I do not think that this is the correct way to construe TM—a point I have more to say about in §5.5.

This section has provided an explanation of TM—a view of self-knowledge which I have argued can be thought of as introspective. I will now attempt to provide argumentative support for thinking that this view is true.

## **5.2 Arguments in Favour of the Transparency Method**

So far, an explanation of TM has been provided, and several contentious issues associated with the view have been identified. We have not, however, considered what argumentative support can be given for TM. It is to this task I will now turn.<sup>11</sup> In my view, there are two main arguments that can be given for TM. I will call the first of these the *phenomenological argument*. According to this argument, TM best describes the ‘what-it-is-likeness’ involved with coming to know what one believes. The thought here is that if one were to be asked to describe, from the first-person point of view, what the process of self-ascribing a belief feels like, one would describe a process like TM.<sup>12</sup>

---

<sup>11</sup> Silins points out that Evans—who we have drawn upon to formulate TM—offers ‘little to defend the Transparency Thesis’ (2012, p. 302).

<sup>12</sup> For a related discussion about the phenomenology of belief, involving TM, see Markos Valaris (2014).

To support this claim, it is useful to think about the concept of what it means to believe something. Recall the two quotations that I opened the chapter with. The first was from Robert Audi, who claims that belief, like action, can be described as an object of ‘decision and deliberation’ ([2001] 2015, p. 27). In my view, this not only captures something important about the concept of belief, but it also captures something important about the phenomenology of coming to know what one believes. For example, if someone were to ask me whether I believe that it snows in Sweden, I typically would deliberate—that is, I would think about the climate of Scandinavia—before then deciding that it does. I would then ascribe to myself the belief that it does snow in Sweden. Here my attention is focused upon what the belief is about—in this case, whether it snows in Sweden—which accords with the characterisation of TM that was given above.

Since I conceded above that explicit deliberation need not always be involved in the belief ascription process, what should we say about cases where explicit deliberation is absent? In such cases, the subject will still typically take the content of the belief in question to be true. This thought is echoed in the second quotation that I opened the chapter with—namely, the quotation from Bernard Williams, who says that “‘I believe that *p*’ carries, in general, a claim that *p* is true; that is, it is a qualified way of asserting that *p* is true’ (1973, p. 137). Here I think Williams, like Audi, not only captures something important about the concept of belief, but also something important about the phenomenology of believing something. When I am asked whether I believe that it snows in Sweden, what I am doing, typically, is deciding whether the proposition ‘it snows in Sweden’ is true.

While I think that the phenomenological argument is important, I also acknowledge that it cannot, by itself, establish the truth of TM. Rival theorists, such as inner perception theorists, may agree that the what-its-likeness of belief-ascription does accord with TM, but respond that (a) this is not enough to show that TM is true; or (b) deny that this falsifies the inner perception view. Furthermore, philosophers who think that intuition or phenomenology are unreliable guides to discovering truths about the mind (see, e.g., Dennett 1991; Carruthers 2011), will find such style of argumentation unpersuasive. Despite these concerns, I do think that the phenomenological argument is important, and further appeals to it will be made as we proceed.

I will now consider a second—and in my view stronger—argument for TM, which I will call the *Moore's paradox argument*. According to this argument, the implications of Moore's paradox provide us with good grounds for accepting TM. By appealing to Moore's paradox, I am following several recent philosophers (see, e.g., Shoemaker 1996; Moran 2001; Byrne 2005a; Gallois 2011; Silins 2012) who think that the paradox can tell us something important about self-knowledge.

I will first explain what Moore's Paradox (hereafter, 'MP') is. MP arises when sentences of the following form are considered: '*P*, but I don't believe that *P*'; 'I believe that *P*, but it is not the case that *P*'; and '*P* is true, but I don't believe *P* is true'.<sup>13</sup> Consider, for example, that a native English speaker were to assert the following sentence:

---

<sup>13</sup> Moore's Paradox is named after the English philosopher G.E. Moore. As Mitchell Green and John Williams (2007, p. 5) point out, Moore is careful to distinguish between paradox and absurdity. Moore claims it would be absurd for a speaker to utter a MP sentence; and it is paradoxical that such an absurd sentence is not contradictory. The following MP sentence is an example that Moore himself provides: 'I went to the pictures last week, but I don't believe I did' (1942, p. 543).

(1) Canberra is the capital city of Australia, but I don't believe it is.

Now, as Moore (1942) famously pointed out, a speaker who avows (1), may be speaking absurdly, but she need not be speaking falsely. While there is some controversy about why this is so, I think, along with Moore—and Sydney Shoemaker (1996, pp. 75–77)—that this is because while what the speaker *says* in (1) is not a contradiction, what she implies is. What the speaker says in (1) could be true because the sentence consists of two independent claims: one is a factual claim about the world that the speaker judges to be true—namely, that Canberra is the capital city of Australia; and the second is a psychological self-ascription—namely, the speaker's belief that Canberra is the capital city of Australia. A paradox arises in (1) because of what the speaker *implies* in (1), which appears to be the following:

(2) I believe that Canberra is the capital city of Australia.

And,

(3) I do not believe that Canberra is the capital city of Australia.

Unlike in the case of (1), a contradiction would arise if one were to accept (2) and (3) simultaneously. What explains this difference? The answer is that (2) and (3) are stated explicitly in terms of the subject's beliefs. This is not the case in (1). The sentence in (1) comprises of (i) a fact about the world that a subject judges to be true, and (ii) a fact about the subject's psychology—in this case a

belief. This does not entail a contradiction because one's judgement that *P*, and one's belief that *P*, are independent. To see why this is so, we can break (1) up into the following two sentences:

(1a) I judge (accept it as true) that Canberra is the capital city of Australia.<sup>14</sup>

And,

(1b) I do not believe that Canberra is the capital city of Australia.

Now, while no contradiction arises here, it would be absurd for someone to hold (1a) and (1b) simultaneously. This is because, as our discussion in §5.1 attested, judgement is an important part of what it means to believe something. Someone who holds (1a) and (1b), simultaneously, would appear to be conceptually confused about what it means to believe something.<sup>15</sup>

I think that MP supports TM because it highlights just how conceptually connected judgement (accepting the truth of a proposition) is to belief. If one judges that Canberra really is the capital city of Australia, and does not believe it is—e.g., someone who accepts (1a) and (1b)—then such a person appears conceptually confused about what it means to believe something. This gives us good grounds for thinking that a speaker who judges that *P* is true, and is

---

<sup>14</sup> Here I understand the act of judging to involve the acceptance—which may sometimes entail explicit deliberation—that a proposition *P* is true. In other words, when *S* judges that *P*, *S* thinks that *P* is true.

<sup>15</sup> We could imagine some cases where no conceptual confusion would arise. André Gallois (2011, p. 146 ft. 1), for example, points out that an eliminativist who thinks beliefs do not exist may judge that *P*, and not believe that *P*—as there are no such thing as beliefs. And as Alan Hájek points out, someone who rejects very concept of truth ought to be prepared to say '[i]t is raining, but I believe that "it is raining" is not true (for I reject the very notion of truth)' (2007, p. 225). Because I think that beliefs exist and truth exists, like many other philosophers, these are issues that I will not be concerned with here.

familiar with the concept of belief, will also believe that *P*. This was how we formulated TM.

In order to develop this argument further, it is important to point out the limitations of MP. Most importantly, it needs to be pointed out that such a paradox only arises from the first-person point of view. As Moore pointed out, no absurdity would arise in a case like the following:

(4) Canberra is the capital city of Australia, but Tony doesn't believe it is.<sup>16</sup>

And neither is it absurd for a speaker to avow the following:

(5) Canberra is the capital city of Australia, but I did not believe it was last week.

There is nothing absurd about either (4) or (5). If I were to avow (4), then what I would be doing is judging that the capital city of Australia is Canberra; and also making a claim about someone else's psychology. If I avow (5), then I am expressing my current judgement that the capital city of Australia is Canberra; and also making a claim about what I believed the week before. In neither avowal does MP arise; and in neither avowal, do I display any conceptual confusion about the nature of belief.<sup>17</sup>

---

<sup>16</sup> Assuming, of course, that I am not Tony.

<sup>17</sup> MP sentences arise from the first-person point of view, and only apply to current judgements. Notice that these are two of the conditions that I said were required for a process to count as introspective. This speaks in favour of the current view counting as introspective.

How does this discussion about MP help to support TM? It does so in the following way. If one judges that Canberra is the capital city of Australia, and at the same time does not actually believe that Canberra is the capital city of Australia, then one is in the position of the speaker who asserts a MP sentence. One would appear to be conceptually confused about what a belief is. I contend that this supports the following:

(TM) A subject can acquire knowledge of what she believes by attending to a question about the world. The question ‘Do I believe that *P*?’ is *transparent* to the question ‘Is it true that *P*?’

So far, I’ve argued that MP supports TM: the view that someone who judges that *P* is true will typically believe that *P*.<sup>18</sup> One question that remains to be addressed is the issue of infallibility. I noted above that TM does not entail such a thesis. However, if there really is such a congruous connection between judging that *P* and believing that *P*, then might TM be committed to the infallibility thesis? If this is so, then this would count as a major difficulty for the view.<sup>19</sup> What is the relationship between judging that *P* and believing that *P*, and how can judgement and belief break down?

One suggestion for why one can judge that *P* and still not believe that *P* is offered by Shoemaker, who claims that failures of rationality, such as self-

---

<sup>18</sup> It would also be absurd for someone to judge that *P* is the case, and withhold belief that *P*. Consider someone who were to avow the following: ‘I judge that Canberra is the capital city of Australia, but I neither believe it or disbelieve it’. If I accept that Canberra is the capital city but I neither believe or disbelieve it is, then, again, I seem to be conceptually confused about what it means to believe something, even though no contradiction arises.

<sup>19</sup> In chapter 2, I argued that the infallibility view is implausible.

deception, can lead to ‘failures of access’ (1996, p. 71).<sup>20</sup> In such cases, one may judge that *P*, but still not believe that *P*. Others disagree with this contention. Georges Rey (2013), for instance, thinks that failures of rationality are not the only way in which judgement and belief can come apart, and maintains that mismatches between judgement and belief are far more widespread. For example, Rey thinks that ordinary notions like free will, causation, truth, and knowledge, are deeply problematic, yet, nevertheless, Rey claims he cannot help *believing* in them. Rey’s point here is that while he judges that such notions are problematic, and potentially non-existent, he continues to believe in them. Rey thinks that this gives support to the claim that the connection between judgement and belief is merely ‘contingent’ (2013, p. 271). If Rey is right here, then pressure will be applied to the key claim of TM, which is: judging that *P* is true can typically give one knowledge that one believes that *P*.<sup>21</sup> Let us consider two examples of such sentences of the sort that Rey mentions above:

(6) I judge that free will does not exist, but I cannot help believing it does.

(7) I judge that causation does not exist, but I cannot help believing it does.

According to Rey, examples such as (6) and (7) show that Shoemaker’s point is too strong because there are times where a rational person, who also understands

---

<sup>20</sup> Silins argues for the weaker claim that TM gives one ‘justification to believe that one believes that *p* whenever one judges that *p*’ (2012, p. 303).

<sup>21</sup> This, of course, depends on what is meant by ‘typically’. The sense in which I understand the word ‘typically’ here is in the standard dictionary sense—namely, *in most cases*.

the concept of belief, can judge that  $P$  is the case and yet not believe that  $P$ . Rey's point is not that we can come up with counterexamples to TM—the view is not, after all, committed to the infallibility thesis. His point is, rather, that utterances such as (6) and (7) are common enough in everyday life such that they put pressure on the claim that TM 'typically' produces knowledge of what one believes if followed. In response to this concern, let us consider sentences (6) and (7) in some more detail. In my view, even if we admit that (6) and (7) do show that judgement and belief can sometimes come apart, this is not enough to undermine the general tenability of TM. This is because the cases that Rey cites are not representative of our everyday beliefs. The examples he cites involve difficult philosophical theses which are hard to think about. These cases appear to be the exception rather than the norm.

Let us first consider (6). Here Rey imagines a case where someone who is convinced that free will doesn't exist, and thus *judges* that free will does not exist; and yet still *believes* that free will exists, because she cannot help but doing so. Rey considers such a case to be a paradigmatic example of belief and judgement coming apart that is not the result of irrationality or self-deception. Such a case appears to show that mismatches between judgement and belief can arise with more pervasiveness than Shoemaker suggests.

In response to this charge, I will consider (6) in some more detail. Let us imagine that Ben is an incompatibilist about free will—that is, Ben accepts the following conditional: if determinism is true, then free will does not exist. Suppose at some point that Ben becomes intellectually convinced, perhaps by reading a paper by an eminent physicist, that determinism is true. Given Ben's incompatibilism, we can suppose that when he becomes convinced that

determinism is true, Ben comes to judge that free will does not exist. In Rey's view, what someone in Ben's position—that is, someone who judges that free will does not exist—will have trouble believing is that free will does not exist. This is because even after being convinced that free will is an illusion, it is plausible to suppose that Ben will continue to act like he has free will. He will continue to agonise over important life decisions, such as what school he should send his children, or what career path he should take; and he will still have to decide whether to get out of bed each morning. This would seem to suggest that Ben still believes in free will, which does not line up with what he judges to be the case.

Even if we accept that this is the right thing to say in this case, I do not think such an example undermines the general plausibility of TM. For one, the case described above is far from typical. It involves a person who is attempting to square a counterintuitive philosophical doctrine with their everyday behaviour. Rey might be right in thinking that such a case shows that judgement and belief *sometimes* come apart, but the case is certainly not like most of our beliefs. It is better seen as an exception to the norm, and so doesn't undermine the claim that our beliefs are typically doxastically integrated, rather than splintered.

A further point that is worth considering, that Rey ignores, is that the case above may not involve a disparity between judgement and belief at all. Although I will say more about this issue in chapter 6, it is worth mentioning here in order to show that it is far from uncontroversial how to interpret such purported mismatches. Another plausible way to interpret Ben's situation can be described as follows:

(6a) I believe that free will doesn't exist, but I cannot help acting like it does.

(6a) differs from (6) because it doesn't involve a mismatch between what Ben judges and what Ben believes: in (6a) Ben judges that free will doesn't exist and he also believes it doesn't. What Ben will struggle to do, however, if he sincerely avows (6a), is act in a way that is consonant with his belief. That is, even if Ben accepts that free will is an illusion, he must continue to deliberate and agonise over various decisions. While this may be a paradoxical psychological state for Ben to occupy, it wouldn't provide us with a case where Ben judges one thing and believes another. While I grant that such an interpretation is controversial, and requires further argumentation, I think it does show that such an example is not as straightforwardly problematic for TM in the way that Rey imagines.<sup>22</sup>

I think a similar story can be told with respect to (7). Here we are presented with a case involving someone who is struggling to incorporate a counterintuitive philosophical doctrine, that they take to be true, into their everyday life. Even if Rey is right that a subject who judges that causation doesn't exist in the world will fail, or find it difficult, to believe that it doesn't, I do not think that this provides us with a good reason to think that one's judgements do not typically line up with one's beliefs.

To see why this is so, let us consider the example in more detail. Let us suppose that Ben is also an eliminativist about causation—that is, Ben is convinced that causation doesn't exist in the world. Thus, Ben doesn't think that sentences such as 'smoking *causes* lung cancer' and 'overexposure to the sun

---

<sup>22</sup> I consider this issue in more detail in chapter 6.4.

*causes* skin cancer' are literally true. According to Rey, even if Ben was to sincerely judge that causation doesn't exist, he will have a hard time believing it doesn't. That is, he will still likely continue to believe that smoking causes lung cancer and overexposure to the sun causes skin cancer. Even if we accept Rey's interpretation, I do not think such a result gives us reason to doubt that our judgements typically line up with our beliefs. The case involving causation, like the one involving free will, is problematic because it involves a counterintuitive philosophical doctrine that is not easily squared with our everyday experience. We can agree with Rey that such cases show that mismatches are not just limited to episodes of self-deception or irrationality, but I do not think they should be seen as anything more than exceptions to the norm. Typically, most things that we believe are not like this.

While I grant this is a plausible way to interpret the case, I also think it's worth considering whether such a case could be construed as one in which a mismatch between one's judgement and belief does not arise. Another way of describing the psychology of someone in the case above is as follows:

(7a) I believe that causation does not exist, but I cannot help acting like it does.

If this is the right thing to say in the case above, then we do not have an example of a judgement and belief coming apart. Someone who avows (7a) may still judge that that causation doesn't exist, and will also believe it does not; what they will not be able to do, however, is act like it doesn't. A subject who really

does not believe that causation exists, much like the free will denier, will still have to make decisions and attempt to interact with the world.

We do not need to accept this claim, however, to accept TM as a tenable account of self-knowledge. As I said above, we should not construe TM as an infallible guide to self-knowledge. So, it should be allowed that sometimes one's judgement that *P* will not line up with one's belief about *P*. Despite the possibility that judgment and belief can come apart, I still maintain that a strong case can be made for the claim that a rational agent—who is also conceptually competent with respect to what a belief is—who judges that *P* will also *typically* believe that *P*.

In order to develop TM in more detail, I will now focus on what it means to 'judge that *P*'. Since human beings are rational creatures, when we judge, we do so for reasons. I will now discuss this matter within the context of the rationalist versus empiricist debate.

### **5.3 Rationalist Versus Empiricist Approaches to the Transparency Method**

I will now return to a question that was raised at the beginning of this chapter—namely, the question of how TM should be understood within the context of a recent debate between rationalists and empiricists.<sup>23</sup> This is not a debate about what the correct method of self-knowledge is—for example, a debate about whether TM or the inner sense view is true. The debate is, rather, as Brie Gertler

---

<sup>23</sup> Following Brie Gertler (2011a), in her book *Self-Knowledge*, I will use term 'rationalism' for the view that rational agency makes an epistemic contribution to self-knowledge. Since the release of this book, however, Gertler (2016) now prefers to use the term 'agentialism' to describe this view. She claims that this allows us to distinguish between authors who invoke rationality in explaining self-knowledge, without being committed to the view that rationality makes an epistemic contribution to self-knowledge (see, e.g., Shoemaker 1994; Gallois 1996); and authors who think rationality does make an epistemic contribution to self-knowledge (see, e.g., Moran 2001; Boyle 2009, 2011). A third term for this view is employed by Matthew Boyle, who calls it 'reflexive' self-knowledge—since in his view a subject 'must be able to reflect on her grounds for holding a given claim true' (2009, p.150). While there are merits in all three approaches, I will choose to use the term 'rationalism' in what follows, as it contracts well with empiricism.

points out, ‘at its core, a dispute about the *epistemic basis* for knowledge of one’s [propositional] attitudes’ (2016, p. 4).

According to rationalist approaches, our capacity to achieve self-knowledge of *some* of our propositional attitudes, such as some of our beliefs—and even some of our desires, or some of our intentions—stems from our ability to act as rational agents. More specially, it stems from our ability to hold certain propositional attitudes on the basis of reasons. This contrasts with empiricist approaches, which construe the process of achieving self-knowledge of these same propositional attitudes in ways that *are* reducible to empiricist factors such as detection or observation. Empiricist approaches claim that knowledge of a propositional attitude must always be gained empirically and never rationally. Rationalists will allow that knowledge of one’s own propositional attitudes (beliefs, and even some desires and intentions) can sometimes be gained empirically; but they will deny that this is always the case.

It is crucial to point out that rationalism, as I understand the view, does not entail the thesis that rational agency will *always* contribute to the self-knowledge of *every* mental state one is capable of being in; rather, rationalism entails the thesis that there exists a set of mental states, such that our self-knowledge of these mental states requires that rationality is involved. As we are only concerned with belief at this stage, I will not discuss the further question ‘What is the nature of this set of mental states?’ until chapters 7 and 8. What I am concerned with in this chapter is the claim that self-knowledge of some of our beliefs, with respect to TM, requires rational agency. Let me first say a bit more about the nature of this debate

Rationalists (see, e.g., Burge 1996; Moran 2001; Bilgrami 2006; Zimmerman 2008; Boyle 2009), broadly speaking, reject the idea that self-knowledge of our propositional attitudes can *always* be characterised in empiricist ways—that is, in terms of internal observation or some kind of inward glance. One of the main reasons that these authors hold this view is because they think that empiricist ways (e.g., detection and observation) are inadequate to account for the self-knowledge we can have of certain mental states.<sup>24</sup> This is because empiricist notions would distance a subject from the reasons that she believed something, or intended to do something. This is taken to be a problem, as Gertler points out, because ‘a thinker who *detects* a belief or intention...[would be] alienated from that attitude’ (2016, p. 2).<sup>25</sup>

It is worth repeating that rationalists do not argue that all self-knowledge requires rational agency. This is important to point out, because, as I will show, in §5.5, some philosophers have attempted to show that rationalism is implausible because of the fact that some self-knowledge does not require agency. This is not a position rationalists need to be committed to. For example, consider my belief that I am feeling cold. According to TM, I can know that I believe this by judging that I am feeling cold. Such a process does not seem to require me to invoke rational agency at all. All I need to do is pay attention to my cold sensation. Here empiricist notions of detection and observation seem sufficient to know that I believe this.

When we think about other kinds of beliefs, that are not based on sensations, problems begin to arise for the empiricist view. Consider my belief that Canberra is the capital city of Australia. According to TM, I can know that I

---

<sup>24</sup> As Gertler (2016) points out, contemporary accounts of rationalism are inspired by a broadly Kantian approach to reason and agency.

<sup>25</sup> Gertler herself does not accept rationalism.

believe this by judging that Canberra is the capital city of Australia. Unlike in the case of my belief that I am cold, it seems that I do need to do more than simply detect or observe something. Unlike with my belief that I am cold, there are relevant normative factors that are important for me to attend to. For example, I ought to make sure that I have reasons for this belief; and I ought to be ready to abandon this belief given new evidence. Such a requirement appears difficult to account for in terms of detection or observation. This is because my belief that Canberra is the capital city of Australia is, to use the terminology of Thomas Scanlon, a ‘judgment-sensitive attitude’ (1998, chapter 1). According to Scanlon, a judgement-sensitive attitude is a member of the class of mental states:

that an ideally rational person would come to have whenever that person judged there to be sufficient reason for them, and that would, in an ideally rational person, ‘extinguish’ when that person judged them not to be supported by reasons of the appropriate kind (1998, p. 20).

My belief that Canberra is the capital city of Australia is judgement-sensitive because it is sensitive to the reasons I have for holding it. Suppose that tomorrow I found out that the capital of Australia had been changed to Brisbane. I would then have sufficient reason to stop believing that Canberra is the capital city of Australia. Since my belief ought to accord with the evidence, I should be prepared to change it. Such normative requirements are not relevant to my perceptual beliefs—such as my belief that I am feeling cold. As Scanlon puts it, ‘no reason in the standard normative sense can be demanded’ (1998, p. 20) for why I believe that I am cold. The only reason I am required to have is that I am cold.

Empiricists do not deny that agency is *related* to judgement-sensitive attitudes. What they do claim is that our knowledge of our judgement-sensitive attitudes should be reducible to ‘empirical justification or warrant’, as Gertler (2016, p. 4) puts it. Empiricists (see, e.g., Armstrong 1968; Lycan 1996; Goldman 2006; Byrne 2005a; Gerlter 2016) maintain that we can have knowledge of our judgement-sensitive attitudes, such as our beliefs, by detecting or observing an internal mental state. For example, an inner sense theorist, such as Lycan, thinks that one can know what one believes by using one’s own internal scanner. Whereas empiricist interpreters of TM, such as Alex Byrne (2005a) and Jordi Fernández (2013), do not posit internal scanners, but they do think that TM can be accounted for in terms of observation and self-detection.

To be clear, then, there at least two opponents I have in mind here: (1) empiricists who posit internal scanners and (2) empiricist interpreters of TM who don’t posit internal scanners but do posit self-observation and self-detection. Byrne, a proponent of (2), for instance, argues that there exists a mechanism for detecting one’s mental states, but it does ‘not “resemble perception”’ (2005a, p. 80). This differs from proponents of (1)—such as Lycan (1996)—who think the mechanism for detecting one’s mental states (an internal scanner) does resemble perception. Proponents of (1) and (2) both differ from rationalists in that they think rational agency *never* contributes to self-knowledge. In what follows, I will seek to challenge proponents of (2) who attempt to account for TM without appealing to rational agency. In doing so, I also seek to challenge proponents of (1), who also reject the view that rational agency sometimes contributes to self-knowledge.

#### 5.4 Gertler's Challenge

How, then, can we adjudicate between the rationalist and the empiricist version of TM? We can do so by examining a challenge that Brie Gertler—an empiricist—has recently issued to rationalists which she calls the ‘the rationalist’s burden’ (2011a, p. 257). In my view, Gertler’s challenge succinctly shows just what needs to be done in order to defend rationalism. She explains the burden as follows: ‘[i]n order to challenge empiricist theories, rationalists must demonstrate that rational agency makes a particular *kind* of epistemic contribution to self-knowledge, one that is irreducibly normative’ (2011a, p. 257). By ‘epistemic contribution’, Gertler means contribution to knowledge. In her view, it is easy to see how observation and detection—which are characteristics of empiricist approaches—can make an epistemic contribution to self-knowledge because they are ‘recognisably epistemic’ (2011a, p. 257). To see why this is, let us consider an example. Suppose that Tom wants to know whether he needs to bring his umbrella on his way to work. To do this he needs to know if there is a high chance of it raining. In order to acquire such knowledge, Tom might: *look* at the weather outside his window; *listen* to the weather forecast on the radio, or visit the Bureau of Meteorology website and *see* what the forecast is. What such ways of acquiring knowledge all have in common is that they are paradigmatically empiricist—that is, they involve observation or detection of one’s environment. Gertler thinks that since observation and detection *are* recognisably epistemic, in the sense that they are typically associated with knowledge, it’s easy to see how they could contribute to self-knowledge. For example, if one wanted to know whether one believed that snow is white, it is easy to see how looking inside could provide such self-

knowledge. Gertler's challenge to rationalists is that they need to show how rational agency—something that is not recognisably epistemic—can contribute to self-knowledge.

Rationalists can obviously respond to Gertler by acknowledging that rational agency *is* not a typical epistemic notion, but point out that all this shows is that self-knowledge is unique in this respect. One rationalist, Akeel Bilgrami, for example, says exactly this. He states '[t]here is not much that is standard about self-knowledge. It stands out as special among all the other kinds of knowledge that we have' (2006, p. 290). The challenge, then, for rationalists, is to show how some self-knowledge is irreducibly normative—meaning that, it cannot be explained by epistemic notions such as observation and detection. A theory is not a rationalist one, according to Gertler, 'if the normative factors it invokes can be reduced to (or replaced by) non-normative factors, without loss of explanatory force' (2011a, p. 258). If one can have *knowledge* of what one believes, or what one intends, without appealing to rational agency, without loss of explanatory force, then there is no reason to favour the rationalist approach.

We can now summarise this challenge. According to Gertler, the rationalist must, in order to challenge the empiricist approach to self-knowledge, show that the:

normative features of rational agency (i) make a crucial epistemic contribution to knowledge of one's own attitudes, one that is (ii) irreducible to the epistemic factors (evidence, reliability, etc.) countenanced by empiricism (Gertler 2011a, p. 265).

In what remains of this chapter, I will address this challenge by using TM as test case. I will argue that, with respect to our judgement-sensitive beliefs, rational agency is required for self-knowledge of what we believe.

## 5.5 Transparency and Rationality

In response to Gertler's challenge, I will now argue that (a) the normative features of rational agency can make an *epistemic contribution* to self-knowledge. I argue that that this provides support for the thesis that (b) the empiricist account of TM cannot adequately account for the self-knowledge that rational agents can have of their own judgement-sensitive beliefs. In arguing for (a), I am following Richard Moran (2001, 2012) and Matthew Boyle (2009, 2011), who think that rational agency provides an essential contribution to self-knowledge. Moran, for instance, argues that by relying *only* on 'purely epistemic' factors, such as observation and detection, one's capacity for self-knowledge would be severely diminished. He says,

[t]he special features of first-person awareness cannot be understood by thinking of it purely in terms of epistemic access (whether quasi-perceptual or not) to a special realm which only one person has entry. Rather, we must think of it in terms of the special responsibilities the person has in virtue of the mental life in question being *his own* (2001, p. 32).

An important idea that is expressed by Moran in this passage, is the idea that by reducing the process of self-knowledge to epistemic access, one would not be able to account for the fact that one's mental life is one's own. As I understand Moran here, one's mental life being one's own is crucial, because of the responsibilities that one has as an agent with respect to the way that one's own mental states are constituted. This is something that Matthew Boyle also points out when he says that we have 'an ability to know our own minds by actively shaping their contents' (2009, p. 134). If Moran and Boyle are right here, that agency is essential to self-knowledge, then this is something that empiricist approaches will struggle to account for.

Although I agree with Moran and Boyle that there is a set of mental states where it is appropriate to construe the process of self-knowledge in such a way, I think it is also important to point out, before continuing this line of thought, that not all self-knowledge—not even all beliefs, for that matter—should be understood in this way. This point is worth stressing because some philosophers have been critical of the rationalist approach (as I discuss in chapter 7) in light of the fact that such an approach to self-knowledge appears to be limited in its application.

In order to explain this issue in further detail, let us consider two of my current ‘perceptual’ beliefs. Consider, for instance, my belief that I am feeling cold; and my belief that I am seeing a brown coffee mug in front of me. Although in one sense these mental states are *mine*, in the sense that they are occurring to *me*, they are not mine in the sense that I cannot actively shape the contents of these beliefs. Although it is true that I can turn the heating in the room up, in order to not feel cold; and it is true that I have the ability to close my eyes, so that I do not see the mug, I am not responsible for how the mental states themselves are constituted. Moreover, my self-knowledge of these beliefs appears reducible to empiricist factors such as observation or detection, without the loss of any explanatory force. With respect to these specific beliefs, the rationalist approach may seem misplaced.

To see why the existence of such beliefs do not pose a problem for the rationalist approach, let us recall Scanlon’s distinction between judgement-sensitive attitudes and non-judgement-sensitive attitudes (see §5.3). Recall it was said that an ideally rational person does not need to judge, for instance, whether there is sufficient reason, in the normative sense, for believing that she is

perceiving a mug—she just sees it.<sup>26</sup> This is in direct contrast to judgement-sensitive beliefs. Consider my belief that the Earth is spherical, for example. This belief is judgement-sensitive because I *ought* to have reasons, in the normative sense, for why I have this belief. The key point here is that I do not just have this belief for any old reason—or for no reason at all, for that matter. I ought to consider what scientists have had to say about the shape of the Earth; or consider what the photographs taken from space reveal about the shape of the Earth; or see if there is any relevant counterevidence. It is this sense that I am responsible for what I believe, which makes me an appropriate target of criticism if I believe that the Earth is flat, for instance, despite my awareness of the evidence to the contrary. Such normative requirements do not seem applicable with respect to my perceptual beliefs.

The main criticism of the empiricist approach that I will advance here is that the empiricist approach to TM treats judgement-sensitive beliefs like they are non-judgement-sensitive beliefs. This is problematic, I argue, because, as we have seen, judgement-sensitive beliefs have a certain normative element to them. Unless we are prepared to reject the very notion of judgement-sensitive attitudes, I think that we need to take seriously the notion that agency can contribute to self-knowledge.

In order to make this objection more explicit, I will now re-examine TM, and compare the rationalist and empiricist approaches to TM with each other. Specifically, I will discuss Byrne's view, who claims that, 'at least with respect to belief, the inner-sense theory is partly right. There is an inner mechanism for detecting one's beliefs' (2005a, p. 100). Although I will discuss Byrne's view in

---

<sup>26</sup> Here I am talking about *perceiving*. Whether that perception is veridical is, of course, another question.

greater depth in chapter 8, what I now wish to focus on is Byrne's claim that 'detection' can provide knowledge of one's own judgement-sensitive beliefs.

It is important to point out that like Byrne I agree that the question 'Do I believe that *P*?' is *transparent* to the question 'Is it true that *P*?' So, we would both agree, for instance, that I can know that I believe that the Earth is spherical by attending to the question 'Is it true that the Earth is spherical?' Using this belief as an example, we can understand the difference between the empiricist approach to TM (hereafter, 'E-TM') and the rationalist approach to TM (hereafter, 'R-TM') as follows:

(E-TM) I judge, in a way that is reducible to observation and detection, that the Earth is spherical.

(R-TM) I judge, in a way that not reducible to observation and detection, and includes rational agency, that the Earth is spherical.

What is *common* to both approaches is the acceptance of transparency—the thesis that that one can know what one believes by attending to a question about the world. According to Byrne, the *difference* between the two approaches is that R-TM always involves '*self-constitution*' and E-TM always involves '*self-detection*' (2005a, p. 83). Self-constitution, according to Byrne, is the view that one needs to make up one's mind about *P*, or attend to the question 'Is it true that *P*?', in a 'deliberative spirit' (2005a, p. 84), in order to know what one believes. While Byrne thinks that such an approach can sometimes yield self-knowledge, he thinks that in most cases of belief-ascription this not what actually occurs. Byrne thinks, therefore, that such a 'conclusion is overdrawn' (2005a, p. 84)—adding that for this reason, Evans' example about the third world war—i.e., that

one can know whether one believes there will be a third world war by deliberating about whether there will a third world war—is ‘misleading’ (2005a, p. 85).<sup>27</sup>

Although Byrne is not clear about how deliberation could be integrated into his own view, he does appear to construe R-TM as the view that all of one’s belief-ascriptions involve deliberation, or making up one’s mind. Byrne provides the following example, to show that such a position is untenable. He notes that he has believed for some time that Richard Moran is the author of *Authority and Estrangement*, and so doesn’t need to *make up his mind* or *deliberative* about whether Moran is the author of this book, to know that he believes it is so—his mind is already made up (2005a, p. 85). In my view, while Byrne is right about this, the implications of this point are not far reaching. I think that Byrne’s criticism of the claim that one must ‘make up one’s mind’ to know what one believes is somewhat of a red herring. Byrne’s focus on this point distracts us from the real issue, which is that self-detection cannot account for the fact that rational agency is required for self-knowledge of our judgement-sensitive beliefs.<sup>28</sup> Let us reconsider Byrne’s own example to see why this is so.

Like Byrne, I also happen to believe that Richard Moran is the author of *Authority and Estrangement*. And like Byrne, I made my mind up about this issue long ago. I, therefore, do not need to actively deliberate right now, in the

---

<sup>27</sup> It may appear to the reader that the way in which I have characterised Byrne’s view makes him a rationalist. Recall that Byrne claims that answering the question ‘Is it true that *P*?’ in a deliberative spirit is *not always* the way in which one typically comes to have self-knowledge of what they believe. Even if Byrne is right here, couldn’t it still be the case that answering the question ‘Is it true that *P*?’ in a deliberative spirit will *sometimes* yield self-knowledge? Wouldn’t that make Byrne a rationalist, seeing that, presumably, rationalists can disagree over the extent to which self-knowledge involves rationality? In my view, the answer to this question is: no. It does not follow from the fact the deliberation is sometimes involved with self-knowledge, that rational agency is. An inner sense theorist, for example, who thinks that one can gain knowledge of what one believes, by an internal monitor, will not deny that deliberating about whether *P* is true can lead one to believe that *P* is true. What they will deny is that *knowledge* of such a belief requires rational agency. They would insist it always involves internal detection.

<sup>28</sup> Recall that in §5.1, I argued that the Evans inspired view should not entail the thesis that explicit deliberation is required every time that one achieves self-knowledge. So here I agree with Byrne.

sense that I need to make up my mind, to know that I do believe this. It does not follow from this, however, that reason or agency can be fully discounted from the process of knowing that I believe this. I still need to be aware of the reasons I have for holding the belief. If I lack good reasons for holding this belief, or if new counterevidence becomes available to me—e.g., I were to learn that *Authority and Estrangement* was the product of a computer program that generated philosophy books—I ought not continue to believe that Richard Moran was the author of *Authority and Estrangement*.

So, even if it is true that I have already ‘made up my mind’ in the literal sense that Byrne alludes to, this does not show that such normativity can be discounted. What I am claiming here is that order to judge that Richard Moran was the author of *Authority and Estrangement*, I need to accept that proposition as true. It is hard to see how I can do this if I am restricted to self-detection as described in E-TM.

Once we divorce R-TM from the thesis that one must deliberate, or make up one’s mind, every time that one attempts to achieve self-knowledge, I think that Byrne’s criticism loses its force. This is a point that even Moran himself concedes. He says, ‘the role I give to the deliberative stance is not meant to suggest that most of our beliefs are actually formed through explicit deliberation or reasoning’ (2004, p. 458). I agree with Moran that this is the right thing to say here. Most people’s beliefs about the country they live in, for instance, are unlikely to be the result of explicit reasoning. Again, however, this does not show that reason can be wholly discounted. My belief that I live in Australia may not be the result of deliberative reasoning, but I still need to be aware of reasons for why I have this belief.

By equating R-TM with the thesis that one needs to make up one's mind, or deliberate, in every instance of belief ascription, Byrne has mischaracterised R-TM. While a proponent of R-TM will certainly grant that making up one's mind, or explicit deliberation, will sometimes be involved in coming to have self-knowledge of what one believes—such as when new beliefs are formed—they need not be committed to the thesis that self-knowledge of what one believes always requires one to make up one's mind every time. What the proponent of R-TM should be committed to is the thesis that rational agency cannot be discounted from one's attempt to gain self-knowledge of what one believes. An alternative way of construing the rationality requirement is to appeal to the notion of commitment—something that Moran captures in the following:

as I conceive of myself as a rational agent, my awareness of my belief is awareness of my commitment to its truth, a commitment to something that transcends any description of my psychological state. And the expression of this commitment lies in the fact that my reports on my belief are obligated to conform to the condition of transparency: that I can report on my *belief* about X by considering (nothing but) X itself (2001, p. 84).

For Moran, a commitment to X, involves an *active* willingness to accept certain claims about X—that is, to weigh up evidence for X, or consider reasons why X is true. While this may sometimes involve deliberation or making up one's mind—and sometimes self-constitution—it need not always. My acceptance of the claim that the Earth is spherical involves a commitment of mine that the proposition 'the Earth is spherical' is true. If I follow E-TM, and self-detect that I judge that the Earth is spherical, I appear to forgo the reasons I have for holding the claim to be true.

Whilst the rationalist approach to self-knowledge needs to be developed further (the topic of chapters 7 and 8), there are three main theses that I hope to have shown the plausibility of here: (i) following TM can give one knowledge of what one believes; (ii) the rationalist view should not be thought of as the view that self-constitution, making up one's mind, or explicit deliberation, is involved in *every* belief-ascription; and (iii) rational agency provides an essential contribution to self-knowledge of one's own judgement-sensitive beliefs.

## **5.6 Conclusion**

This chapter has been focused on defending TM. In §5.1, I explained TM, and showed how the view could be considered as a form of introspection. In §5.2, I put forward two arguments in favour of TM. In §5.3, I introduced the rationalist versus empiricist debate, and showed how it is relevant to our concerns about TM. In §5.4, I discussed a challenge that Gertler has given to rationalists. In §5.5, I attempted to respond to Gertler by showing that rational agency provides an essential contribution to self-knowledge, thus motivating the acceptance of the rationalist approach to TM. In order to develop this view further, I will, in the next chapter, address a series of objections that have recently been made with respect to this view.

## Chapter 6

### Objections to the Rationalist Interpretation of the Transparency Method

In the previous chapter, I defended the transparency method (hereafter, ‘TM’). On this view, one can determine whether one believes that *P*, by attending to the question ‘Is it true that *P*?’<sup>1</sup> I also argued for rationalism—the thesis that the normative features of rational agency sometimes make an epistemic contribution to self-knowledge. Let us call the combination of these two theses the rationalist interpretation of TM. Before I attempt to extend the application of the rationalist interpretation of TM to other non-doxastic propositional attitudes (e.g., intentions, desires, wishes, hopes, and so on), the topic of chapters 7 and 8, I will first address a series of recent objections that have been made to the view.

In §6.1, I address the *homo philosophicus* objection—the objection that the rationalist interpretation of TM overstates what it is for humans to have self-knowledge of what they believe. In §6.2, I address the corruption objection—the objection that by attending to a question about the world, as is required by TM, one may potentially corrupt what one believes. In §6.3, I address the epistemic stance objection—the objection that rational agency often requires one to take a purely epistemic stance towards one’s own mental states. In §6.4, I address the matching problem for TM—the claim that the connection between judgement and belief is not always as congruous as TM theorists maintain it is.

---

<sup>1</sup> I have argued that TM should be thought of as an introspective approach to self-knowledge. This means that I think it can explain first-person authority (see chapter 2).

## 6.1 The *Homo Philosophicus* Objection

Quassim Cassam objects that rationalism overstates what it is for human beings to achieve self-knowledge of what they believe (as well as what they desire, intend, and so on). According to Cassam, while rationalism might describe the way that '*homo philosophicus*, a model epistemic citizen' (2014, p. 2), would come to have self-knowledge of what she believes, for instance, it is not the way that regular human beings typically achieve self-knowledge. Let us call this *the homo philosophicus objection*.<sup>2</sup>

Although Cassam objects to rationalism in general, he raises the *homo philosophicus* objection with specific reference to TM. Because there are rationalist accounts of self-knowledge in the literature that do not explicitly endorse TM (see, e.g., Burge 1996), it is unclear to what extent Cassam's objection can be generalised to all accounts of rationalism. As things stand, however, Cassam's objection is directly relevant to the rationalist interpretation of TM, so it is an objection that needs addressing here.

In order to adequately assess this objection, let us recall the formulation of TM, with respect to belief, that I gave in the previous chapter:

(TM) A subject can acquire knowledge of what she believes by attending to a question about the world. The question 'Do I believe that *P*?' is *transparent* to the question 'Is it true that *P*?'

I also argued that since humans are rational agents, when they judge that *P* is the

---

<sup>2</sup> Cassam's invocation of *homo philosophicus* is inspired by a distinction that is made in economics between *homo sapiens* and *homo economicus*—a perfectly rational agent who thinks and decides in an exemplary way. Like her economic cousin, *homo philosophicus* is a perfectly rational agent who, according to Cassam, reasons critically, believes what she ought rationally to believe, and doesn't suffer from self-ignorance (2014, pp. 52-53).

case—with respect to their judgement-sensitive beliefs—they should do so for reasons, in the normative sense. I argued that the process of answering the question ‘Is it true that *P*?’, with respect to judgement-sensitive beliefs, cannot be reduced to self-detection or observation. Following Moran (2001), I argued that one needs to commit oneself to the truth that *P*, which requires one to be aware of the reasons why it is that one judges that *P*.

In Cassam’s view, the rationalist interpretation of TM overstates what it is for humans (or at least most humans) to have self-knowledge of what they believe. In order to see why Cassam thinks this is the case, it is important for us to see how he formulates the rationalist interpretation of TM. Cassam’s own construal of the rationalist interpretation of TM differs slightly from the way that I have construed it above. Because of this, I argue that Cassam’s objection misses the mark. In Cassam’s (2014, p. 4) view, it is the following formulation, given by David Finkelstein, which describes rationalism.

The question whether I believe that *P* is, for me, transparent to the question of what I ought rationally to believe—i.e. to the question of whether the reasons require me to believe that *P*. I can answer the former question by answering the latter (2012, p. 103).

Let us call Finkelstein’s alternative formulation of TM—which Cassam uses to formulate the *homo philosophicus* objection—TM\*. The main difference between TM and TM\* is that according to TM\*, the process of achieving self-knowledge is described in terms of what one *ought rationally to believe*. This differs in a subtle, but important, way from TM.

According to Cassam, coming to know what one ought to believe about *P* is something that only a *model epistemic citizen* could determine. Since, as

Cassam points out, most of us are far from model epistemic citizens, this would put the process of coming to know what one believes, for instance, out of reach of most people. Cassam thinks this leads to what he calls the ‘Substitution problem for Rationalism’ (2014, p. 104). As Cassam sees it, TM\* requires that one substitute an easy question, ‘Do I believe that *P*?’, with a more difficult question: ‘Do the reasons require me to believe that *P*?’ To give an example, Cassam claims that the question ‘Do the reasons require me to adore my dog?’ is not only a more difficult question than ‘Do I adore my dog’, but it is the ‘*wrong* question’ (2014, p.105) to ask if one wants to know whether one adores one’s dog.

Cassam makes the same point by considering another example, from Krista Lawlor (2009), about a mother who is attempting to gain knowledge of what she desires. In this example, we are to imagine Katherine, a mother who is struggling to answer the following question, one for which she thinks there exists a matter of fact answer to: ‘Do I want another child?’ Cassam says:

[n]otice how odd it would be to answer this question by asking herself whether she ought rationality to want another child. She might ask herself this question if she conceives of herself as *homo philosophicus* but if she is a well-adjusted human being there are lots of other things she can and will do to answer her question’ (2014, pp. 142–143).

In both examples, I am in complete agreement with Cassam. I think that if one wanted to know what one believes, adores, or desires in these cases, then attempting to determine what the reasons require one to believe, adore, or desire would be the wrong thing to do. This is because we often believe or desire what we ought not to. For the model epistemic citizen who always believes what she

ought to, or desires what she ought to, such a process could yield self-knowledge—but that is clearly not going to work for most people. The real question is: ‘Does TM\* accurately describe the transparency procedure?’ In my view, it does not. The examples that Cassam brings up here are red herrings, and are not pertinent to our current concerns.

In order to further illustrate the difference between TM and TM\*, let us consider an example, involving both views. Suppose Alvin is asked whether he believes that Barack Obama was a fair president. On TM, as I have formulated it, Alvin would be able to know whether he does believe this by attending to a question about the world—namely, ‘Was Barack Obama a fair president?’ Now, because such a belief is judgement-sensitive, I’ve argued that rational agency cannot be discounted. This means that one will need to have reasons for why one believes such a proposition. One must do more than just treat the question ‘Was Barack Obama a fair president?’ as a brute stimulus.

According to TM\*, the question that is transparent to Alvin’s psychological mental state—in this case a belief—is not his judgement about whether Barack Obama was a fair president, but rather a question of whether the reasons require him to believe that Barack Obama was a fair president. This is clearly a more difficult, and demanding, question for Alvin to answer, compared to what is described in TM. In order to answer such a question, Alvin may need to think long and hard, and potentially gather external evidence. Furthermore, Alvin may not even know how to determine what he ought to believe. Descriptively speaking, this is surely not how belief ascription works, as we do not typically struggle to determine what we believe. I contend that Cassam has misconstrued TM, and has torn down a straw man.

One may respond to my criticism of Cassam's formulation of the rationalist interpretation of TM by saying that it is I who have mischaracterised the view—and it is TM\* that should be adopted. In response to this concern, it is important to recall one of the initial motivations for embracing TM: Moore's paradox. Recall our discussion in chapter 5. There I said it would, typically, be absurd for someone to judge that *P* is true, and simultaneously believe not-*P*. For example, it would be absurd for someone to sincerely say, 'It is true that Africa is a continent, but I don't believe that Africa is a continent'.<sup>3</sup> It was such an absurdity, I argued, that provided support for TM—the thesis that one can know what one believes about *P*, by attending to the question 'Is *P* true?'

On Cassam's formulation of TM\*, conversely, no such absurdity materialises. There is nothing absurd about someone who ought rationally to believe not-*P*, and yet believes that *P* is true. On Cassam's view, the following example is both possible, and unproblematic. Suppose that David is attempting to answer Michael's question 'Do you believe that the Moon landing was a hoax?' According to TM\* this requires David to answer the question 'Do the reasons require me to believe that the Moon landing was a hoax?''<sup>4</sup> Now, since there are not any good reasons for accepting this conspiracy theory, David should answer 'No' to the latter question. However, suppose that David were to still believe that the Moon landing was a hoax, and is also indifferent to what he ought to believe. If this is so, then David doesn't have any sound reasons to believe that the Moon landing was a hoax; nevertheless, he believes it was a hoax anyway. Now while David's reasoning is faulty here, he is not conceptually confused about what a

---

<sup>3</sup> Moore's paradox arises in sentences such as this one because, while the sentence may be absurd, it could still be true. This is because it doesn't involve a contradiction. It involves a judgement about a proposition concerning the world, and a psychological self-ascription.

<sup>4</sup> On TM, as I construe it, if one wanted to know whether one believed that the Moon landing was a hoax, one would answer the question: 'Is it true that the Moon landing was a hoax?'

belief is. Moreover, no absurdity is present. To believe that *P*, even though one ought not to, is not only possible, but actual in many situations. We often believe what we ought not to. Without support from Moore's paradox, it is difficult to see why we should accept TM\* in the first place.

In summary, I think that Cassam has misconstrued the rationalist interpretation of TM. I agree that self-knowledge would be overly demanding if TM\* were true, but I have argued that there are good reasons not to accept TM\*.

## 6.2 The Corruption Objection

The second objection I will examine is one which pertains to a specific feature of TM—namely, the feature which requires one to attend to a question about the world, in order to know what one believes. According to some, attending to a question about the world, i.e., 'Is *P* true?', in order to determine whether one believes that *P*, may corrupt, or alter, what one actually believes about *P*. Let us call this the *corruption objection*.<sup>5</sup>

The corruption objection can be explicated in greater detail by considering an example. Suppose Jane is asked at time *t*<sub>1</sub> whether she believes that long haul flights are bad for the environment, because they contribute to climate change. If Jane follows TM, in order to gain knowledge of what she believes about this claim, then she may ask herself the following question: 'Are long haul flights bad for the environment, because they contribute to climate change?' Proponents of the corruption objection do not doubt that Jane may gain knowledge about what she believes *after* asking herself this question at time *t*<sub>2</sub>. What they do claim is that if Jane really is interested in knowing what she

---

<sup>5</sup> As I interpret the corruption objection, it is applicable to both rationalist and empiricist interpretations of TM.

believed at  $t_1$ , then deliberating about whether long haul flights are bad for the environment should be the last thing she should do—as it may change what she actually believes at  $t_1$ . According to some philosophers (see, e.g., Shah and Velleman 2005, Reed 2010; Gertler 2012a; Cassam 2014) this shows that TM will often fail to be a good way to determine what one currently believes.<sup>6</sup>

This objection can be summarised by considering the following from Nishi Shah and J. David Velleman.

The question “Do I believe that  $p$ ” can mean either “Do I already believe that  $p$  (that is, antecedently to considering this question)?” or “Do I now believe that  $p$  (that is, now that I am answering the question)?”...one can answer the question *whether I now believe that  $p$*  by forming a conscious belief with respect to  $p$ , whereupon one’s consciousness of that belief will provide the answer; but one cannot answer the question *whether I already believe that  $p$*  in a way that begins with forming the belief (2005, p. 506).

According to the proponents of this objection, one must refrain from any kind of deliberation, if one wishes to know what one believes at the time the question is asked. Shah and Velleman, for instance, say ‘asking oneself *whether  $p$*  must be a brute stimulus in this case rather than an invitation to reasoning’ (2005, p. 506); and Gertler says, ‘if this method is to reveal our pre-existing beliefs, we must not gather new evidence concerning  $p$ . That is, *we must limit ourselves to looking inward*’ (2011b, pp. 132-133).<sup>7</sup> Such comments are a direct challenge to the view of self-knowledge I have been defending so far.

The first thing to say in response to the corruption objection is to point out that the process of ‘reasoning’ or ‘gathering evidence’ will not *always* corrupt what one believes. One’s long-standing beliefs, for instance, will not always change, or be corrupted, just because one attends to the content of what

---

<sup>6</sup> Cassam, alternatively, calls this the problem of ‘antecedent belief’ (2010, p. 90).

<sup>7</sup> I understand the term ‘brute stimulus’ to mean that one must not engage with the content of the belief.

one's belief is about. This may seem like a trivial point, but it is one that is underappreciated by the above authors, whose point about corruption is, in my view, overstated. Shah and Velleman, for instance, claim we *must* treat the question of what we believe as a brute stimulus; and Gertler claims that we *must not gather new evidence*. It is tempting to interpret such claims—with their usage of the word 'must'—as suggesting that reasoning or gathering evidence will always provide an obstacle to self-knowledge. If this is the right way to interpret these claims, then the objection clearly fails.

In order to see why this is the case, let us consider an example. Imagine that Professor Randall teaches a class on environmental ethics that he gives every semester. Suppose that each semester, after his first lecture on climate change, his students invariably ask him the same question: 'Do you believe that Earth will be inhabited by humans in 100 years?' Being an optimist, Professor Randall always replies 'Yes, I do believe the Earth will be inhabited by humans then'. Now, if Professor Randall were to follow TM, in order to know what he believes about this matter, each time he is asked this question, there is no reason to think that following TM would 'corrupt' what he believes, just because TM requires him to consider the question itself. His optimism means that he will give the same answer each time the question is asked.

Although supporters of the corruption objection may concede this point, their central concern remains unanswered—namely, what about cases where deliberation will corrupt what one believes? If we interpret the corruption objection as the objection that reasoning and evidence gathering will sometimes corrupt what one believes, then the objection has more force. What about a case, for instance, where Professor Randall is asked 'Do you believe that Earth will be

inhabited by humans in 100 years?', and instead of answering 'Yes', he first deliberates by recalling a recently published paper that he read that morning, which makes him reflect upon the question. Suppose that after such reflection, he gives the pessimistic answer of 'No' to his students. In this case, it seems right to say that before the question was asked, at time  $t_1$ , Professor Randall believed that the Earth *would* be inhabited by humans in 100 years; and then, at time  $t_2$ , after he deliberated about the question, he no longer did.

Now, while such a case is certainly possible, I do not think that it has the negative implications for TM that some have suggested. To see why this is so, I will consider Shah and Velleman's and Gertler's suggestion that what we need to do in such cases is to treat questions about what we believe as *brute stimuli*. What I will argue here is that their solution gives rise to an even greater problem than the problem of corruption, and thus proponents of the corruption objection trade a small problem for a large one.

To see why this is so, let us try to imagine professor Randall treating the question 'Do you believe that Earth will be inhabited by humans in 100 years?' as a brute stimulus. What would it mean for him to do so? One possibility is that the professor must 'look within', while refraining from any engagement with the content of the question at all. One problem I have with this suggestion is that it would require Professor Randall to refrain from any serious engagement with the content of the belief, meaning that he would have to ignore the reasons why he accepted that belief in the first place. Recall that in the last chapter, I argued one of the essential characteristics of belief is that it involves an acceptance of a claim as true. As Bernard Williams points out, to avow "I believe that  $p$ " itself

carries, in general, a claim that p is true. To say “I believe that p” conveys the message that p is the case’ (1973, p. 137).

The problem with treating a question about what one believes as a brute stimulus, is that one will be forced to avoid considering the reasons why one accepts the belief in the first place. As Richard Moran points out ‘[s]imply hearing oneself coming out with something in response to a brute stimulus will provide no more reason for thinking this represents one’s beliefs about something than if one were to sneeze in response to the stimulus’ (2012, p. 221). If Professor Randall refrains from engaging with the question about whether humans will inhabit the Earth in the future, then he may avoid corrupting what he believes, but he is left with an even greater problem—namely, the problem of how he can be sure that he even believes that humans will inhabit the Earth in the first place. In other words, how can he be sure that it is a *belief* he has, rather than say a desire, or just a random thought?

What we need to acknowledge, in order to adequately respond to the corruption objection, is that ‘corruption’ is just a problem with belief self-ascription in general. Sometimes the very act of asking someone what they believe will corrupt or alter what they believe. So, while this is certainly a problem for TM, other views of self-knowledge, such as the inner perception view, will also be affected by it. If Professor Randall is asked ‘Do you believe that Earth will be inhabited by humans in 100 years?’ and the inner sense view is correct for example, then he may still corrupt what he believes if he first considers the recent article he read, before answering the question.<sup>8</sup>

---

<sup>8</sup> Unless he is explicitly told ‘tell me what you believe, but do not think about the content of the belief at all—that is, do not engage with the content of the belief’. This would be a very strange thing to ask someone to do, and it is not clear one would understand what is being asked, if one was to be given such a command.

In summary, the corruption objection does not have serious implications for TM. Sometimes the process of answering a question about what one believes will lead one to ‘corrupt’ or ‘change’ what one believes. However, this is merely a consequence of how belief ascription typically works. It is no more a problem for TM than it is for other theories of self-knowledge.

### 6.3 The Epistemic Stance Objection

The third objection that I will examine is one that Barron Reed makes, with respect to rationalism.<sup>9</sup> While Reed finds much that is ‘illuminating’ (2010, p. 165) about rationalism, he argues that it ‘cannot be the whole story, for we still must take a purely epistemic stance toward our own mental states on *many occasions*. Rational agency requires it’ (2010, p. 165, my emphasis).<sup>10</sup> What Reed is objecting to here is the generalizability of the two central theses that constitute rationalism. These were described earlier as (i) rational agency makes a significant epistemic contribution to self-knowledge; and (ii) knowledge of one’s attitudes is irreducible to *typical* epistemic factors described by empiricism (e.g., detection, observation). In Reed’s view, self-knowledge sometimes requires that (i) and (ii) are explicitly bypassed, and a purely epistemic (in the empirical sense) stance is taken towards one’s own mental states. Let us call this the *purely epistemic stance objection*. In order to formulate this objection, Reed

---

<sup>9</sup> Like Cassam, Reed formulates his objection to rationalism with specific reference to TM. This means that Reed’s objection is directly relevant to the rationalist interpretation of TM.

<sup>10</sup> Recall that rationalism is not the thesis that rational agency will *always* contribute to self-knowledge—rather it is the thesis that rational agency will *sometimes* contribute. According to Reed’s objection, as I see it, rationalism is far more circumscribed than many rationalists would take it to be. Reed, for instance, would object to the claim I made in the previous chapter that rationalism is applicable to the set of beliefs that are best referred to as judgement-sensitive beliefs. According to Reed’s objection, there are many occasions where one will need to refrain from rational agency in order to have self-knowledge of one’s judgement-sensitive beliefs. Given the way in which I have defined rationalism—namely, as the thesis that rational agency can sometimes make an epistemic contribution to self-knowledge—Reed would count as a rationalist because he appears to allow that rational agency can sometimes contribute to self-knowledge. Reed would disagree with many rationalists about the scope of rationalism, however. His view appears to be that rationalism is far more circumscribed than many of the view’s proponents would claim.

offers the following case.

Suppose, for example, that Penny is an economist who began her career by writing several first-rate papers on taxation policy. Over the years, though, her interests have changed, and she now works on issues related to global poverty. But, recently, her department has hired a new economist, Bill, who specializes in taxation policy. Happy to have a colleague with some common interests, Bill asks Penny about her views on taxation. As they begin to discuss some of the finer points of a progressive income tax, Penny realizes that she does not remember all of the details of the view she laid out in her early papers (2010, p. 176).

Reed argues that this case is problematic for the rationalist interpretation of TM in the following way. He considers what would happen if Bill were to ask Penny whether she believes that *P*—where *P* is a complex position about income tax that Penny formed a view about years ago, but has now simply forgotten. Reed thinks that if Penny were to follow TM, and attend to the question ‘Is *P* true?’, Penny would have to judge that she does not believe that *P*, because she would not be able to recall the specific details of *P*. Reed argues that the question of whether she believes that *P* cannot be ‘answered by considering nothing but the evidence relevant to *whether p*’ (2010, p. 176). Reed thinks this is because Penny would be unable to commit herself to the truth of *P*, and would have to concede that she does not believe that *P*.

In Reed’s view, what Penny needs to do, in order to have self-knowledge of what she believes, is to attend to her previous judgement that *P*—a judgement that was made years ago. This would, according to Reed, involve bypassing any engagement with the content of the belief that *P*—that is, the complex position about income tax—and instead treat the question ‘Is *P* true?’ in purely epistemic terms (e.g., in terms of detection, or observation of the mental state itself). Thus,

Reed thinks it would be wrong for Penny to follow TM, in order to know what she believes, as she wouldn't be able to say that she believed that *P*. And given that the most plausible thing to say is that Penny still believes that *P*, this gives the wrong answer.

The first thing to say in response to Reed's objection is that Penny *does* have evidence to ground her judgement that *P* is true—namely, her memory that she once believed that *P*. Although Penny will not be able to articulate the reasons why she was once convinced that *P* is true, her memory that she once believed that *P* is a reason for her to judge that *P* is true.<sup>11</sup> Whether this reason is a good one—that is, whether her memory is veridical, or whether her original position was well thought out—is a separate matter that need not concern us here.<sup>12</sup> The case involving Penny is not a paradigmatic case of someone bypassing the process of judging whether *P* is true, as Reed maintains it is. If Penny really was bypassing the process of engaging with *P*, she would presumably not defer to her memory of once having defended *P*. It is only by engaging with *P*, that she realises that she cannot remember the details of *P*. And thus, she defers to her memory.

Reed is aware of this kind of response, but dismisses it because he doesn't think that it captures the relationship that Penny has towards her earlier work on taxation policy. Reed thinks that if Penny really was abiding by TM in this way—that is, by relying on her own memory that her previous expert self once endorsed—then she may as well defer to the opinion of another expert

---

<sup>11</sup> It might be objected that deferring to memory is a form of detection or observation, but this is not clearly so. When one judges that *P* is true, because one remembers having once thought that *P* is true, one is simply using their memory as a way of coming to judge that *P* is true.

<sup>12</sup> One can, for instance, judge that *P* is true for absurd reasons, and then, by following TM, have knowledge that one believes that *P*. For example, one can judge that the Earth is flat because it looks flat, and have a true belief that one believes that the Earth is flat.

about whether *P* is true. Reed claims, '[i]f she had, let us suppose, a book on taxation written by the economist generally regarded as the acknowledged master of the field, she just as well might have consulted it (2010, p. 177). In other words, Reed is suggesting that if the right thing to say in this case is that Penny judges that *P*, because she once believed that *P*, then Penny may as well attend to the question 'Does expert *Y* say that *P* is true?' If she judges that the expert does, then she should believe it. Reed thinks that such a result is problematic because the expertise of another doesn't seem relevant to what Penny already believes.

In my view, Reed's response does not work in the way that he thinks it does. While Reed is right to be concerned with what justifies a belief, this is a concern quite separate from our current concerns about TM. Let me explain. Consider a case where one does rely on the testimony of an expert for believing that *P*. Consider for example, that an expert doctor writes, in her best-selling book, that the liver performs over 500 different functions. Not being an expert on biology, suppose I read this book and accept that this claim is true and, thus, attribute this belief to myself. This example is unproblematic, because I have no earlier belief about the number of functions the liver has. Now, the question is: why is such an appeal to expertise any less tenable, in the case where I already have formed a belief about *P* in the past?

Even if Penny is being asked whether she believes that *P*, and defers to an expert, she will still have reasons for accepting that *P*. Suppose she replies to Bill by saying 'I cannot remember the details of *P*, but I do know of an expert who endorses *P*, so I judge that *P* is true'. The question of whether this provides Penny with legitimate justification to believe that *P* is a separate question, and it

does not show that one should refrain from rational activity. According to TM, one can know that one believes that  $P$ , by attending to the question ‘Is it true that  $P$ ?’ This is exactly what Penny has done here.

In summary, I do not think that Reed’s example is successful in showing that we must take a purely epistemic stance towards our judgement-sensitive beliefs on various occasions.<sup>13</sup> All Reed’s example shows is that sometimes the only reason we may have for thinking  $P$  is true is that we once held  $P$  to be true, or that an expert holds  $P$  to be true. Whether these are good reasons for accepting that ‘ $P$  is true’ is an important question, but ultimately a different question from whether TM and rational agency is the right way to think about self-knowledge of one’s judgement-sensitive beliefs.

#### **6.4 When Judgement and Belief Come Apart: The Matching Problem**

The fourth objection that I will examine is an objection to the defining characteristic of TM—namely, that one’s judgement that  $P$  can give one knowledge about whether one believes that  $P$ . I briefly raised this objection in chapter 5.5, where I looked at the concern that the connection between one’s judgement that  $P$  and one’s belief that  $P$  is merely a contingent one. In this section, I will return to this issue and examine similar claims made by authors who cite examples of judgement and belief coming apart, in order to undermine the general plausibility of TM. Following Cassam (2014, p. 117), I will call this the *matching problem* for TM. In order to explain this problem in sufficient detail, I will first consider two cases that exemplify the problem. I then argue that

---

<sup>13</sup> Recall that in order to take a purely epistemic stance, with respect to gaining self-knowledge of one’s own mental states, one must refrain from any sort of rational activity. One must rely purely on self-detection or on inner observation of a psychological mental state.

such cases, as well as cases like them, do not have the negative implications for TM that some (see, e.g., Schwitzgebel 2012b; Gertler 2011b; Cassam 2014) have thought.

Before looking at these two examples, I need to introduce the conception of dispositional beliefs. Let us say that *S* dispositionally believes that spilling the salt brings bad luck, for example, if *S*, *ceteris paribus*, feels fear when she spills the salt; throws a pinch of salt over her shoulder when salt is spilled; starts sweating when the salt is spilt; and so on.<sup>14</sup> Dispositional beliefs can be contrasted with occurrent beliefs—which have been our main focus so far—because they can be attributed to a person even when that person is not conscious of the belief. *S* can be said to dispositionally believe that spilling the salt brings bad luck, for instance, even when she is asleep.<sup>15</sup>

According to Brie Gertler, cases where subjects judge that *P*, but behave like they dispositionally believe not-*P*, provide grounds for doubting the claim that TM can give one access to one's dispositional beliefs, which in Gertler's view are 'arguably more central examples of belief than occurrent judgments' (2011b, p. 126). If this is correct, then the scope and generalisability of TM will be jeopardised. In order to challenge this claim, I will now consider two cases of mismatch in more detail.

---

<sup>14</sup> This example is from Gertler (2011b, p. 134).

<sup>15</sup> It is important not to confuse the concept of dispositional beliefs, with *dispositionalism*, a theory about what it is to believe something. A representationalist theorist, or functionalist theorist, about belief will also grant the reality of dispositional beliefs. See Schwitzgebel (2015) for a detailed account of the various positions one can take, with respect to what it means to believe something. Here I attempt to be as neutral as possible about these different views.

#### 6.4.1 Two Cases of Mismatch

The first case I will examine draws upon an example discussed by Gertler (2011b). In this case, we are to imagine a person, Nick, who was raised to believe that spilling salt will bring bad luck. Nick was also raised to believe that bad luck can be reversed by dropping a pinch of salt over his shoulder. Now an adult, Nick has come to realise that this superstitious belief ought to be abandoned. When he is asked whether he thinks that spilling the salt brings bad luck he replies ‘No’. However, as Gertler notes, whenever Nick spills salt he is overcome with a sense of doom, and is compelled to throw drops of salt over his shoulder, in addition to exhibiting other behaviour that suggests he still believes that spilling salt brings bad luck.

Gertler thinks that this case poses a problem for TM in the following way. If Nick attends to the question ‘Does spilling salt bring bad luck?’, he would answer, ‘No, it does not—that is nothing but a superstition’. Given what we have said so far about TM, by attending to such a question about the world, Nick should have transparent access to his belief that spilling the salt does not bring bad luck. According to Gertler, however, this is not what Nick dispositionally believes. She says ‘[t]his inclination is best explained by attributing to Nick the belief that spilling salt brings bad luck’ (2011b, p. 135). There is a thus *matching problem*. Nick judges that spilling the salt does not bring bad luck, but he behaves as if it does, and thus he arguably believes that it does bring bad luck.

A similar problem arises in a second case that is discussed by Eric Schwitzgebel (2010), which involves a university professor, Juliet, who judges that all races are of equal intelligence. Schwitzgebel describes the case in such a

way that this is not just a passing thought that the professor has. Juliet has studied the arguments for racial equality, and is convinced by them. She is also prepared to argue vehemently for intellectual equality, and has done so on several occasions. Yet, as Schwitzgebel describes the case, her behaviour diverges from what she judges to be the case: she is racist in many of her actions. It affects her grading, the way that she holds class discussions, and her behaviour in hiring new staff. Again, we have a *matching problem*. Juliet judges that the races are of equal intelligence, but she doesn't behave like she does.<sup>16</sup>

#### **6.4.2 Possible Responses to the Matching Problem**

We have looked at two different cases that purport to display a disparity between (i) a subject's judgement that *P* and (ii) what that subject actually believes about *P*. How should such cases be understood with respect to the matching problem for TM? One possibility to consider is that such cases are mischaracterised as mismatches. We might say that in Nick's case, Nick doesn't really believe that spilling the salt brings back luck. And in Juliet's case, we might say that Juliet does believe that all races are of equal intelligence. If this response is tenable, then there is no matching problem. A second strategy would be to concede that such cases, as well as others like them, do display a discrepancy between a subject's judgement that *P* and their belief that *P*, and then show how such mismatches are not enough to undermine the general plausibility of TM.

In order to explore these two different approaches, I will draw upon a recent paper from Schwitzgebel (2010), who has usefully catalogued a series of views that provide competing explanations of how such purported mismatches

---

<sup>16</sup> Unlike Gertler, Schwitzgebel is not specifically concerned with providing a counterexample to TM. He is, rather, interested in how we should describe the psychology of subjects in such cases.

should be characterised. Such views offer different answers to the question: ‘How should we understand the psychology of someone who judges that *P*, but behaves like they believe not-*P*?’ The views that Schwitzgebel discuss are as follows:

- (1) The *pro-judgement view* (see, e.g., Zimmerman, 2007; Gendler, 2008). On this view, the subject believes that *P*, and also does not believe not-*P*. For example, *S* believes that Canberra is the capital city of Australia, and also does not believe that Canberra is not the capital city of Australia.<sup>17</sup>
- (2) The *anti-judgement view* (see, e.g., Hunter 2009, Gertler 2011b). On this view, the subject does not believe that *P*, and instead believes not-*P*. For example, *S* does not believe that Canberra is the capital city of Australia, and instead believes that Canberra is *not* the capital city of Australia.
- (3) The *shifting view* (see, e.g., Rowbottom 2007). On this view, the subject shifts from believing *P* and believing not-*P*. For example, *S* shifts from believing that Canberra is the capital city of Australia, to believing that Canberra is *not* the capital city of Australia.
- (4) The *contradictory belief view* (see, e.g., Gertler 2011b, Bilgrami 2006). On this view, the subject believes both *P* and not-*P*. For example, *S* believes that Canberra is the capital city of Australia, and also believes that Canberra is *not* the capital city of Australia.

---

<sup>17</sup> Schwitzgebel notes that in each case it is determinately true that *S* believes that *P* or *S* believes that not-*P*.

- (5) The *in-between belief view* (see, e.g., Schwitzgebel 2010). On this view, a subject does not determinately believe *P* or not-*P*. The case is best described as one where the subject is ‘in-between’ believing *P* and not believing not-*P*. For example, it is not quite right to say that *S* determinately believes that Canberra is the capital city of Australia or that *S* believes that Canberra is not the capital city of Australia.

Although these views are not characterised in terms of exclusivity—in the sense that they apply to every case of mismatch—there has been, as Miri Albahari point out, a tendency to ‘shoehorn all the discordant cases into their preferred analysis’ (2014, p. 702). Given the complexity and variety of situations in which mismatches can be said to occur, as will be examined below, I think we are better off treating these views as rival explanations in *certain situations*, rather than rival explanations that cover all cases. That is, I agree with Albahari that we should accept the ‘*the contextual view*’ (2014, p. 703). According to Albahari, this view involves ‘overturning [the] assumption of uniformity...On the contextual view, which analysis applies to which case depends on the discordancy case at hand’ (2014, p. 701). This means that while there will be cases where any of (1)–(5) may be the best explanation, we should not expect that any one view will explain every case.

Given that I do not accept the claim that any particular one of these views should be uniformly adopted in all cases, we need to consider each view separately. For purposes of clarity, it will be useful to divide the above views into two sets. I will call the first set *non-mismatch* explanations. This set includes views (1) and (3). Such explanations do not pose a problem for TM, because they

do not assert that any mismatch between a subject's judgement that  $P$  and their belief that  $P$  has occurred. I will call the second set *mismatch explanations*. This set includes views (2) and (4). Such views pose a problem for TM, because they involve cases where a subject judges that  $P$  and believes not- $P$ . I will avoid examining (5). This is not because I think the view is implausible, but rather, it is because I think that insofar as (5) raises a matching problem, it can be responded to in the same way that I respond to the other mismatch explanations. In cases where a person *in-between* judges that  $P$ , and then *in-between* believes that  $P$ , there would not be any mismatch between what the person judges and what the person believes.

### 6.4.3 Non-Mismatch Explanations

One way to characterise cases that appear to exemplify the matching problem, is to deny that such cases really do display a disparity between a subject's judgement that  $P$  and what that subjects believes about  $P$ . In the first case we considered, we can say that Nick *doesn't* believe that spilling the salt brings bad luck. In the second case, we can say that Juliet really *does* believe that all the races are of equal intelligence. If such explanations are right, then the mismatch problem does not arise.

Let us examine this type of response in a bit more detail. If we recall from the list of views that we considered above, there are two ways in which this might occur. One way is to accept the shifting view. On this view, a subject shifts from believing that  $P$  to believing not- $P$ . For example, in the first case, it could be that when Nick judges, at time  $t_1$ , that spilling the salt does not bring bad luck, he really does believe that it will not. When he behaves as if he

believes that spilling the salt brings bad luck, at some further time  $t_2$ , Nick really does believe that spilling the salt does bring bad luck. His belief changes from  $P$ , at time  $t_1$ ; to not- $P$ , at time  $t_2$ .

In the second case, Juliet might judge, at time  $t_1$ , that all the races are intellectually equivalent, and really believe that they are. Once Juliet behaves like she believes that they are not, at a further time  $t_2$ —e.g., when she is giving a lecture, marking papers, or hiring new staff—her belief changes, such that she believes that the races are not intellectually equivalent. When Juliet is arguing for racial equality her belief could change back again. If such an account is plausible there is no mismatch between her judgement and what she believes.

Before assessing the applicability of the shifting view, with respect to these two cases, I will first consider a case where the shifting view is well suited. Consider, for example, a meat eating individual, Tom, who attends a debate about the ethics of eating meat. Suppose that before the debate, Tom judges that eating meat is morally permissible, and determinately believes that eating meat is morally permissible. While listening to the debate, Tom learns new facts about the treatment of animals in some slaughterhouses, and comes to judge that eating meat is immoral. Suppose that after the debate, at a local diner, Tom sits down with his friends and orders a steak. Here the shifting view is a plausible way of explaining this mismatch. When Tom judges during the debate that eating meat is immoral, we can suppose that he really does believe that it is. We can imagine that he would not eat meat at this time, if he was offered some. A few hours later, when Tom is hungry, and enough time has passed such that the thoughts of the slaughterhouse are distant, Tom judges that eating meat is morally permissible, and his belief shifts back to believing that eating meat is morally permissible.

How plausible is the shifting view, with respect to the other two cases mentioned above? The answer to this question depends on how the specific details of the cases are framed; it also depends on how the specific details of the subject's psychology are described. In the case involving Tom, I was careful to stipulate that enough time has passed from the time of his judgement that *P*, and his belief that not-*P*. I also stipulated a reason for the change in attitude. If similar stipulations were to be made in the Nick and Juliet cases, then the shifting view may explain the purported disparity. However, if these cases are such that the time between the subject's judgement that *P*, and the discordant not-*P* behaviour are closer together, and there is no explicit change in the subject's psychology, then this strategy is less likely. The shifting view can explain some examples of purported mismatch, but not all.

A second way in which the non-mismatch approach may be applicable is by appealing to the *pro-judgement view*. Unlike the shifting view, this view can accommodate cases where there is no significant time difference between one's conscious judgement that *P* and one's behaviour that indicates not-*P*. According to the pro-judgement view: a subject who judges that *P*, and behaves like they believe not-*P*, believes that *P* and fails to believe not-*P*. One way in which the pro-judgement view may be realised is by appealing to what Tamar Gendler (2008) calls *aliefs*. According to Gendler, an 'alief' is a piece of terminology that can be used to describe the psychology of someone whose behaviour appears to contradict what she explicitly endorses to be the case, without the need to ascribe contradictory beliefs to that person. Gendler characterises an alief as follows:

...paradigmatic alief is a mental state with associatively linked content that is representational, affective and behavioral, and that is activated—consciously or nonconsciously—by features of the subject's internal or ambient environment. Aliefs may be either occurrent or dispositional (2008, p. 642).

To explain aliefs further, let us consider an example. Imagine a person, Suzy, who is shaking and trembling while she is standing on the Grand Canyon Skywalk—a cantilever bridge with a glass walkway, that enables people to look down into the Grand Canyon.<sup>18</sup> The question is, ‘Does Suzy believe that the Skywalk is unsafe?’ On the one hand, it does seem plausible to ascribe this belief to Suzy. After all, if we observe Suzy’s behaviour—such as her shaking and trembling—there appears to be good grounds for attributing this (dispositional) belief to her. On the other hand, there are good reasons for not ascribing to her such a belief. Surely Suzy doesn’t really believe that the bridge is unsafe—after all, she has voluntarily walked onto the Skywalk. Why would she walk out if she believed it was unsafe? Gendler’s solution to this question is to say that in such a case Suzy *really does believe that the bridge is safe*, however she *alieves that it is unsafe*.

Aliefs—dispositional and occurrent—can be described as habitual, or instinctual, emotional or behavioural reactions to circumstances, both real and imagined, that passively occur to a subject. Unlike typical beliefs, aliefs are not the result of any acceptance of a claim to be true—the kind that would result after deliberation.<sup>19</sup> According to Gendler, the content of someone’s alief, who is trembling on the Skywalk, may be something like ‘[r]eally high up, long long way down. Not a safe place to be! Get off!!’ (2008, p. 635). This is an

---

<sup>18</sup> This is one of the examples Gendler (2008) uses to explicate the concept of an alief. The bridge is over 200 metres from the ground.

<sup>19</sup> I am not saying that all beliefs require deliberation, of course. I currently believe that a computer is in front of me, for example, and this does not involve much in the way of rational deliberation.

instinctual, automatic, reaction to the situation that simply occurs to her. This differs from Suzy's belief that the bridge is safe—a belief that is sensitive to what she judges to be true. If Suzy receives news that part of the bridge has collapsed, for example, she should change her belief about the bridge being safe.

By accepting the reality of aliefs, we can avoid attributing to Suzy contradictory beliefs—or at least attributing a mismatch between what she judges to be true and what she believes. This allows us to say that when Suzy judges that the walkway is safe, she believes it is safe. What she does not believe is that the walkway is safe—her behaviour is evidence of this. One advantage that the pro judgement view has over the shifting view, specifically in the Skywalk example, is that it can be applied in cases where a subject simultaneously judges that *P* and behaves in way to suggest they believe not-*P*. We do *not* need to posit any significant time gap between a person's judgement, and a person's discordant behaviour.

Can the pro judgement view be applied to the previous cases we looked at—namely, the salt spilling case and the racist professor case? Might the discordant behaviour present in these cases be ascribed as aliefs? As I said with respect to the shifting view, I think that the answer to this question depends on what is stipulated in each case. I certainly grant that it is possible, but perhaps not likely, that in the salt spilling case Nick may have the following alief: 'salt spilled. Activate salt throwing over shoulder mechanism now!', and at the same time Nick determinately does believe that spilling the salt is just a superstition.

One problem with thinking that aliefs can explain the disparity between judgement and behaviour in many other cases, is that sometimes such discordant behaviour appears to satisfy all the conditions required to count as a dispositional

belief. It may be fine to say in the Skywalk case that Suzy does not believe the bridge is unsafe (at least if she's not phobic), but there is a worry that if we start to deny that certain other instances of behaviour do not count as beliefs, then we risk being inconsistent in our attribution of belief. This is a point that Gertler makes (2011, p. 137), and one that I think we need to take seriously. Gertler claims that Nick's salt throwing behaviour, for example, appears to fulfil all that is required for a belief to count as a dispositional belief. If we deny that such behaviours really are beliefs, then a problem arises—namely, that our concept of dispositional belief starts to break down. If we deny that Nick believes that spilling the salt brings bad luck, then the worry is that we would also have to deny other cases, which really do seem like paradigmatic cases of belief.

I have argued that *sometimes* when *S* judges that *P*, and behaves in a way to suggest that she believe not-*P*, *S* may not necessarily believe not-*P*. There are times where non-mismatching options—e.g., the shifting view and the pro judgement view—are viable explanations. In such cases, there is no issue for TM. I have also tried to show the limits of these approaches. I will now move on to the non-mismatch approach, which does provide a problem for TM.

#### **6.4.4 Mismatch Explanations**

So far, we've looked at two different ways in which the non-mismatch explanation may be the right way to describe a case where someone judges that *P*, and then behaves like they believe not-*P*. In cases where such an explanation is correct, there will be no matching problem for TM. I will now consider the mismatch explanation, which does pose a problem for TM. This will involve looking at the anti-judgement view—where a subject fails to believe that *P*, and

believes not-*P*; and the contradictory belief view—where a subject believes that *P* and also believes not-*P*. My aim here is not to show that such explanations are more or less plausible than non-mismatch explanations, but rather, I will argue that (i) *some* cases should be described as mismatches; and (ii) such occurrences are not enough to undermine the plausibility of TM.

The first view I will consider is the anti-judgement view. On this view, when a subject judges that *P*, and behaves in a way that suggests they believe not-*P*, that subject fails to believe that *P*, and actually believes not-*P*. In the salt spilling case, for instance, this would involve Nick failing to believe that spilling the salt doesn't bring back luck, and would involve him believing that it does. Gertler (2011, p. 137) thinks that (i) Nick's superstition satisfies all the conditions that are required for it to count as a belief and (ii) such a continued belief explanation fits well with what is called belief perseverance in the social psychology literature. This phenomenon, according to the online Psychology Dictionary, involves a 'tendency to persist with one's held beliefs despite the fact that the information is inaccurate or that evidence shows otherwise' (Nugent 2013). If this is the right way to describe the salt spilling case, then we can say that Nick continues to believe that spilling the salt brings bad luck, even though he is aware of the reasons for he should not hold the belief. Gertler claims that since the phenomenon of belief perseverance is widespread, and mundane, TM will commonly fail to yield self-knowledge if it is followed.<sup>20</sup>

One problem with Gertler's claim that belief perseverance supports the anti-judgement view, is the fact that belief perseverance, as we have defined it above, is described in terms of objective, or mind independent, reasons—not

---

<sup>20</sup> Again, the line of argument here is putting pressure on the claim that TM will 'typically' yield self-knowledge. This is because the situation described is so mundane.

what the subject herself judges. Belief perseverance need not necessarily involve a situation where someone judges that *P* is true, and yet believes not-*P*. To see why this is the case, let us consider an example. Suppose that Jack fails to acknowledge that his spouse, Jill, is being unfaithful. We can imagine that Jack fails to recognise the signs of Jill's infidelity—signs that are obvious to all of Jack's friends. In this situation, Jack may still *judge* that Jill is being faithful to him, and *believe* it to be so. While it may be true that Jack ought to believe that Jill is being unfaithful, given the evidence he is aware of, his belief *perseveres*. Such a case, however, doesn't support the claim that Jack really does believe that Jill is being unfaithful. Jack's failure to judge that Jill is being unfaithful explains why he continues to believe it. Here Jack's belief perseveres even though he has reason to abandon it. So, while belief perseverance may indeed be mundane, not all cases of it would involve a mismatch between judgement and belief.<sup>21</sup>

The case involving Nick, therefore, is disanalogous to the case involving Jack—even though both may be described as examples of belief perseverance.<sup>22</sup> Unlike in Jack's case, where judgement and belief do line up, Nick's belief that spilling the salt brings bad luck does not, as stipulated by Gertler, line up with his judgement that it does not. While Gertler's appeal to belief perseverance may explain why Nick continues to believe that spilling the salt brings bad luck, it doesn't explain why Nick's judgement and belief fail to line up. It is, thus, unclear why Gertler thinks that the general phenomenon of belief preservation supports the anti-judgement view. While belief perseverance may explain why people continue to believe things even after they have good evidence not to, it is not obvious how it can explain mismatches between judgement and belief.

---

<sup>21</sup> I am not claiming that Gertler explicitly claims this, however.

<sup>22</sup> I take the case involving Jack to be a typical case.

The main problem I have with the anti-judgement view, is not that it is undersupported by the phenomenon of belief perseverance. It is that it ignores what is explicitly judged by the subject. What I think Gertler underappreciates here is the role that deliberation and rationality play in belief formation and maintenance. In chapter 5, I argued that to believe something, typically, is to hold that something to be true. I did not claim that this holds necessarily, but I did say that there is a presumption that someone who legitimately judges *P* to be true, also believes that *P*. Now I don't claim that this is always the case, but I think that by denigrating this feature of belief, we are dismissing an important part of what it means to believe something. This is a point that Schwitzgebel echoes, when he says that the anti-judgement view

omits what the subject explicitly endorses, how she is disposed to judge the overall state of affairs all things considered, what side she would take in an argument, how she is disposed to reason about the case in reflective moments, her best conscious assessment of the evidence (2010, p. 542).

To deny that an agent who explicitly endorses that *P* is true, and defends *P* in arguments, does not actually believe that *P* is, I think, problematic; just as it is problematic to deny that an agent who behaves like she believes not-*P* actually does believe not-*P*. Because Gertler does not appear to share this view, she thinks that the case involving Nick 'does not strike us particularly bizarre' (2011, p. 138). However, in my view, the opposite is true: I think the case involving Nick is bizarre. Belief perseverance may be common, as Gertler suggests, but for Nick to explicitly judge that spilling the salt brings back luck—that is, to be convinced that the claim is true—and yet throw salt over his shoulder when he spills the salt, suggests that something out of the ordinary has occurred. As I said

in the previous chapter, in our discussion of Moore's Paradox, there is a certain absurdity involved when someone judges that *P*, e.g., that it is raining, and also believes not-*P*, e.g., that it is not raining. To say that Nick believes not-*P*, when he explicitly endorses *P*, is problematic in a way that I think Gertler does not fully appreciate.<sup>23</sup>

Despite my criticisms of the anti-judgement view, I do not claim that it should be ruled out. It may, for instance, provide an appropriate explanation in some cases—such as cases involving self-deception. We can imagine, for example, in the salt spilling case that Nick—in the company of his scientist colleagues—is asked whether he thinks that spilling the salt will bring bad luck. Aware of the criticism that will face him if he admits to believing such a claim, we can suppose Nick says ‘No, *it is* just a superstition’. Let us also assume that Nick actually judges that it isn't. Nevertheless, when he spills the salt, Nick immediately throws a pinch of salt over his shoulder, and grimaces in sudden consternation. Here it may be plausible to say that Nick always did believe that spilling salt will bring bad luck, even though he judged that it did not. It may be that because of Nick's fear of the condemnation that he would receive from his colleagues that he self-deceives himself into thinking that he doesn't really believe spilling the salt brings back luck. So, despite the fact that Nick really does judge that spilling the salt does not really bring back luck, he believes that it does. This is an interpretation of the case where the anti-judgement view may be applicable.

The second mismatch option I will examine is the contradictory belief view. This view says that when a subject judges that *P* and behaves in a way that

---

<sup>23</sup> Gertler does not consider the anti-judgement view to be the only plausible option here. She thinks that the contradictory belief view could also be the right way to interpret the Nick case.

suggests she believes not-*P*, that subject believes that *P*, and also believes not-*P*. One advantage that this view has over the anti-judgement view, is that it can account for the fact that what the subject explicitly endorses is reflected in what she believes—as just discussed above.

Let us consider the contradictory belief view, with respect to Schwitzgebel's example involving the racist professor. Here we can say that Juliet judges that all the races are intellectually equivalent, and believes that they are; and at the same time, she believes, unconsciously, that all the races are not intellectually equivalent. We can suppose that Juliet is unaware of this latter belief because it has been repressed, for example. It may be too uncomfortable for Juliet to admit to herself that she has this belief, because she would have to accept facts about herself that she would be ashamed of.

The contradictory belief view may also be a plausible way to explain other cases that arise in the social psychology literature, which also display a disparity between what a subject explicitly endorses about *P*, and how that subject behaves with respect to *P*. Consider explicit association tests, for instance. These are psychological tests that can sometimes show a disparity between what a subject endorses and how they behave. For example, when subjects are asked to associate black and white faces with positive and negative words, they are sometimes quicker to associate certain coloured faces with positive words (see, e.g., Greenwald et al. 2009). Or because of a racial stereotype, subjects are more likely to identify an ambiguous object in a picture as a gun, when the face they saw associated with the object was black, compared to when it was white (see, e.g., Payne 2006). Such data is interesting because it can come from subjects who endorse positive views about egalitarianism, or who

explicitly object to racism. Subjects can be surprised by such results, suggesting that the subjects were unaware, or are unconscious, of parts of their own psychology.<sup>24</sup>

Now, if one were to follow TM, in order to have knowledge of such prejudices, one would fail to secure self-knowledge of them. For instance, suppose someone with an unconscious racist attitude was to judge that all of the races are of intellectual equivalence, and then proceed to ascribe the belief to themselves that all of the races are of intellectual equivalence. Such an attribution of a belief would be in contradiction with what the subject unconsciously believes, and so it would seem that TM would fail to give one knowledge of what one believes. So, here we have a case where a subject's judgement does not provide a good guide to what that subject believes.

The quick response to this objection is that such subjects who follow TM may still have a true belief about what they believe. Recall that on the contradictory belief view, we can still say that a subject who judges that *P*, will believe that *P*. So, subjects will still have a true belief about their psychology. However, they will also believe not-*P*, which they will not be able to have knowledge of by following TM. So, we are still left with the problem that TM fails to yield self-knowledge in such cases. Gertler (2011b, p. 141) in response to this issue, rightly points out that even if we accept the contradictory belief view, we are still faced with a situation where TM fails.

Even though Gertler is surely right here, I do not think that her objection is enough to undermine TM. One reason for thinking this is that TM, like any other theory of self-knowledge, should allow for the fact that errors are possible.

---

<sup>24</sup> See Neil Levy (2014, 2017) for a discussion about how such data relates to moral responsibility.

It should be compatible with the fact that self-deception, or certain biases, can lead one to self-ignorance about what one believes. Gertler's insistence that the contradictory belief view is problematic for TM appears to be predicated upon the assumption that TM is committed to the thesis that such kinds of errors are not possible. I've argued (see chapter 2) that regardless of the view of self-knowledge one adopts, error is something that needs to be accounted for.

In order to elaborate upon this issue, let us consider what Moran—one of the leading proponents of TM—has said about the limitations of the method. Moran notes that there will be times where TM will fail, and a subject's awareness of their own psychology will have to come from a non-introspective source. For example, let us imagine a case where someone is unaware of the resentment they feel towards a sibling. Suppose that this person has followed TM, in order to have knowledge of this resentment, and has failed to secure self-knowledge of her resentment. Moran says the following about such a person's psychology:

when she reflects on the world directed question itself, whether she has indeed been betrayed by this person, she may find out that the answer is no or can't be settled one way or the other. So, transparency fails because she cannot learn of this attitude of hers by reflection on the object of that attitude. She can only learn of it in a fully theoretical manner, taking an empirical stance toward herself as a particular psychological subject (2001, p. 85).

As Moran illustrates in this passage, there will be times when TM fails, because there will be times when our judgements about the contents of our mental states will be affected by our psychological shortcomings. Whether such cases are best described in terms of the anti-judgement view or the contradictory belief view, we should not expect that TM will be able to give us access to all of our mental

states. In the case that Moran describes, a therapeutic, or non-introspective approach, will sometimes be needed in order for one to have knowledge of a mental state.<sup>25</sup>

Furthermore, if self-knowledge is tied to rationality, in the way I have been describing, then such shortcomings of TM should be expected. The following from Sydney Shoemaker is helpful in explicating the connection between self-knowledge, rationality, and failures of self-awareness.

If our special access to our mental states is tied to rationality...we have an explanation of how it can be the case both that the access must be appreciable and that it is less than perfect. It must be appreciable because an appreciable degree of rationality, and the first-person access that comes with it, is required for the very existence of mental states of the sort we have. It is less than perfect because our rationality is less than perfect. Failures of rationality, such as those involved in self-deception, can bring with them failures in access. But if we try to suppose such failure the rule rather than the exception, we overstep the bounds of intelligibility (1996, p. 71).

There are two important points that Shoemaker makes here that I think are worth highlighting. The first is his focus on rationality. Because we are far from perfectly rational (we are not *homo philosophicus*, in Cassam's terms), and we are prone to episodes of self-deception, as well as other cognitive limitations and biases, there will be times where we will fail to achieve self-knowledge. The examples discussed in this section attest to this.

The second, and perhaps more important point that Shoemaker makes, is one about the implications of such errors. I agree with Shoemaker that we need to be careful not to infer that, just because it is possible for failures of access to occur, that such failures of access are the norm (rather than the exception). The

---

<sup>25</sup> See Bilgrami's (2006) Appendix 1 – 'When Self-knowledge Is Not Special (with a Short Essay on Psychoanalysis)' for a discussion about such cases, where one can only achieve self-knowledge in an empirical, third-person, manner.

strategy of providing examples of mismatch between judgement and belief, in order to undermine TM, can only work if it can be established that such errors are in fact the norm. Such a claim has not been established in experimental psychology, or in everyday experience.

In summary, I've argued that the mismatch objection is not enough to undermine TM. I have granted that sometimes when a subject judges that *P*, she will believe not-*P*. As I have shown, however, such a result does not show that TM is an unreliable guide to self-knowledge. It only shows that error is possible. As I have also said, error is something that any theory of self-knowledge that seeks to explain self-awareness should be able to account for.

## **6.5 Conclusion**

In this chapter, I considered four distinct objections that face the rationalist account of TM. I have argued that such objections are not enough to undermine TM. One concern that I did not address here is how TM can account for the awareness one can have of other types of propositional attitudes such as intentions, desires, and so on. According to some philosophers, TM is limited in this respect. At best, it can only be extended to a small sub-class of beliefs and is, thus, not a good general theory for how it is we come to have knowledge of our propositional attitudes. In the next chapter, I respond to this objection by showing how TM can account for the awareness we have of propositional attitudes other than belief.

## Chapter 7

### Extending the Transparency Method Beyond Belief—Part One: The Scope of Rationalism

In the last two chapters, I have defended a view of self-knowledge called the transparency method (hereafter, ‘TM’). I have also argued for rationalism—the thesis that rational agency can sometimes make an epistemic contribution to self-knowledge. In chapter 5, I argued that TM can account for the introspective way in which one can achieve knowledge of what one believes; and in chapter 6, I was concerned with answering various objections to TM. In this chapter, I will address a problem that has been lurking in the background of our discussion and has yet to be addressed. This is the concern that TM is limited in its applicability: that TM cannot be generalised to account for the self-knowledge one can have of one’s other propositional attitudes—such as one’s desires, intentions, wishes, hopes, fears, and so on. According to this objection, TM can only yield self-knowledge of what one believes.<sup>1</sup> Call this the *generality objection*.<sup>2</sup>

In order to adequately address the generality objection, it is important that the rationalism versus empiricism debate also be taken into consideration. While somewhat orthogonal to the question of whether TM can be generalised to other types of mental states, the debate has a bearing upon the attempts to extend TM, along with the various arguments for upholding the generality objection. This is because these arguments are often given with specific reference to either an empiricist or rationalist interpretation of TM. So, in order to adequately assess

---

<sup>1</sup> Proponents of this objection include Shaun Nichols and Stephen Stich (2003), David Finkelstein (2003), and Peter Carruthers (2011).

<sup>2</sup> I follow Quassim Cassam (2014) with this terminology.

the generality objection, we will need to be aware of how the two interpretations of TM differ.

I begin in §7.1 by formulating the generality objection. I distinguish two different versions of the objection, and state the one on which I will focus. I then assess the reasons typically given for upholding such a formulation of the generality objection. In §7.2, I will offer a positive account of where I think the scope of rationalism lies. I argue that the set of mental states that are referred to as judgement-sensitive attitudes fall within its purview. In section §7.3, I offer some critical remarks on one recent attempt to extend TM beyond belief. Such a criticism will set the stage for the next chapter, where I offer a positive response to the generality objection.

## **7.1 The Generality Objection**

Before showing how TM can be extended beyond belief, I will first say a bit more about what constitutes the generality objection. Given that the objection may be formulated in at least two different ways, as I will now demonstrate, it is important for us to clarify which of these formulations our investigation will focus upon.<sup>3</sup> Here I will consider two different versions.

### **7.1.1 Generality Objection One: Perceptual Beliefs**

The first generality objection can be formulated as follows.

---

<sup>3</sup> The generality objection applies to both rationalist and empiricist interpretation of TM.

*Generality objection-perceptual beliefs* (hereafter, ‘GO-PB’): There is a class (or set) of mental states that one cannot acquire knowledge of by following TM. One will not be able to acquire knowledge of one’s perceptual beliefs—e.g., one’s belief that one is in pain; one’s belief that one is feeling hungry; or one’s belief that one is hearing the Lyre Bird sing—by following TM

Dorit Bar-On is an exemplar of this view. She raises this type of objection in her discussion of the ‘limited applicability’ (2004, p. 136) of TM. Bar-On (2004, pp. 136–137) claims that while TM could conceivably work for some beliefs—since when one believes that *P*, one typically endorses the content of *P*—and even to some other propositional attitudes with intentional content (e.g., desires, intentions, and so on), she does not think that TM can be easily extended to one’s sensations, or one’s perceptual beliefs.

Bar-On asks ‘what commitments would be involved in avowing hunger, thirst, non-specific rage or joy, pain, and so on? None, it seems’ (2004, p. 137). Bar-On’s point is that the commitment requirement (the accepting of *P*, the judging that *P*, the deliberation of whether or not *P*, the making up of one’s mind about *P*) that proponents of the rationalist interpretation of TM propose, is poorly suited to explaining the way in which one attains knowledge of one’s perceptual beliefs.<sup>4</sup> To put Bar-On’s point more succinctly, when one believes that one has a headache, one does not typically have to deliberate, or commit oneself to the

---

<sup>4</sup> It is important to note that Bar-On (2004, p.136) is specifically targeting Moran’s (2001) transparency-as-commitment model here. Such a requirement is explicated by Moran as the following: ‘as I conceive of myself as a rational agent, my awareness of my belief is awareness of my commitment to its truth, a commitment to something that transcends any description of my psychological state’ (2001, p. 84). As I will show, not all TM theorists need to be committed this view.

truth of any facts about the world—one just has the experience of the headache. Bar-On’s claim, then, is that in excluding perceptual states from the range of states that can be known by following TM, TM appears to have quite a limited applicably.

A proponent of TM can respond to Bar-On’s objection in two ways. First, they could say that while the criticism itself is sound, the objection, as a whole, fails because it does not apply to all interpretations of TM. Proponents of the empiricist interpretation of TM (see, e.g., Byrne 2005a, Fernández 2013) who construe TM in broadly empiricist terms, will be untroubled by Bar-On’s objection because they do not endorse the rationalist’s commitment requirement. Alex Byrne, for instance, who is a proponent of the empiricist interpretation of TM, makes essentially the same criticism as Bar-On with respect to the rationalist interpretation. In an example concerning the seeing of a cat, he says, ‘*seeing the cat* is not in *any* sense a matter of making up one’s mind, or “coming to some sort of resolution”’ (2005a, p. 85). Byrne’s point here is that one does not need to commit oneself to any facts about the world, or deliberate about the nature of cats to know that one believes one is seeing a cat—one just has the experience. Unlike Bar-On, however, Byrne only takes this to be a problem for the rationalist interpretation of TM, rather than with TM itself.

Since Bar-On does not distinguish between the rationalist and the empiricist interpretation of TM, she fails to account for the fact that not all TM theorists will appeal to the notion of commitment. Thus, Bar-On’s criticism has limited application—that is, it will only affect rationalist accounts of TM. Since I agree with Richard Moran (2001) and Boyle (2009) that commitment plays an important role in the acquisition of some types of self-knowledge, and thus

disagree with Byrne that rational agency can be altogether discounted from self-knowledge, such an objection is still relevant to the position that I am proposing.

How, then, should a proponent of the rationalist interpretation of TM respond to Bar-On's objection? In my view, one should agree with Bar-On, and also Byrne, that the commitment requirement is not applicable to perceptual beliefs, but disagree with them that this fact is problematic for the rationalist interpretation of TM. Here one can say that since my sensations, and the beliefs that arise from them—such as my belief that I am in pain, or my belief that I am seeing a plane in the sky—simply just occur to me, without deliberation, it is not surprising that rational agency does not contribute to my self-knowledge of them. The idea here is that perceptual beliefs do not fall under the purview of the rationalist theory. Thus, all that Bar-On and Byrne's criticisms succeed in showing, is that there are certain limits to the rationalist interpretation of TM, as opposed to there being an inherent problem with the theory itself. Thus, I think GO-PB fails, because rationalism is consistent with the claim that one can have knowledge of one's perceptual beliefs, without appealing to rational agency or commitment.

Bar-On anticipates such a reply, and argues that it will not work because this would mean that we would have to accept a non-uniform account of self-knowledge: a rationalist account for some mental states and an empiricist account for others. According to Bar-On, a good theory of self-knowledge is one that applies to '*both* intentional and non-intentional avowals alike' (2004, p.144). In other words, Bar-On thinks that by conceding that the rationalist theory cannot be extended to perceptual beliefs, a defender of the rationalist interpretation of TM would need to embrace a hybrid-theory—a thesis that she thinks is

untenable. Following Matthew Boyle, let us call the demand that a theory of self-knowledge must be uniform the *uniformity assumption*. According to Boyle, this is ‘the demand that a satisfactory account of our self-knowledge should be fundamentally uniform, explaining all cases of “first-person authority” in the same basic way’ (2009, p. 141).

If we choose to accept this assumption, we may respond to the above concerns in two ways, as Boyle himself points out. First, we could say that *all* introspective knowledge is achieved in ways that are broadly rationalistic. This would be problematic, for reasons I have given above, as it is not clear how perceptual beliefs, for example, can be subsumed into the rationalist account. The second option would be to say that *no* self-knowledge is achieved in ways that are broadly rationalistic. This too, however, is problematic for those wanting to defend rationalism—as it amounts to rejecting rationalism.

One way out of this dilemma is to simply reject the uniformity assumption. This would involve rejecting the thesis that an account of self-knowledge *must* be the same for sensory, or perceptual, mental states as it is for ‘intentional states’ or propositional attitudes. The pertinent question we need to ask then is ‘Are there any sound reasons for accepting the uniformity assumption?’ In my view, there are not. While I grant that a completed account of self-knowledge could turn out be uniform in its nature, this is not a restriction we should impose upon a theory of self-knowledge from the outset, even if it would make the theory more parsimonious. Although I accept the view that parsimony, or simplicity, is a theoretical virtue, a view’s simplicity only becomes virtuous when that theory can explain a set of a data equally as well as, or in a way superior to, a more complicated theory. Without first examining the set of

data that a theory purports to explain, simplicity is something that cannot be privileged.<sup>5</sup>

In summary, GO-PB is only a problem for the rationalist interpretation of TM, if we assume that a theory of self-knowledge must be uniform. I have suggested that this assumption is not warranted without first looking at the data.

### **7.1.2 Generality Objection Two: Propositional Attitudes**

I will now turn to the second formulation of the generality objection. This is the objection that TM cannot be extended beyond belief. A more detailed exposition of the objection can be given as follows.

*Generality objection-propositional attitudes* (hereafter, ‘GO-PA’): TM cannot be generalised from belief to account for the self-knowledge that one can have of other types of propositional attitudes—such as one’s desires, intentions, wishes, and so on.

GO-PA differs from GO-PB in that it is applicable to both interpretations of TM. If TM is only capable of giving a subject knowledge of some of their own beliefs, it cannot offer us a good general theory of introspection. In what remains of this chapter, it will be my aim to address this objection.<sup>6</sup> I will argue that, at best, the reasons typically given for upholding GO-PA point towards a lacuna in

---

<sup>5</sup> As Boyle (2009) points out, many have made the uniformity assumption without providing much in the way of argumentation for it. Boyle follows Kant (([1781] 1998)) and Donald Davidson (1984), who, Boyle notes, did not make the uniformity assumption.

<sup>6</sup> Gordon call this the ‘belief only’ (2007, p.155) objection

our understanding of TM, rather than to any inherent problem with the view itself.<sup>7</sup>

The first example of GO-PA that I will consider is one given by Shaun Nichols and Stephen Stich (2003). Despite the fact that Nichols and Stich think there are some virtues of TM, they do not think that TM can be generalised to account for the self-knowledge one can have of one's non-belief propositional attitudes.<sup>8</sup> With specific reference to Robert Gordon's (1996) 'self-assent routine'—Gordon's terminology for what we have called TM—Nichols and Stich argue that,

[t]here is no plausible way of recasting these questions so that they are questions about the world rather than about one's mental state. As a result, the assent routine strategy [TM] strikes us as clearly inadequate as a general theory of self-awareness (2003, p. 194).

The claim that Nichols and Stich are making in this passage is that the transparency thesis—the thesis that one can acquire knowledge of one's own psychology by attending to a question about the world—cannot account for non-belief mental states, because no 'worldly' questions exist that correspond to what one desires or what one intends to do, for example. The same objection is made by David Finkelstein, who says the following:

it is difficult to claim that the self-ascription of belief provides a model of self-knowledge that can be used in order to understand our awareness of our own, say, desires because there seems to be no "outward directed" question that bears the kind of relation to "Do I want X?" that the question "Is it the case that *p*?" bears to "Do I believe that *p*?" (2003, p. 161).

---

<sup>7</sup> A more radical view is that there are no mental states at all that can be known by following TM. I am not aware of any philosopher in the literature who has explicitly advanced this view.

<sup>8</sup> Nichols and Stich are not clear about what these virtues are, however.

Similarly, Peter Carruthers says, '[it] is unclear how such accounts could generalize to many other types of attitude besides belief and judgment' (2011, p. 84).<sup>9</sup>

According to some philosophers, then, TM cannot be extended from beliefs to other propositional attitudes because there is no plausible way to extend TM; it is difficult to claim that TM could be extended; and it is unclear how TM could be extended. While I do not find such reasons, by themselves, to be strong reasons for upholding GO-PA, I do think they provide a significant challenge to the TM theorist, by placing the burden of proof squarely upon them.

Before attempting to respond to such a challenge, I will offer some thoughts on why I think that this objection has been found so appealing in the first place. Although I do think that TM can be extended beyond belief to other propositional attitudes, I also think it is important to recognise the extent to which beliefs are representative of these other attitudes, and the extent to which they are not. As Sarah Paul has recently pointed out, some attitudes, such as intention, are not, at first glance, ideally suited to be explained by TM. She says that the

leading theories [TM theories] were not developed with intention specifically in mind, and I do not think they well account for it. The tendency of theories of self-knowledge either to treat all mental state types indiscriminately or to focus on belief and

---

<sup>9</sup> Carruthers—like some other proponents of GO-PA—still grants that following TM will sometimes yield self-knowledge. For instance, he states 'there is one class of judgments for which an outward-looking account [by the term 'outward-looking account' Carruthers is referring to TM] really can work' (2011, p. 83). For Carruthers, this is the set of mental states best referred to as 'perpetually-embedded judgements' (2011, p. 83). Recall from our discussion in chapter 3, that these are judgements, or beliefs, about perceptual content, such as the belief that one is seeing the waves crash upon the shore, or one's belief that one is hearing the lyrebird, and so on. Carruthers thinks that anyone who sees, for example, that a toy is broken will immediately be able to have knowledge about one of their beliefs—namely, that one believes that the toy is broken. It is interesting to compare Carruthers' criticism, that TM can only apply to sensory beliefs, with Bar-On's criticism—namely, that there is no way to extend TM to sensory beliefs. The existence of these two very different objections show why this chapter's attempt to clarify foundational issues about extending TM beyond belief is important.

suppose that a similar story must apply to intention has led to the epistemology of intention getting short shrift (2012, p. 328).

Now, Paul is not claiming in this passage that TM cannot be extended to intentions, but rather, she is pointing out that unless such differences between the various attitudes are adequately accounted for, we will not be in a position to formulate a complete theory. I suspect that this is one of the main reasons why TM's critics have sought to dismiss it, and why they think it cannot account for other states.

What we must acknowledge, then, is that while belief may indeed be representative of these other states, it is, in a key respect, unique amongst propositional attitudes. This is because beliefs, unlike attitudes such as desire and intention, aim at truth.<sup>10</sup> My belief that the moon orbits the earth ought to line up with the fact that the moon orbits the earth. My belief will be true if and only if the moon actually does orbit the earth; if not it will be false. As John Searle has pointed out, '[b]eliefs...are true or false, depending on whether the content of the belief matches an independently existing reality' (2001, pp. 36–37). This relationship is captured well in instances where one is seeking to link one's judgment that a fact about the world true with a corresponding psychological state, such as one's belief. However, the same sort of attending to the world does not appear applicable to intentions, desires, and other attitudes. This is because such mental states have different conditions of satisfaction, or different direction of fit (see, e.g., Searle 2001, p. 37; Velleman 2000). Desires, intentions, wishes, and hopes, do not aim at some truth that is out there in the world, but are rather

---

<sup>10</sup> The claim that beliefs 'aim at truth' is from Bernard Williams (1973). This is the orthodox view, held by many contemporary philosophers. We need not be committed to the claim that this is the *only* aim of belief. For example, Daniel Whiting (2012) in his article 'Does Belief Aim (Only) at Truth?' thinks there might be other aims of belief, which are often neglected by philosophers.

are *fulfilled* or *frustrated* depending on whether, in actuality, they are realised or not. Neither my desire to eat dinner at a Thai restaurant, nor my intention to finish writing this chapter, are facts out there in the world that I can judge to be the case or not. Nevertheless, as I will go on to explain, there are still outward-directed questions that one can attend to, in order to achieve knowledge of such states—some of which will be sensitive to rational deliberation.

A second, and slightly different, argument for upholding GO-PA is given by Eric Schwitzgebel. Like the above authors, Schwitzgebel concedes that TM can *sometimes* yield self-knowledge. He says ‘[f]or some of our attitudes I am inclined toward a version of what is sometimes called a “transparency” view’ (2012, p. 190). Schwitzgebel thinks that TM can work for fairly simple attitudes, such as his belief that ‘it doesn’t rain much in April in California’ (2012 p. 191).<sup>11</sup> However, for more complex mental states, Schwitzgebel is sceptical that TM can yield self-knowledge.

For instance, Schwitzgebel thinks that we have poor knowledge of the attitudes that pertain to our central values. One example he gives is sexism. Many men in academia, he claims, profess that men and women are equal. And yet their behaviour seems to suggest that they do not believe this claim. The reason that this is problematic for TM is because, as Schwitzgebel points out, such subjects seem to judge that the sexes are equal, and yet behave in ways that suggest they do not believe this.<sup>12</sup> If, as TM states, one can determine one’s belief about *P* by judging that *P*, why would such a mismatch occur? Although this clearly a problem for TM—an objection we called the matching problem in

---

<sup>11</sup> This scepticism differs from Carruthers’, because the belief that it doesn’t rain much in California in April is not a perceptual belief.

<sup>12</sup> Schwitzgebel cites data from Sally Haslagner (2008), in order to substantiate this claim.

chapter 6—it is a problem that faces *all* theories of introspection. This is because TM, like any other theory of self-knowledge, should not be not committed to an infallibility thesis. So, Schwitzgebel’s objection, while certainly an issue, is not unique to TM.

We have looked at two distinct arguments for GO-PA. With respect to the first argument, which stated there is no plausible way of extending TM beyond belief, I claimed that a TM theorist can respond by providing an account of how such an extension is possible. With respect to the second argument—Schwitzgebel’s argument that TM can only work for simple beliefs—I have claimed that no account of introspection should embrace an infallibility thesis, and so TM is compatible with the fact that sometimes it will fail. I will now begin to show how, contra GO-PA, TM can be extended beyond belief. The first step in this task will be to address issues pertaining to the scope of rationalism.

## **7.2 The Scope of the Rationalistic Interpretation of the Transparency Method**

Now that we have identified which formulation of the generality objection our investigation will be focused upon—GO-PA—we can now move on to discuss the prospects of extending TM. The first step of this task requires us to consider the scope of the rationalist interpretation of TM.<sup>13</sup> This issue is important to address, as was shown above, because the rationalist interpretation of TM will not be applicable to every type of propositional attitude one is capable of being in. Recall that one’s perceptual belief that one is seeing the waves crash upon the

---

<sup>13</sup> Although I shall only be discussing rationalism with respect to TM, what I will have to say here will apply equally to rationalists who do not endorse TM (see, e.g., Burge 1996).

shore, is not something that will require the exercise of rationality. So, while I agree with Robert Gordon, who writes that there is ‘in fact, an algorithm for generating, for each propositional attitude type, including hope, a corresponding type of lower level utterance that may be used as input to an assent routine [(‘assent routine’ is another term to describe ‘TM’)]’ (2007, p.156), we still need to determine what the scope of rationalism is. That is, we need to answer the question: ‘What set of mental states is the rationalist interpretation of TM applicable to?’ Let us call this the scope question for rationalism.<sup>14</sup>

In what follows, I will provide an answer to the scope question for rationalism, by reintroducing the concept of judgement-sensitive attitudes. (Recall that this was briefly discussed in chapter 5.) By the term ‘judgement-sensitive attitudes’, I mean, following Thomas Scanlon (1998), the set of attitudes that are sensitive to rational deliberation and evaluation. Such mental states, in contrast to standard perceptual experiences (e.g., my seeing of my computer screen while sitting at my desk), require that one has *reasons* for holding or discarding them.<sup>15</sup>

---

<sup>14</sup> Rationalists who endorse TM cannot simply answer ‘all propositional attitudes’ to the scope question for rationalism because not *every* propositional attitude, of the familiar types we have been discussing (e.g., beliefs, desires, intentions, hopes), will require adherence to rational norms. One will, as I will argue below, be able to acquire knowledge of *some* of one’s own propositional attitudes in ways that are broadly empiricist. For example, I do not need to adhere to any rational norms to acquire knowledge of my desire to drink a cold glass of water on a hot day. Like a standard sensory experience, I just seem to have it. In addition to there being a scope question for rationalism, there is also a question of scope pertaining to the nature of propositional attitudes—namely, the question about what mental states count as propositional attitudes. For instance, in the recent literature on propositional attitudes, there has arisen some controversy over the nature of perception. Tim Crane for example, has recently asked the question ‘is perception a propositional attitude?’ (2009, p. 452). Although he does not think it is, he points to other philosophers, such as John McDowell (1994) and Alex Byrne (2005b), who claim it is. McDowell and Byrne argue that perceptual experiences contain propositional content and, thus, perception should be thought of as a propositional attitude, alongside intentions, beliefs, desires and so on. On their view, my seeing the aurora borealis would be classified as a propositional attitude. We do not need to be concerned with this debate here, however, because the TM theorist is not committed to view that TM can only be extended to propositional attitudes—whatever that set turns out to consist of. That is, the question of whether TM is applicable to perception does not rest on the question of whether perception is a propositional attitude or not.

<sup>15</sup> I should mention that Scanlon himself is not directly involved in the debate about TM.

In chapter 5, I said that because such mental states have this normative component to them, it is plausible to suppose that they constitute the set of mental states that fall within the scope of rationalism. I say only *plausible* here because is it, obviously, a further claim—one that requires further argumentation—that rational norms make an *epistemic* contribution to the knowledge of such mental states. Empiricists, such as Brie Gertler (2016), would agree that such mental states carry with them this normative component—that is, she claims that ‘*beliefs should conform to evidence*’ (2016, p. 14)—however, she thinks that one can have *knowledge* of them in ways that are broadly empiricist.<sup>16</sup>

In what remains of this chapter, I will lay the groundwork for my defence of the claim that the scope of rationalism consists of the set of judgement sensitive attitudes. I thus aim to challenge empiricists like Gertler (2011b, 2016) and Byrne (2005a) who think that such mental states can be known in broadly empiricist ways. First, let me say a bit more about what is meant by the term ‘judgement-sensitive attitudes’ (hereafter, ‘JSAs’). Following Scanlon, we can give a more precise description as follows:

(JSAs): Judgement-sensitive attitudes constitute the class of attitudes for which reasons, in the standard normative sense, can be asked. A judgement-sensitive attitude is one in which it will make sense to ask *why* a subject has that attitude. It is one that will require the subject to have reasons for holding or discarding that attitude.<sup>17</sup>

---

<sup>16</sup> This may seem like a small difference, but much hangs on it. As I will describe in the next section, the epistemic contribution that rational agency provides to self-knowledge can be seen as the defining characteristic of rationalism.

<sup>17</sup> For this characterisation, I have drawn upon Scanlon’s discussion of JSAs that takes place in his book *What We Owe to Each Other* (see 1998, pp.18–21).

The locution ‘reasons in the standard normative sense’ refers to reasons in the sense pertaining to agency and responsibility. We can contrast this usage of reasons with ‘reasons in the purely *descriptive* sense’, which can be thought of as explanatory, which is to say historical, reasons for why someone holds an attitude. For example, the purely descriptive reason for why Andrew believes that Pluto is no longer a planet may be because his schoolteacher told him so. Reasons in the standard normative sense, in contrast, are caught up with the subject’s agency—that is, it is incumbent on a subject to seek justification, or suitable grounds, for holding a certain attitude.

Let us consider the attitude belief—to take a familiar example—in order to clarify this conception. Beliefs are JSAs because a subject who believes something will be required to act in certain ways with respect to that belief.<sup>18</sup> As Scanlon says, ‘a person who believes that *P* will tend to have feelings of conviction about *P* when the question arises [and] will normally be prepared to affirm *P*’ (Scanlon 1998, p. 21, my emphasis). Someone who believes that *P* will also be required to provide justification for why they believe that *P*; and be willing to modify their belief about *P* if counterevidence is provided. Suppose I believe that there is a black hole in the middle of the Milky-Way Galaxy. What does having this belief require of me? Obviously, I should be prepared to answer in the affirmative when I am asked whether I think there is a black hole at the centre of the Milky-Way Galaxy. But I should also be prepared to use this belief as a premise for further reasoning. Given that I believe that there is a black hole outside of our solar system, I can reason that there must be more to the universe than just our solar system. I should also be on the lookout for counterevidence to

---

<sup>18</sup> Not all beliefs, of course, are JSAs. My belief that I am seeing the sunset doesn’t seem tied to rationality.

my belief: such as peer-reviewed scientific papers that purport to overturn the claim.<sup>19</sup> JSAs, as Scanlon puts it, are:

attitudes that an ideally rational person would come to have whenever that person *judged there to be sufficient reasons* for them and that would, in an ideally rational person, extinguish when that person judged them not to be supported by reasons of the appropriate kind' (1998, p. 20 my emphasis).

So, when I have reasons, such as scientific evidence, for thinking that there is more to the universe than our solar system, my belief that our solar system constitutes the universe should extinguish, because I have sufficient reason to abandon the belief.

The same principle holds for many other types of mental states. Take intentions, for example. Having an intention to attend a film on the weekend, for example, *is not* something that I can be completely indifferent towards. As Scanlon puts it, 'a person who intends to do A will not only feel favorably disposed, on balance, to that course of action, but will also tend to be on the lookout for ways to carry out this intention' (1998, p. 21). Suppose I have an intention to attend the New Year's Eve celebrations in Times Square. Typically, if I actually do have this intention, I would spend the preceding months working out flight routes to New York; ask for time off work in December; and be on the lookout for conflicts in my schedule. If I were to book a flight to Manila for New Year's Eve; not bother to even look at flights to New York; and show complete

---

<sup>19</sup> This seems to make the process of forming a belief a very cognitively demanding task—one that seems to put belief out of reach of young children and non-human animals. One should not get this impression, however. While I think that young children and non-human animals can have beliefs, I think it is highly unlikely that they can have judgement-sensitive beliefs. That is, I find it implausible to think that their beliefs would be sensitive to rational deliberation. Harry Frankfurt, in an influential paper, 'Freedom of the Will and the Concept of a Person', makes a similar distinction. He thinks that adult humans' 'capacity for reflective self-evaluation that is manifested in the formation of second-order desires' (1971, p. 7) makes them morally responsible for their actions.

indifference to the idea of going to New York, it would be difficult to say that I actually did have the intention to spend New Year's Eve in Times Square.<sup>20</sup>

The same is true of many other types of attitudes, such as desires, hopes, wishes and so on. I do not say *all*, however. As mentioned above, there will be some attitudes that do not require any form of deliberation or rational assessment on the part of the agent—such as in the case of perceptual beliefs. It does not make sense, typically, for one to question why one has a belief that one is currently seeing a red ball. I do not mean a casual explanation such as: 'because I opened my eyes and it was in front of me'. I mean, rather, it does not make sense to ask the kinds of normative questions of the sort we have been discussing above. The perceptual experience of seeing the red ball is the only justification one requires for holding such a belief.<sup>21</sup>

Following Moran, I will call such mental states 'brute' attitudes (2001, p. 115). The term 'brute' will be recognised by those familiar with the idea of brute facts. Such facts, as commonly described in metaphysics and epistemology, are facts that can be explained without any appeal to any further facts; meaning they do not require any further explanation and are, thus, seen as fundamental. Given that we can explain why we experience redness, or pain—such as the presence of a red apple, or because I tripped and fell down the stairs—calling such attitudes *brute* may seem like a misnomer. In another sense, though, it is quite appropriate. Our feelings of pain, or the hearing of the rain as it falls on the tin roof, are experiences that happen to us, and we do not need to appeal to any other facts to have knowledge of them; we are, in a sense, simply observers of the 'cognitive

---

<sup>20</sup> Following Michael Bratman, I take the common-sense notion of an intention to be 'inexorably tied to the phenomena of plans and planning' ([1987] 1999, p. 2)

<sup>21</sup> This will be the case even if the red ball is an hallucination—it is the experience that is important here.

show'.<sup>22</sup> As Moran says, such experiences are '[l]ike an alien intruder, they must simply be responded to '(2001, p. 114).<sup>23</sup>

This brute attitude/judgement-sensitive attitude distinction is also applicable to other attitudes such as desire. In order to elaborate upon this point, let us consider a distinction that Thomas Nagel makes in his book *The Possibility of Altruism* between 'motivated' and 'unmotivated' desires.<sup>24</sup> Nagel claims that while certain desires are arrived at by deliberation, there are some that can be said to simply occur within us. Hunger, for example, is a desire that simply assails us—that is, we just seem to experience it. Even though we can explain its presence by a lack of food, it does not arise from any deliberation, nor is it sensitive to reasons for holding or discarding it. Thus, Nagel calls it unmotivated. Other desires, Nagel claims, 'need not simply assail us' ([1970] 1978, p. 29). The desire to shop for groceries *because* there is nothing in the fridge is, in contrast, a motivated desire because it is sensitive to deliberation.<sup>25</sup> It is sensitive to beliefs I have about the world, and can potentially be exhausted, such as when I judge there to be no reason to shop for groceries because the fridge is full.

Not all commentators, however, seem to fully appreciate this distinction. Neil Sinhababu, for example, writes in a way to suggest that all of our desires are brute.

You might think it's shameful and stupid to not desire someone because of height, weight, race, age, hair color, or past romantic history. But no reasoning can lead you from this wishing and believing to desiring. A process that changed our desires this way would probably count as reasoning. But humans can't do it (2017a, p. 40).

---

<sup>22</sup> I borrow this expression from Gertler (2016, p.1)

<sup>23</sup> For more on brute facts see Eric Barnes (1994).

<sup>24</sup> As Nagel himself notes, this was pointed out by Aristotle in his *Nicomachean Ethics*, Book 3, Chapter 3.

<sup>25</sup> This example is Nagel's ([1970] 1978, p. 29)

In my view, Sinhababu's claim about human rationality is far too pessimistic. Even if Sinhababu is right that desires such as these cannot be altered by reasons—a big if—it does not follow there are no judgement-sensitive desires. There are trivial everyday examples that show that there are. Consider my desire to fill the car up with petrol, for example. If I come to believe that in actual fact my car's tank is full, then my desire to fill my car up with petrol should evaporate when I come to form this belief; just as is my desire to go food shopping should evaporate if my belief that there is no food in the fridge changes. Motivated desires are sensitive to what I believe, and show that there is sometimes a role for reasoning to play when it comes to what I desire. What I think Sinhababu's passage usefully shows is how just radical it is to eschew rational agency from desire. The question whether empiricists are committed to such a view, is something I consider in the next chapter.<sup>26</sup>

Nagel's motivated/unmotivated distinction, along with Scanlon's judgement sensitive/non-judgement sensitive attitude distinction, helps to highlight the fact that some attitudes are sensitive to rational deliberation. One cannot be indifferent, or unreceptive, to reasons for why one is in possession of some attitudes. Let us summarise this distinction as follows. There are (i) *judgement-sensitive attitudes*—attitudes which require one to have reasons, in the standard normative sense, for why one has them; and (ii) *brute attitudes*—attitudes for which reasons, in the standard normative sense, cannot be legitimately asked for. With this distinction now in place, I want to return to our

---

<sup>26</sup> Such a view is consonant with Hume's account of belief. Hume says 'belief is more properly an act of the sensitive, than of the cognitive part of our nature' (T 1.2.2.3; SBN 183); and 'when the mind...passes from the idea or impression of one object to the idea or belief of another, it is not determined by reason' (1.3.6.12; SBN 92).

discussion of rationalism that was first brought up in chapter 5. Following Brie Gertler, I said that rationalism is constituted by the following two claims:

the normative features of rational agency: (i) make a crucial epistemic contribution to knowledge of one's attitudes, one that is (ii) irreducible to the epistemic factors (evidence, reliability, etc.) countenanced by empiricism (2011a, p.258).

It is my view that we are capable of procuring knowledge of our JSAs only *because* of our ability to be rational agents—that is, because we have the capacity to hold mental states for reasons. I thus agree with condition (ii), which says that the methods of empiricism—such as evidence and observation—will sometimes not be enough to procure knowledge of our attitudes. Relying on such factors would make us mere ‘observers of a passing cognitive show’ as Gertler (2016, p. 1) puts it.<sup>27</sup> For reasons I went into above, being a mere observer fails to do justice to the fact that some propositional attitudes are mental states we hold *for reasons*.

In order to help summarise what has been claimed so far in this chapter, let us consider the following table:

---

<sup>27</sup> André Gallois (1996, p. 127) makes a similar distinction, when he notes that some mental states are subject to the will.

**Table 3: The Scope of the Transparency Method**

| Self-Knowledge via the Transparency Method                          | Rationalism | Empiricism |
|---|-------------|------------|
| Sensory Mental States (e.g., pain, perception, hearing).            |             | ✓          |
| Judgement-Sensitive Attitudes (e.g., beliefs, intentions, desires). | ✓           |            |
| Brute Attitudes (e.g., brute desires, perceptual beliefs).          |             | ✓          |

Table 3 helps to summarise the main points of this section, which can be thought of as my answer to the scope question. Each row represents one of the three different classes of mental states that I have discussed so far; and each column gives my answer to the question of whether or not that class of mental states can be known in ways that are either broadly rationalist or empiricist. The table also helps to map some disagreements that have arisen so far. Early in this chapter, I disagreed with Bar-On that an account of self-knowledge must be uniform. The presence of the ticks in different columns represents this disagreement. Another disagreement which can be seen in this table is the one I had with Byrne about the rationalist versus empiricist interpretation of TM. Although I agree with Byrne that one can have knowledge of one's mental states by using TM, I disagree with him that this can *always* be done in ways that are broadly

empiricist. This disagreement is shown by the fact that I have placed a tick in the rationalism column on the JSA row.

Now that my view about where the scope of the rationalist view lies has been laid out, I will move on to provide argumentation for the claim that TM can actually be extended to such states.

### **7.3 The Prospects of Extension**

So far, I have put forward a view about where I think the scope of rationalism lies. I have not, however, offered much in the way of argumentative support for the claim that TM can actually be extended beyond belief. Given that some authors deny that TM can be extended in such a way (recall our discussion in §7.1.2), a satisfactory response to the generality objection will not have been offered until an account of such an extension has been provided. In what remains of this chapter, and in the next, I will attend to this task.

Although I am certainly not the first to address these issues, I should note that, at present—and to the best of my knowledge—no complete account has been given in the literature. The generality objection, therefore, points to somewhat of a lacuna in the transparency theory—one that is recognised by some of the theory’s leading proponents. Akeel Bilgrami, for example, in his book *Self-Knowledge and Resentment*, largely focuses on the attitudes belief and desire and says that ‘[e]xtending the account given here to the other intentional states and eventually to qualitative states of mind is an important task but must await another occasion’ (2006, p.1). And Richard Moran holds that, while belief is ‘representative’ (2012, p. 214) of the way in which we acquire knowledge of

our attitudes in general he, much like Bilgrami, deals mostly with desires and belief. Others have, in the last few years, focused on giving an account of individual attitudes like *intention* (see, e.g., Paul 2012) and *wishes* (see, e.g., Barz 2015).

Let us call this the *incompleteness problem* for TM. Given that our space here is limited, and the issue complex, I do not pretend to offer the final word on the matter. Instead, I will concern myself with two, more manageable, objectives. First, I will look at some individual mental states and show how one can achieve knowledge of them by following TM; and second, I will show that many of these types of mental states will require adherence to rational norms, thus, providing further support for rationalism.

Let me now finish this chapter by offering some critical remarks on one potential solution to the incompleteness problem, which has been discussed recently by Quassim Cassam in his book *Self-Knowledge for Humans*.<sup>28</sup> Although Cassam does not himself endorse this solution, he does consider it a *potential* way of responding to what he calls the ‘Generality Problem from Rationalism’ (2014, p.103).<sup>29</sup> Cassam claims that *rationalism*—by which he is really referring to the rationalist interpretation of TM—could be extended beyond belief, to other attitudes, by attending to various ought questions.<sup>30</sup>

In order to assess this approach, let me first explain how Cassam construes the rationalist interpretation of TM—as this is important for understanding how his solution works. He claims that one can determine the

---

<sup>28</sup> See chapter 9 of Cassam’s book *Self-Knowledge for Humans* (2014) titled ‘Looking Outwards’.

<sup>29</sup> The approach that Cassam discusses is one originally formulated by David Finkelstein (2012). Cassam clearly has the rationalist interpretation of TM in sight here. However, it is not clear whether he thinks that the empiricist interpretation of TM faces the same problem.

<sup>30</sup> I emphasise the term rationalism here because, as I said, there are other ways of defending rationalism, e.g., Burge (1996), which do not reference TM.

answer to a psychological question, e.g., ‘Do I believe that *P*’ by answering a question about what one ought to do. On his view, the question ‘Ought I to believe that *P*’ is transparent to the question ‘Do I believe that *P*?’<sup>31</sup> On this construal of TM, I could determine whether I believe that Neil Armstrong landed on the moon, by asking myself the question ‘Ought I to believe that Neil Armstrong landed on the moon?’ Given this model, Cassam thinks that TM might then be extended to other mental states by asking similar ‘ought’ questions. To get a better idea of how this method works, let us consider some examples. One can determine whether one intends to go to the football match, by asking oneself: ‘Ought I intend to go to the football match?’; one can determine whether one wishes that Manchester United will win the football match, by asking oneself the question ‘Ought I to wish that Manchester United will win the football match?’; one can determine whether one desires go to the football, by asking oneself ‘Ought I to desire to go to the football match?’; and so on.

Now, although this may indeed be a *potential* solution to the incompleteness problem for rationalism, it is, for reasons I will now go into, untenable. One problem, as Cassam himself points out, is that gaining knowledge of what one ought to desire, for example, is much harder than gaining knowledge of what one does desire. Attempting to determine whether one ought to desire to give 100 dollars to Oxfam, is a harder question than determining whether one actually does desire to give 100 dollars to Oxfam. The same is true for other types other types of self-knowledge. Thus, Cassam thinks this leads to what he calls the ‘Substitution Problem for Rationalism’ (2014, p.104). Cassam argues that only *homo philosophicus* (a perfectly rational human) could hope to achieve

---

<sup>31</sup> I should stress that Cassam himself does not actually accept this view.

self-knowledge of their mental states in this way. He concludes that TM does not offer us a compelling account of how most people can achieve knowledge of their mental states.<sup>32</sup>

I will respond to Cassam by first agreeing with him that such a method does indeed make self-knowledge of our attitudes look like a very onerous task—one that would put it out of reach of most people. It is true that determining what one ought to believe is a more difficult task than determining what one does believe. So, it would be hard to argue that this is our normal route to self-knowledge—given that we do achieve such self-knowledge. The main problem with this strategy, however, is that it is predicated upon a mischaracterisation of TM. So, while I think Cassam is right to reject this interpretation of TM, I do not think that such a rejection has any serious implications for either the rationalist or empiricist interpretation of TM. This is because Cassam has torn down a strawman—one that does not faithfully represent either version of TM.

To see why this is the case, let us recall our discussion of Moore's paradox, first given in chapter 5. Recall that we noted that there was something absurd about someone who judges that Canberra is the capital city of Australia and yet at the same time believes that the capital city of Australia is not Canberra. This absurdity was instrumental in constructing an argument for why we should accept TM. Notice, however, that such an absurdity does not arise when one's psychological mental state fails to line up with what one ought to do. To illustrate this point, let us consider the following conversation between Sally and Molly. Suppose Molly asks Sally if she intends to see the Rolling Stones in concert. Suppose Sally replies, 'Yes, I *do* intend to go. However, I ought not to

---

<sup>32</sup> This objection was looked at in the last chapter (see chapter 6.1).

go—I have a Philosophy exam in the morning that I really need to pass, which I have not yet studied for. However, I still *intend* to go’. Whatever we might say about Sally’s psychology here, there is nothing obviously absurd in what she says. Clearly, if she wants to pass the exam, then it would make sense for her to go home and study. Similarly, it would be sensible for smokers who wish to reduce their risk of getting lung cancer to give up cigarettes. Such people are not irrational, in the sense that they are accepting an apparent contradiction; they can be more accurately described as suffering from weakness of the will, or *akrasia*, which is, as Searle points out, as common as ‘wine is in France’ (2001, p. 10). Such a person differs from someone who judges that it is raining, and yet believes that it is not raining—someone who is more paradigmatically irrational.<sup>33</sup>

Such examples show that Cassam is really targeting a different, and implausible, version of TM. Recall that according to the way that I have defined TM, one can know what one believes about *P*, by attending to the question ‘Is it true that *P*?’ I argued, in chapter 5, that such a position is supported by the fact that it would be absurd for someone to say, ‘I judge that *P* is true, but I don’t believe that *P* is true’. This is because when one judges that *P*, one typically also believes that *P*. On Cassam’s construal of TM, however, one does not get such absurd sounding sentences. This shows that Cassam has offered us a mischaracterisation of TM. In the next chapter, I will give an alternative response to the generality problem that avoids the worries that Cassam brings up here.

---

<sup>33</sup> See Al Mele (2009) ‘Weakness of will and akrasia’ for more on the issue.

## **7.4 Conclusion**

In this chapter, I have examined two different formulations of the generality objection. I argued that only one of them—generality objection: judgement-sensitive attitudes—was relevant to the prospects of extending TM. I then gave an answer to the scope question for rationalism. I argued that the set of attitudes called judgement-sensitive fall within its purview. Next, I examined Cassam’s attempt to extend TM beyond belief. I argued that such an account was predicated upon a mischaracterisation of TM, and was thus untenable. In the next chapter, I provide an alternative account to Cassam’s, where I offer a positive account of how TM can be extended beyond belief.



## Chapter 8

### Extending the Transparency Method Beyond Belief—Part Two:

#### Rationalism versus Empiricism

The phenomena of self-knowledge...are themselves based as much in asymmetries of responsibility and commitment as they are in difference in capacities, or in cognitive access.  
Richard Moran (2001, p. 64)

In chapter 7, I considered the prospects of extending the transparency method (hereafter, ‘TM’) beyond belief to include non-doxastic mental states such as intentions, desires, wishes, hopes, and so on. Recall that, according to TM, one does not need to *look inside* of one’s mind to procure knowledge of one’s own mental states. One can, rather, attend to ‘outward-directed’ phenomena such as states of affairs, intentional objects, and reasons in favour of believing something. While many think that TM can account for the self-knowledge one can have of one’s basic beliefs, such as one’s belief that it is raining, fewer think that TM can be generalised to account for the self-knowledge one can have of other types of propositional attitudes—such as one’s desires, intentions, wishes, and so on.

In this chapter, I will challenge the generality objection by showing how TM can be extended to such mental states. By this, I mean there are instances of each category of mental state (such as desires, intentions, etc.) where TM is applicable. Since there are two distinct ways in which to interpret TM—a rationalist and empiricist way—I will also be concerned with comparing and contrasting the two approaches. Although it may be tempting to view the rationalist versus empiricist debate as one that is orthogonal to the task of

extending TM, I will show that this is not the case. I will argue that in order to adequately respond to the generality objection, this debate also needs to be taken into consideration.

Although both approaches to TM converge over what the right *method* of self-knowledge is, they diverge over the question ‘Can self-knowledge of our judgement-sensitive attitudes be achieved without invoking rationality?’ To anticipate, I will argue that the empiricist version of TM is beset by an objection that I will call the *passivity objection*. This is the objection that the empiricist interpretation of TM cannot account for the fact that judgement-sensitive attitudes are mental states that subjects hold for reasons.

I will proceed as follows. In §8.1, I give an exposition of the passivity objection. In §8.2, I lay out the details of Alex Byrne’s empiricist approach to TM—the main view I will be contrasting with the rationalist approach to TM. In §8.3, I show how Byrne’s account of judgement-sensitive belief is beset by the passivity objection. I then argue that the rationalist account avoids this objection. In §8.4, I do the same for desire. In §8.5, I do the same for intention. In §8.6, I consider how we might go about extending TM to other mental states, in a way that is compatible with the rationalist account.

### **8.1 The Passivity Objection**

The passivity objection is the objection that empiricist approaches to self-knowledge cannot account for the relationship that agents have between themselves and the judgement-sensitive attitudes that they hold. Such a relationship, as I suggested in the chapter 7, consists of an agent’s rational capacity for holding certain mental states for reasons—thus making them in an

important respect responsible for what they believe, desire, intend, and so on. In a recent discussion of this issue, Brie Gertler has succinctly articulated the objection as follows:

‘[e]mpiricist approaches to self-knowledge face an influential objection: that empiricism portrays us as mere observers of a passing cognitive show, and thereby neglects the fact that believing and intending are things we *do*, for *reasons*’ (2016, p. 1).<sup>1</sup>

As I have suggested in the previous chapter, such an objection will not apply to all mental states. The fact that one *passively observes or detects* one’s perceptual beliefs—e.g., one’s belief that one is seeing a black swan, or one’s belief that one is hearing the last call for drinks will not give rise to the passivity objection. This is because, with such states, it is not incumbent on agents to have reasons, in the normative sense, for why they have them. It is enough that one justifies one’s non-judgement-sensitive beliefs or desires with the passive experience itself.

In what follows, I will provide argumentative support for the claim that the passivity objection is particularly troublesome for the empiricist approach to TM. I will do so by focusing on Alex Byrne’s empiricist approach to TM. I will argue that his account of TM characterises judgement-sensitive attitudes as mere passive events that simply occur to us. I will suggest that this is not an adequate way to account for the self-knowledge of such mental states. Although Byrne’s view is, of course, not the only empiricist account to be found in the literature, we can, to a certain extent, treat his view as representative of the empiricist approach because the passivity objection picks out a problem that I think is

---

<sup>1</sup> See Gertler (2016) for a reply to the passivity objection. She does not find the objection convincing.

present in all empiricist accounts.<sup>2</sup>

One way that empiricists have responded to the passivity objection is to say that it is rationalism that mischaracterises the way in which the self-knowledge of our judgement-sensitive attitudes is obtained. Brie Gertler (2016), for example, thinks the passivity objection fails because empiricists can do justice to the normative requirements that certain mental states have. In her view, since human beings have a capacity for self-reflection, they can modify their beliefs and intentions when they judge necessary—thus giving them the power to abandon a certain mental state when reason requires it.

In what follows, I will argue this reply is unconvincing. Expanding upon some ideas recently expounded by Richard Moran (2012), I will argue that if one attempts to follow TM *without* appealing to one's own capacity for *rational agency*, one will be *estranged* from the mental state that one is seeking to gain knowledge of. So, for example, if one attempts to answer the question 'Do I believe that *P*?' by answering the question 'Is *P* true?' without appealing to rational agency, one will be forced to treat the latter question as if it were a brute stimulus. In other words, one will be forced to treat the question 'Is *P* true?' as something to be detected or observed, independent of the reasons one has for accepting *P* to be true. This is something that Moran argues—convincingly in my view—is at odds with the conception of what it means to believe something. This point is expanded upon by Moran in the following passage:

the “passivity of belief” is the reverse side of a person's rational agency as a believer, for it is because one's beliefs are the expression of one's rational relation to the world that they cannot be simply “chosen.” If what I believe were not answerable to the ways the world is, then I could indeed treat my beliefs as states

---

<sup>2</sup> For another empiricist account of TM, see Jordi Fernández (2013).

which I could seek to produce in myself for reasons unrelated to their truth (2012, p. 233).

Moran's point here is that by reducing the process of self-knowledge to one of simply choosing or detecting, one will not be able to account for the fact that one's judgement-sensitive beliefs are mental states we hold for reasons. In other words, we might say that a judgement-sensitive belief, and the reasons one has for holding that belief, are inextricable. While Moran mostly focuses on belief, I will argue that his main point can be generalised to other judgement sensitive attitudes—such as intentions, desires, hopes, wishes, and so on. If this is right, then the passivity objection can be thought of a general problem that faces *any* account of self-knowledge that attempts to account for judgement-sensitive attitudes without appealing to rational agency.

Although Moran has been influential in the past few years, by drawing attention to this issue, there are important philosophers who anticipated a similar view. Perhaps the most notable of these is Kant—as Robert Brandom says in the following.

One of Kant's big ideas is that what distinguishes judgments and intentional actions from the performances of nondiscursive creatures is that judgments and intentional actions are things we are in a distinctive sense *responsible* for, they are commitments of ours, they are exercises of *authority* on the part of their subjects. As such normative statuses, they are things our *rational entitlement* to which is always potentially at issue. That is, they come with a standing obligation to have *reasons* for them (2015, pp. 2–3).

What I wish to draw attention to in this passage is the idea of commitment. In chapter 5, I argued that it was right to think of judgement-sensitive belief as a form of a commitment. I suggested that what it means to believe a proposition is

to be committed to the truth of that proposition. What Brandom is suggesting here is that the idea of commitment is not something that is only applicable to belief—it can be applied to other types of self-knowledge too. Drawing upon this idea, I will argue that commitment is something that is involved with other judgement-sensitive attitudes, such as desire, intention, and so on. I will show that this poses a significant problem for empiricist approaches to TM.

In what follows, I will consider several different examples of judgement-sensitive attitudes. My plan is to compare and contrast the rationalist approach to TM with the empiricist approach. I will argue that the passivity objection arises for the empiricist approach to TM, when judgement-sensitive attitudes are considered. To keep this comparative task manageable, I will focus solely on TM, rather than comparing several other empiricist accounts of self-knowledge such as the inner sense view. Although this means ignoring the numerous other empiricist approaches in the literature, this is not necessarily a negative thing, as the passivity objection can be formulated quite broadly. Before this comparison can begin, I will offer a summary of Byrne's view.

## **8.2 Transparency, Empiricism and Rule Following**

In a series of recent papers, Alex Byrne (2005a, 2011b, 2012) has sought to extend the application of TM beyond belief to other propositional attitudes such as desire and intention. Because Byrne's view represents a broadly *empiricist approach* to extending TM, it is a challenge to the broadly *rationalist approach* I have been proposing. Whilst Byrne is not the only philosopher to have defended an empiricist approach to TM, I will direct my critical remarks of the empiricist approach primarily towards his view. There are two reasons for doing this.

Firstly, his view is, to my mind, the most developed empiricist account of TM to be found in the literature: Byrne accounts for several different individual types of mental states rather than just belief.<sup>3</sup> And secondly, since the passivity objection is quite general—such that it targets all broadly empiricist accounts—only one account will be needed to demonstrate just how the objection manifests. Although we are already familiar with how TM works (see chapter 5), Byrne’s interpretation is unique in that it accounts for transparency with what he calls a ‘rule following procedure’ (2005a, p. 12). To assess his view, I will first give a brief account of what this consists of.

Byrne’s main aim, to my mind, is to account for the phenomenon of transparency *without* involving rational agency. This can be surmised by the following passage, where he says ‘[t]here *is* a mechanism for detecting one’s mental states but...in an important respect it does not “resemble perception”’ (2005a, p. 80). Two points require further explanation here. First, the ‘mechanism’ that Byrne is referring to here is none other than TM. And second, Byrne thinks of TM as a form of ‘detecting’—which, while being distinct from *inner perception*, makes it broadly empiricist. He elaborates upon this point as follows:

‘[i]f this procedure [TM] can yield self-knowledge, and if it involves the (casual) *detection* of...belief...then this would be an instance of the “broad perceptual model” without being Ryleanism or the inner sense theory’ (2005a, p. 93).

The key to understanding this view is via what Byrne calls a rule following procedure. This can be understood by giving some examples of what Byrne calls ‘epistemic rules’ (2005a, p. 93). These are rules that, if followed,

---

<sup>3</sup> This is Byrne’s attempt to offer a response to the generality objection.

*tend* to produce knowledge. Such a rule can be thought of as a conditional with the following form:

‘R: If conditions *C* obtains, believe that [*P*]’ (2005a, p. 94, my emphasis).<sup>4</sup>

Byrne gives the following example of such a rule:

‘DOORBELL If the doorbell rings, believe that there is someone at the door’ (2005a, p. 94).

I agree with Byrne that DOORBELL is a good rule: anyone who follows it will typically form a true belief that someone is at the door. But how exactly does one follow it? One does by so by first recognising that the antecedent is true (the conditions *C* that obtains). In this case, *C* would be the fact that the doorbell is ringing within one’s earshot. After recognising that *C* has obtained, the next step would be to adopt the belief that there is someone at the door. These two steps are all that it takes for one to have successfully followed the rule. In Byrne’s view, DOORBELL will typically lead one to true beliefs about whether there is someone at the door.

All of this is not to say that the rule guarantees that one’s belief will be true. If one follows the rule, and self-ascribes the belief that there is someone at the door, then that belief could still turn out to be false. It may be that the doorbell is malfunctioning and thus one’s belief that someone is at the door will

---

<sup>4</sup> R stands for ‘epistemic rule’.

be false. So, DOORBELL will not always produce true beliefs if followed. Nevertheless, Byrne thinks that DOORBELL is still a *good rule*, however, because it ‘*tends to produce knowledge about one’s visitors*’ (2005a, p.94, my emphasis).

We can compare DOORBELL, a good rule, with a *bad rule*. A bad rule is one that tends not to produce true beliefs if followed. Consider the following, for example:

POLITICIAN: If a politician says *P* is the case, believe that *P* is the case.

POLITICIAN differs from DOORBELL because it would not tend to produce true beliefs, in the rule follower, in the same way that DOORBELL would. This is because following POLITICIAN will often lead one to form false beliefs—politicians often say false things, after all. This is *not* to say that one can *never* achieve a true belief about the world by following POLITICIAN—following it will sometimes produce true beliefs about the world—but it will not be as reliable as DOORBELL.

Now that the concept of an epistemic rule has been explicated, I will show how Byrne combines the concept of rule following with *transparency*. In Byrne’s view, epistemic rules can also account for the way in which one achieves self-knowledge of what one believes (as well as what one intends, desires, so on). Byrne proposes the following epistemic rule for belief, which he calls BEL, as follows:

‘BEL If [*P*], believe that you believe that [*P*]’ (2005a, p. 95).

The antecedent, *P*, which is described in BEL can be thought of as the condition *C* that obtains.<sup>5</sup> *P* stands for a proposition about the world. BEL says that if this condition obtains, namely *P*, then you should believe that you believe the proposition *P* is true. In other words, you are to self-ascribe the belief that *P* to yourself. Let consider an example. Suppose I want to know whether I believe that Canberra is the capital city of Australia. I can do so by following BEL. First I attempt to determine whether Canberra is the capital city. To do so I would typically judge or accept that Canberra is the capital city of Australia. Once I have judged that such a condition has obtained, I then attribute the belief to myself that I believe that Canberra is the capital of Australia. Such a process provides a way for me acquire self-knowledge of what I believe about Australia's capital.

BEL differs in an important respect from the previous two epistemic rules we have looked at—DOORBELL and POLITICIAN—because it is a psychological rule. This means that it purports to reveal *knowledge* about one's mind.<sup>6</sup> At first pass, this rule seems to be inferior to both previous rules we have examined. How can a fact about the world, such as the location of a nation's capital city, be a reliable guide for acquiring knowledge of an individual's mental state? As Matthew Boyle states, inferring a fact about one's own psychology from a fact about the world seems 'mad' (2011, p. 230).<sup>7</sup> The previous rules we have looked at—DOORBELL and POLITICIAN—which did not guarantee knowledge, at least provide a relevant connection from antecedent to consequent.

---

<sup>5</sup> The relation one stands in to condition *C*, that is analogous to hearing the doorbell, is that one is making a judgement—that is, one is accepting the truth of the antecedent.

<sup>6</sup> The other two rules, recall, were about gaining knowledge about who was at the door, or whether claims about the world were true.

<sup>7</sup> See Gallois (1996, p.76) for a discussion of the same point.

For example, in the rule DOORBELL, hearing the doorbell ring seems to be a good guide to the fact that someone was at the door because the two regularly go together. The same does not seem true of BEL, because facts about the world don't tend to link up in the same way with psychological states.

Byrne's response to this concern is to say that psychological rules can *only* work from the *first-person perspective*.<sup>8</sup> That is, BEL will only reveal knowledge of the mind of the person who is following the rule. So, the connection between facts about the world, and facts about a subject's mind is rendered congruous because it is the rule follower who is endorsing a fact about the world. Byrne states '[s]ince the antecedent of [BEL] expresses the content of the mental state that the rule-follower ends up believing she is in, [BEL] can be called a *transparent* rule' (2011a, p. 112).

Even though I think there is much to agree with in Byrne's interpretation of transparency, I will argue that his account is inadequate in making sense of judgement-sensitive attitudes.<sup>9</sup> This is because, as I will show, his view cannot account for the relationship that agents have between themselves and the judgement-sensitive attitudes that they hold. Recall that this was a problem I identified in §8.1 as the passivity objection.

The plan for the remainder of the chapter will be to look at some individual examples of judgement-sensitive attitudes and compare Byrne's approach to TM to the rationalist one I am developing. I will begin by looking at judgement-sensitive belief, before turning to judgement-sensitive desire, and then to intention. I will then close with a general strategy about how to extend the application of TM to other mental states.

---

<sup>8</sup> As was pointed out in chapter 5.

<sup>9</sup> I agree with Byrne, for instance, that such an account will work well for perceptual beliefs.

### 8.3 Judgement-Sensitive Belief

According to TM, one can acquire knowledge of one's own belief—a psychological state—by attending to an 'outward-directed' question about the world. Empiricist approaches to TM construe this process in terms of detection and observation, whereas rationalist approaches are grounded in terms of *rational agency*—focusing on our ability to hold certain beliefs for reasons and, thus, be responsible for what we believe. In what follows, I will consider a typical example of a judgement-sensitive belief, in order to *compare and contrast* these two distinct approaches. I will focus on judgement-sensitive belief because, as I argued in chapter 7, it is only this type of belief to which the rationalist approach is applicable. Non-judgement-sensitive beliefs—such as one's perceptual belief that one is seeing a red tomato—will not require one to exercise reason. I will show how Byrne's account faces the passivity objection, before turning to the rationalist interpretation, which I claim avoids the objection. Recall, from §8.2, that on Byrne's rule following procedure, one can attain knowledge of one's own belief by following the epistemic rule BEL:

'BEL If *P*, believe that you believe that *P*' (2005a, p.95, my emphasis)

Since it will be useful, while examining these two different approaches to TM, to have a typical example of a judgement-sensitive attitude in mind, let us take my ordinary belief that Canberra is the capital city of Australia. Recall from the last chapter that such a mental state is *judgement-sensitive* because anyone who holds such a belief will be obligated to have reasons, in the normative sense, for why one has it. Examples of such reasons are: my map says that Canberra is

the capital, parliament meets in Canberra, the Prime Minister says that Canberra is the capital, and so on.<sup>10</sup>

Now, the question that I am interested in answering is the following: ‘How does one know that one believes that Canberra is the capital city of Australia?’ On Byrne’s construal of TM, one could have knowledge of this belief in the following way:

Step 1: ask oneself if Canberra is the capital city of Australia.

Step 2: if you think that Canberra is the capital city of Australia, attribute the belief to yourself that it is. In other words, believe that you believe that Canberra is the Capital of Australia (a second-order belief).

This two-step process suffices to have followed TM.

While I agree with Byrne that TM can yield knowledge of one’s own belief, I will argue that his empiricist approach is beset by what I have described above as the passivity objection. To see why this is the case, let us look at step 1 in more detail. Recall that this step requires one to judge whether Canberra really is the capital city of Australia. Relying solely on empirical factors—such as self-detection and observation—how can this be done?<sup>11</sup> Here are three possibilities in which I think this could occur:

- (a) The proposition ‘Canberra is the capital of Australia’ feels true to me.
- (b) I detect, in my own mind, the belief that Canberra is the capital. I thus ground my judgement that *it is* the capital based on this belief.

---

<sup>10</sup> Whether these are good reasons is, of course, a separate issue.

<sup>11</sup> Recall that this view adheres to empiricism, as Byrne says in the following: ‘there is an appropriate causal mechanism...the state detected is independent of its detection.’ (2005a, p. 98)

(c) I commit myself to (make up my mind about) the truth of the proposition that Canberra is the capital of Australia. I consider facts relating to the location of the nation's capital and judge that the proposition is true.

Two candidates can immediately be ruled out here, on the basis that they do not fit with Byrne's account of TM. Option (b) can be ruled out because it would require that one use a form of inner sense to first detect what one believes. This strategy would not, strictly speaking, render TM false, but it would make it superfluous: if you need to know what you *believe* in order to follow TM, then there is no point in using TM in the first place to find out what you believe. You would already know what you believe. Option (c) can also be ruled out, as it is the rationalist view, meaning it that cannot be reduced to empiricist factors—e.g. self-detection and observation.

This leaves option (a) as the only remaining way to interpret Byrne's account of TM. This means that one can answer the question 'Is Canberra the capital of Australia?' in a way that conforms to both TM and empiricism. What would it mean to say, then, that a proposition feels true? One way in which this might occur would be if a phenomenological feeling was associated with the proposition 'Canberra is the capital city of Australia' when one conceives it.<sup>12</sup> The feeling would then, presumably, be absent when someone withholds assent—e.g., when someone is asked whether their parents were alive during the

---

<sup>12</sup> The locution 'seems to feel true' may seem vague, but it features in other empiricist theories of belief. Hume, for example, seems to point to a similar phenomenon when he says the following: 'If I see a billiard-ball moving toward another, on a smooth table, I can easily conceive it to stop upon contact. This conception implies no contradiction; but still it feels very differently from that conception by which I represent to myself the impulse and the communication of motion from one ball to another.' ([1748] 1999, ECU 5.2.11; SBN 48). It is a qualitative feeling that makes the difference here for Hume. Reason need *not* be involved.

time of the dinosaurs.

The main issue that I have with this approach is that it doesn't seem to take into account the fact that believing a proposition is something that is done for reasons. In other words, Byrne's account would seem to imply that the process of judging that a proposition is true might be reduced to self-detection or observation. This is the concern that I described above as the passivity objection in §8.1. Now, while it is true on certain occasions, as Boyle has pointed out, one can sometimes be unaware of the 'specific grounds for holding a given belief' (2011, p. 236), this does not cast doubt upon the fact that the question of *why you believe something* is still itself always intelligible. Even if one does not possess any good grounds for believing something, one should still understand that this is typically what is required of belief. If I am asked 'Why do you believe that Canberra is the capital of Australia?' and I respond 'I just do, I have no reason' my reply is likely to raise some confusion. If I cannot offer grounds for why I believe a proposition, I will have no justification for thinking that my belief is true.<sup>13</sup>

The view that TM can be interpreted in a broadly empiricist way is cast into further doubt by considering acts of immediate self-correction. Consider the following example that features in a recent discussion by Richard Moran (2012)—originally proposed by Sydney Shoemaker (2003).<sup>14</sup> Suppose I am asked who the president of the Confederacy was during the American Civil War. Suppose I answer: 'Robert E. Lee', before correcting myself by saying 'Jefferson Davis'—after realising that this was the correct answer. What is noteworthy about this example is that I am not taking my original answer—Robert E. Lee—

---

<sup>13</sup> The reason 'I remember once believing this' could even count as a reason.

<sup>14</sup> This example was also discussed in chapter 7.

as a brute passive perceptual phenomenon to simply be detected. The dissatisfaction with my original answer arises, when I *stop treating the question as a brute stimulus*, and rationally engage with the question. I may recall that it was President Davis who appointed Lee general in chief of the armies of the Confederate states, and so it cannot have been Lee who was the president. Such critical engagement with the content of the question, and the subsequent reasoning that occurs after my initial answer is given, means that I am not treating the question as something to be simply detected or observed. It is true that the empiricist could respond by saying that the original answer stopped *feeling* true, and was replaced by another feeling, but this reply seems unconvincing given the story just told.

Notice that we can contrast this sort of reasoning procedure with a *perceptual belief*—which would not be subject to the passivity objection. Suppose I am asked whether I believe that the coffee mug next to me is red. Recall that I can know this by following the rule BEL: if *P*, believe that you believe that *P*. This involves the following two-step process:

Step 1: do you see a red coffee mug in front of you?

Step 2: if you judge that you do, believe that you believe the coffee mug is red.

Unlike my belief that Canberra is the capital city of Australia, or my belief that the president of the Confederacy during the American Civil War was Jefferson Davis, I do not need to be aware of the reasons why I believe this. My belief is justified by the passive experience of the mug. If I am asked why I

believe this, the only grounds I will require to provide will be the perceptual experience itself. This is something that just occurs to me.

Whilst this may be an appropriate way to account for the knowledge one can have of one's perceptual beliefs, it will not be an appropriate way of accounting for the knowledge one can have of one's judgement-sensitive beliefs. As I argued in chapter 5, in order to determine whether I believe that *P* (where my belief that *P* is a judgement-sensitive belief) I must commit myself the truth of *P*. To do this I must consider facts relating to the truth of *P*—meaning that I must reason whether in fact *P* is true. This, I argued, cannot be achieved in broadly empiricist terms. My beliefs about presidents and capital cities are commitments that I have towards various facts about the world. When I *judge* that Canberra is the capital city of Australia, or that Jefferson Davis was president of the Confederacy, I need to attend to the *reasons* that I have for accepting these propositions about the world, in order to determine whether or not they are true. By foregoing any appeal to rational commitment, I cannot adequately perform this task. I would have to be treat my beliefs, as Moran puts it, like a 'stimulus insulated from the exercise of reason' (2012, p. 231). I will argue in what follows that the same issue arises for those who wish to extend the application of TM to other judgement-sensitive attitudes, without appealing to the exercise of reason.

#### **8.4 Transparency and Desire**

Like Byrne, I agree that TM can be extended to other propositional attitudes such as desire. Unlike Byrne, I do not think the empiricist approach to TM can fully account for the self-knowledge that we can have of what we desire. In this

section, I will offer some criticism of Byrne's empiricist approach to TM for desire. Firstly, I will identify some issues with the 'outward directed' question that Byrne thinks corresponds to the question 'Do I desire to  $\Phi$ ?' I then argue that even if Byrne can find a way to circumvent these issues, his account, as an empiricist one, still faces the passivity objection, when judgement-sensitive desires are considered. I argue that in order to overcome such difficulties, the rationalist version of TM should be accepted.

#### **8.4.1 Byrne's Empiricist Approach to the Transparency Method—Desire**

Let us begin by first looking at Byrne's epistemic rule DES—the rule he thinks can give one knowledge of one's own desires. It can be formulated as follows:

'DES If  $\Phi$ ing is a desirable option, believe that you want to  $\Phi$ ' (2012, p. 177).<sup>15</sup>

Byrne thinks that knowledge of one's own desires is typically obtained by trying to follow DES. Byrne supplements DES with the following: 'DESIRE Knowledge of one's desires is typically obtained by trying to follow DES' (2012, p.177).<sup>16</sup>

One key difference that Byrne claims distinguishes DES from BEL is that it is not self-verifying.<sup>17</sup> This is because, as Byrne correctly recognises, cases of accidie are possible.<sup>18</sup> Suppose that one of Tim's favourite things to do on a

---

<sup>15</sup> Byrne states that he treats the terms 'want' and 'desire' as equivalent (2012, p. 174 ft. 12).

<sup>16</sup> This does not mean that this is the only way in which one can know what one desires (see Byrne 2012, p.177, ft. 18).

<sup>17</sup> Byrne think that the BEL rule is self-verifying in the following sense: anyone who even attempts to follow it, will have a true belief about what they believe.

<sup>18</sup> Cases of accidie are those where an agent recognises that there is an action that she can and ought to bring about, but does not act to bring that action about. For example, one may continue to lie in bed rather than get

Sunday morning is to take his border collie Iggy for a walk through his neighbourhood. Imagine that Tim typically looks forward to this activity all week long. Now, does it follow that if, on Sunday morning, Tim concludes that taking Iggy for a walk is a desirable option that he should believe that he wants to take Iggy for a walk—as is suggested by DES? Not necessarily. Tim might have had one too many glasses of wine the night before, and the idea of getting out of bed and going for a walk makes his head spin.

Despite the possibility of such cases, Byrne still thinks DES is ‘*practically* self-verifying: for the most part’ (2012, p.178). He grants that this does not necessarily mean that DES is a good rule, but he thinks that the ‘burden of proof should be on those who think it is not’ (2012, p. 178). To my mind, however, DES faces several problems that go much deeper than the rule’s tendency to yield true beliefs about what one desires.<sup>19</sup> In fact, I think that the above example, involving accidie, is relatively insignificant insofar as DES’s plausibility goes. I think that if the antecedent in DES is modified such that it says ‘if  $\Phi$ ing is a desirable option *all things being considered*’, then the case involving Tim may not really be a case where one has followed DES and has formed a false belief about what one desires—as Tim may not think that going for a walk with his border collie, all things being considered, is a desirable option.<sup>20</sup>

There are more serious issues with DES, however. First, I think that the connection between ‘finding something a desirable option’ and ‘wanting’ is problematic. Byrne explicitly states that he understands the terms ‘want’ and

---

up and go to appointment one has previously arranged. For a more detailed discussion of the concept of accidie, see Sergio Tenenbaum (2010).

<sup>19</sup> I am not saying that it is a problem, in itself, that the rule has exceptions—after all, we do not want an account of self-knowledge for desire that entails infallibility. The problem is, rather, to explain what is going on in cases where someone finds something a desirable option, and yet does not desire it.

<sup>20</sup> This raises the question of what a mistaken attribution of desire would look like on Byrne’s account.

‘desire’ as equivalent (2012, p. 174 ft. 12). This allows him to connect (i) the believing that one wants to  $\Phi$ , with (ii) the knowledge that one desires to  $\Phi$ . But he also thinks that ‘desirable option’ cannot just mean the same as desire, as by Byrne’s own lights one can find an option desirable and yet not desire it.

The difficult question for Byrne is ‘What does it mean to say that one has judged that  $\Phi$ ing is a desirable option?’ Byrne acknowledges that ‘desired option’ cannot simply mean that one desires an option, as that would make DES circular (2012, p. 178). It would also make DES superfluous. If one had to know what one desired, in order to know that  $\Phi$ ing is a desirable option, then DES may still yield true beliefs about what desires, but it wouldn’t make it a very useful epistemic rule.

What then is Byrne, as an empiricist, left to say here? Surely Byrne must say that in order to judge that the antecedent in DES is true—namely, that  $\Phi$ ing is a desirable option—one must detect or ‘feel’ that  $\Phi$ ing is a desirable option. This might work for non-judgement sensitive desires, such a desire for a cold drink on a hot day, but it is not clear how Byrne could account for judgement-sensitive desires such as John’s desire to study architecture or Susan’s desire to go shopping because there isn’t any food in the fridge. How could one judge that such options are desirable without the exercise of reason? What could be detected or observed that is not a psychological mental state?<sup>21</sup>

To attempt to answer such questions, let us consider a case that Byrne discusses. Byrne imagines that there is a discussion about the mind-body problem that has started in the faculty lounge, and he is deciding whether to step

---

<sup>21</sup> Recall that as a TM theorist, Byrne thinks that ‘[t]here *is* a mechanism for detecting one’s mental states but...in an important respect it does not “resemble perception”’ (2005a, p. 80). This cannot be a matter of detecting a psychological mental state by a form of inner sense, for example, as that is what is entailed by the inner sense view.

in and sort out the conceptual confusion he believes has arisen.<sup>22</sup> Now, although the participants in the discussion might appreciate his insightful remarks, they may also find his interruption quite rude. After some deliberation, Byrne decides to intervene. Now, the question that I am interested in answering here is: how would DES give one knowledge of one's desire to intervene in the discussion in a way that is (i) broadly empiricist and (ii) not a matter of detecting a previously existing mental state or rational commitment? I am not questioning whether or not Byrne would have knowledge of his desire to join the discussion—I agree in this situation he probably does—I'm questioning how following DES can provide knowledge of this fact. According to DES one should proceed as follows:

Step 1: determine whether joining the mind-body discussion is a desirable option.

Step 2: if you judge that it is, then believe that you want to join the mind-body discussion.

According to Byrne, this two-step process will, typically, lead one to knowledge of one's own desire, all things considered. To assess the claim, I will look a bit closer at step 1. What I will focus on here is how one might go about *determining* whether joining the mind-body discussion is a desirable option (a desirable option that is not necessarily the most desirable)—in a way that is broadly empiricist, and not a matter of detecting an existing desire or a rational commitment.

---

<sup>22</sup> See Byrne (2012, p. 179).

One way in which I can envisage step 1 being followed—given such restrictions—is by making the following inference: I notice that I am deliberating about joining the mind body discussion and, thus, I infer that I must find this a desirable option. I may reason as follows: ‘why would I be deliberating about joining in if I didn’t think it was a desirable option?’ This would then lead to me to step 2, where I would attribute the belief to myself that I want to join. While this may be a good inference to make, and one that may lead me to a true belief about what I desire, it seems to me to be an untenable way to interpret TM. One reason for thinking this is because such an approach to TM requires that one take a *passive* stance towards what one desires. One need not be aware of *why* one desires a certain option, on this view. Subsequently, such an inference does not seem to be able to account for the fact that a judgement-sensitive desire is something that is held for reasons. Moreover, it puts the rule follower in only a slightly better position than someone trying to learn of another person’s desire from the third-person point of view.

To make this point clearer, let us consider a slightly modified version of DES—a behaviourist alternative which I will call DES\*. We can formulate it as follows:

DES\*: If you *see* yourself deliberating about whether  $\Phi$ ing is a desirable option, believe you want to  $\Phi$ .

To see how this rule works, let us consider a similar example to the one just described. Suppose that Byrne’s colleague, Alex Blaine, is a researcher in the biology department. Suppose Blaine is walking past his faculty lounge and notices that a discussion about Darwinism has started up. Facing a similar

predicament to Byrne, Blaine contemplates entering. Aware of Byrne's rule following procedure for acquiring self-knowledge of his own desire—and being a behaviourist—Blaine decides to follow the modified BEL\* rule. He follows the rule in the following way:

Step 1: determine if I am behaving like I think joining the Darwinism discussion is a desirable option.

Step 2: if I am, believe that I want to join the Darwinism discussion.

As was the case with DES, following DES\* will sometimes generate true beliefs about one's desires, even if it will not guarantee them. Sometimes (perhaps even often) one's behaviour will correlate with what one desires. But it does not follow from this that DES\* is a good general rule. The fact that I am behaving like I desire to  $\Phi$  may indicate that I do indeed desire to  $\Phi$ , but the rule leaves too much out. For one, the reasons why I desire to  $\Phi$  will not be accounted for by such a rule. Moreover, the rule cannot account for the fact that such a desire is sensitive to various beliefs I have about the world, which could lead me to extinguish the desire, when I judge there to be sufficient reason for doing so.

I have brought up the DES\* rule because I think that it helps to highlight a problem that is also present in DES. The problem is that DES, like DES\*, cannot account for the fact that judgement-sensitive desires are mental states that are held *for reasons*. So even if following DES will sometimes generate true beliefs, this does not mean that it is a good rule. In order to do justice to the concept of judgement-sensitive desires, we need an account of TM that can accommodate such considerations.

#### 8.4.2 The Rationalist Approach to the Transparency Method—Desire

So far, I have raised some issues with Byrne’s attempt to extend TM to desire. One of the problems that I discussed was that his view cannot account for judgement-sensitive desires, being that it is an empiricist approach to TM. In what follows, I argue that this problem can be avoided if we accept the rationalist approach to TM. First, let me say a bit more about the distinction between judgement sensitive desire and non-judgement sensitive desire, in order to show that the distinction is a real, and important, one. As a starting point, let us consider what Tim Schroeder says about the concept in his Stanford Encyclopedia entry on desire:

[t]o desire is to be in a particular state of mind. It is a state of mind familiar to everyone who has ever wanted to drink water or desired to know what has happened to an old friend (2015).

Notice from Schroeder’s passage that there appears to be two distinct types of desires here. Call the former a *brute desire*; and call the latter a *judgement-sensitive desire*. A desire can be thought of as *brute* when it is one that simply assails us—e.g., one’s desire to have a cigarette, or one’s desire to quench one’s thirst. Such desires are mental states that can, as Moran puts it, be thought of as ‘stimulus insulated from the exercise of reason’ (2012, p. 231). We can call a desire *judgement-sensitive* when that desire is subject to the exercise of reason, and can potentially be exhausted when rationality requires it. In the example above, one’s desire to know what happened to an old friend should extinguish when one finds out what happened to that friend.<sup>23</sup> It would be irrational to continue to desire to know what happened to an old friend once one has found

---

<sup>23</sup> Assuming that the desire has been satisfied.

out. This differs from someone who continues to want a drink of water even after one has had a drink—reasoning is not likely to make the desire evaporate. What both types of desire have in common, as Neil Sinhababu points out, is that they are the sorts of mental states that ‘cause pleasure when we sense or imagine them satisfied, and displeasure when we imagine them being dissatisfied’ (2017b, p. 95).

Judgement-sensitive desires *differ* from brute desires because they involve the exercise of reason. Consider the following from Akeel Bilgrami, for instance, who says, ‘to desire something, to believe something, is to think that one *ought* to do or think various things’ (2006 p. 213).<sup>24</sup> So, for example, if one desires to help the poor, one is committed to doing and thinking various things, such as giving money to charity, or volunteering one’s time. If Bilgrami is right here that certain desires do involve a normative component—something I will say more about below—then it is hard to see how self-knowledge of them could be gained in isolation from reason, as the empiricist maintains.

With this distinction in place, we can now attend to the question: ‘What outward directed question one can attend to, in order to know what one desires?’ Speaking very broadly about desire, I propose that the following outward-direct question can be attended to, in order to know what one desires.

- (1) Would  $\Phi$ ing (whether that involves bringing about  $\Phi$ , or  $\Phi$ ’s eventuation) bring me pleasure or satisfaction?

---

<sup>24</sup> Bilgrami does not think that all desires are like this.

To see how answering (1) can lead one to knowledge of what one desires let us consider an example. Suppose I want to know whether I desire to study architecture (a judgement-sensitive attitude). According to (1), I would first ask myself the question ‘Would studying architecture bring me pleasure or satisfaction?’ If I judge that studying architecture would bring me pleasure or satisfaction, then, in my view, I am entitled to believe that I do desire to study architecture. I think the same principle can be applied for brute desires. For instance, suppose someone wants to know whether they desire a cold glass of water on a hot day. On my view, they can do so by attending to the following question: ‘Would a glass of cold water bring me pleasure or satisfaction?’

This is all good and well, but what support can be given for thinking that TM can be applied to desire in the way I that have described? As I said with respect to the propositional attitude belief, I think that Moore’s paradox can lend support. Recall that in chapter 5, I said one could support the application of TM to belief—the view that one’s judgement can be a guide to what one believes—because of the Moore’s paradox sentences that are produced when one avows that one judges that *P*, and believes not-*P*. For instance, it would be absurd for someone to say ‘I judge that Australia is a continent, but I don’t believe it is’.<sup>25</sup> I argued that this shows that when one judges something to be true, they typically believe it is. The same principle is applicable to my account of desire. If I were to avow ‘studying architecture would bring me pleasure or satisfaction, but I don’t desire it’; or ‘drinking a glass of cold water would bring me pleasure or satisfaction, but I don’t desire it’, I would appear to be conceptually confused about what it means to desire something. This mirrors the situation that someone

---

<sup>25</sup> Recall that I still granted that it was possible that one could judge that *P*, and believe not-*P*. I only argued that typically when one judges that *P*, one believes that *P*.

is in when they judge that *P* is true, but don't believe it is. It is such a result, that I think supports (1).

As far as empiricism goes, I think that one could use my account of desire to know what *brute* desire one has. That is, I think one could know what brute desire one has by using the 'kinds of epistemic factors countenanced by empiricism'—as Gertler (2016, p.18) says.<sup>26</sup> However, I think that if one attempts use my account of desire to have knowledge of what judgement-sensitive desire they have, with only empiricist factors, then one will face the passivity objection. What I am claiming here is that in order to answer the question 'Would  $\Phi$ ing bring me pleasure or satisfaction?', when judgement-sensitive desires are considered, one must go beyond purely empiricist factors, and rely on rational agency and commitment.

To expound this idea further, let us consider an example. Let us suppose that Michael has an overriding desire to pursue a career in architecture. According to what I said above, Michael can have knowledge of this desire by attending to the question 'Will pursuing a career in architecture bring me pleasure or satisfaction?' This requires Michael to attend to *outward phenomena*, as opposed to *detecting* a pre-existing psychological mental state, such as a desire to pursue a career in architecture. This means that Michael's attention is directed at the intentional content of his desire. And since this desire is

---

<sup>26</sup> As Gertler points out, empiricists such as Byrne do not need to think of the process by which one achieves self-knowledge as '*observation* in any ordinary sense' (2016, p.2). She adds that '[s]ome empiricists take self-knowledge to consist in self-attributions of attitudes that are empirically warranted by virtue of a reliable link with the attitudes themselves, where the process by which one arrives at self-knowledge need not involve observation in any ordinary sense' (2016, p.2).

judgement-sensitive, it will not be enough to take an observational stance towards such content. Michael will need to exercise his reason.<sup>27</sup>

What does it mean for Michael to exercise his reason with respect to the idea of pursuing a career in architecture? Previously, I discussed the idea of taking an active/deliberative stance with respect to believing something. I discussed Moran's idea of making up one's mind about the propositional content of a belief, in order to have knowledge of one's belief. On this view, if I wanted to know whether I believe that Canberra is the capital city of Australia, I must make up my mind about whether I think that it is. How do I make my mind up about a desire, however?

The idea of commitment is applicable to Michael's desire to become an architect in two key senses. The first relates to Michael's acceptance of various background beliefs that render his desire intelligible. If Michael desires to study architecture, he should accept that he is not already an architect; that architecture is a subject that one can study; that he is eligible to study architecture, and so on. He will also need to be cognizant of his various background beliefs, intentions and other commitments he might have. If he holds the belief 'there is nothing that can be taught at a university that you can't learn by yourself', then his desire to attend university will seem unintelligible.

The second sense of commitment relates to the idea of bringing about, or thinking about (in the sense of accepting certain claims), the object of the desire. The addition of this requirement captures the fact that a person who desires something will, as Schroeder points out, tend 'to act in certain ways, feel in certain ways, and think in certain ways' (2015). If one had an overall desire for

---

<sup>27</sup> The object of Michael's attentive act can be thought of as outward phenomena because Michael is not detecting an inner psychological mental state—that is, he is not detecting a desire he has. Michael's attention is focused upon the prospects of studying architecture.

something, one should, *ceteris paribus*, make legitimate attempts to bring about the state described by the desire.<sup>28</sup> If Michael says that he desires to study architecture, but appears completely indifferent to bringing this state of affairs about, then one could question whether he really does desire this. Michael's commitment to his desire to pursue a career in architecture requires him, for example, to search for universities that offer architecture, prepare his academic record, make various applications, and so on. This requires that he attends to the intentional content of his desire in a way that it not just a matter of detecting a brute stimulus. As Bilgrami describes in the following, there is a normative requirement involved with desire. He says:

to desire something, to believe something, is to think that one *ought* to do or think various things, those things that are entailed by those beliefs and desires by the light of certain normative principles of inference (those codifying deductive rationality, decision-theoretic rationality, perhaps inductive rationality, and also perhaps to some broader form of material inference having to do with the meaning of words as well). It is not to be disposed to do or think those things; it is to think one ought to do and think them. A good word that is often used to describe such internal oughts that are not defined in terms of the corresponding dispositions is that they are 'commitments', commitments to think various things and to do various things (2006, p. 213).

It is important to stress that while Bilgrami's passage is useful in highlighting some of the ways in which desire involves the exercise of reason, this is not applicable to all desires. This is something that appears to be applicable only to judgement-sensitive desires. The main idea from the passage that I wish to draw attention to is that someone who desires something does not simply act *passively* towards that desire. To desire something is to take an *active* stance towards various states of affairs, such as those described above by Bilgrami. It is not clear

---

<sup>28</sup> The desire to do *literally* nothing will still require one to do certain things, such as resisting urges do other things.

how we can integrate this important feature of desire into the empiricist account.<sup>29</sup> What I am claiming here is that it is not easy to see how one could attend to the question ‘Would  $\Phi$ ing bring me pleasure or satisfaction?’, in a way that is sensitive to what we have said above, while at the same time relying on only empiricist factors. Unless we are prepared to reject the very notion of judgement-sensitive desires, it is my view that a rationalist interpretation of TM, like the one I have proposed here, should be accepted.

Our foregoing discussion of commitment and rationality is also relevant to an objection against rationalism made by Quassim Cassam (2014). Recall that in chapter 6.1, I examined Cassam’s objection that rationalism overstates what it is for human beings to have self-knowledge of their mental states. In his view, the rationalist interpretation of TM would require that one attend to the outward-directed question ‘Ought I to desire to  $\Phi$ ?’ in order to know whether one did desire to  $\Phi$ . Such a process, Cassam argues, would put the achievement of self-knowledge of what one desires out of reach of most people. According to the account of desire that I have proposed, the outward-directed question that one should attend to, in order to know what they desire, is not ‘Ought I to desire to  $\Phi$ , but rather it is ‘Would  $\Phi$ ing bring me pleasure or satisfaction? Rationality is required, I have argued, in order to answer the question. It is not a part of the question itself, as Cassam thinks.<sup>30</sup>

---

<sup>29</sup> Although this passage, and our general discussion, has focused on future-directed actions that a subject will be committed to, it is important to clarify that such behaviour will only be a part of what it means to say that one’s desires involve commitment. The prisoner of war who desires that he is released from his cell so that he can return to his homeland may not be in a position to fulfil his desire. Recognizing the hopeless of his situation, he puts his fate in the hands of the guards. It does not follow from this that he must remain passive towards his desire, however. His commitment takes on a different form—one which may consist solely in accepting various counterfactuals such as ‘if I were freed, I would return to my homeland’, and ‘if a case could be made for my release, I would make it’.

<sup>30</sup> This would only apply to judgement-sensitive attitudes.

In summary, the rationalist account of TM is able to capture the fact that judgement-sensitive attitudes are mental states that subjects hold *for reasons*. This is in contrast to Byrne's empiricist account, which struggled to account for this fact. Unless we are prepared to reject the very notion of judgement-sensitive attitudes, we should embrace an account of TM for desire such as the one I have put forward here.

## 8.5 Transparency and Intention

I will now consider how TM can be extended to another judgement-sensitive attitude: intention.<sup>31</sup> As with the previous attitudes we have examined, I first offer a critique of Byrne's empiricist interpretation of TM, before showing why a rationalist interpretation should be preferred.

### 8.5.1 Byrne's Empiricist Approach to the Transparency Method—Intention

Byrne (2011b, p. 216) proposes the following epistemic rule for intention:

INT: if you will  $\Phi$ , believe you intend to  $\Phi$ .<sup>32</sup>

According to Byrne, one can have knowledge of one's own intentions by following INT. To gain a better understanding of how one would follow INT, let us consider an example. Suppose it is true that Bill intends to go to the casino on

---

<sup>31</sup> Whether all intentions are judgement-sensitive is a question I remain agnostic over. The intentions I discuss in what follow certainly are.

<sup>32</sup> Byrne calls this the '*bouletic* schema' (2011b). In his formulation of it, he has as his antecedent: 'I will  $\Phi$ '; and as his consequent, he has: 'I intend to  $\Phi$ ' (2011b, p. 216). For consistency with the other rules discussed here, I have replaced 'I' with 'you'. Byrne considers two other rules, before settling on INT ('INT' is my name for the rule). These include (1) 'if I think  $\Phi$ -ing is the best option, then believe I intend to  $\Phi$ ' (2011b, p.214); and (2) 'if I will intentionally  $\Phi$ , then believe I intend to  $\Phi$ ' (2011b, p. 217).

Saturday night. According to Byrne, Bill can acquire knowledge of his intention by following INT. To apply this rule Bill should follow these two steps:

Step 1: determine if you *will* go the casino on Saturday night.

Step 2: if you judge that you will, believe that you *intend* to go the casino on Saturday night.

Although at first pass this rule seems plausible—since one typically intends to do what one will do—the rule faces several difficulties. One issue arises from the fact that there are times when one does not intend to do what one believes one will do. Consider the following example. Suppose Jane is a recovering gambling addict who is currently trying to overcome her addiction. Jane is also realistic, however. She believes that addictions are not easily dissolved, and she will most likely go to the casino on Saturday night because she is weak-willed. She, thus, believes that she *will* go to the casino on Saturday night, but she does not *intend* to. If Jane were to follow INT she would have a false belief about one of her intentions.

Consider another example—one that Byrne himself discusses—that is raised by Elizabeth Anscombe. Anscombe imagines a case where a poorly prepared student is about to take an exam and thinks to himself ‘I am going to fail in this exam’ ([1957] 2000, p. 2). In such a case, we can think of the student’s thought more like a prediction of what he will do, rather than an intention to fail the exam. We can imagine that in the moments leading up to the exam the student is doing all he can to remember the content that he is being tested on. In these final moments before the exam, he has the intentions to pass,

but being realistic, believes he will fail. Following INT does not give the right answer here either.

Despite the existence of such cases, Byrne still maintains that INT is a good rule, because it is '*practically* strongly self-verifying' (2011b, p. 219). Byrne adds: 'if one reasons in accord with the schema (and is mindful of defeating conditions, for instance the one just noted), then one will arrive at a true belief about one's intention' (2011b, p. 219). While I agree with Byrne that INT will *sometimes* yield true beliefs about one's own intentions, this does not make it a good rule in my view. The main reason is that INT distances the rule follower from the reasons why she intends to do what it is she does. The rule requires one to take a predictive stance towards one's course of behaviour—meaning that one must focus on whether  $\Phi$  *will* happen to oneself, not on the *reasons why* one intends to  $\Phi$ .

To my mind, this doesn't do justice to the concept of an intention. INT does not seem to be able to account for the fact that someone who intends to do  $\Phi$ , will not only believe that they will do  $\Phi$ , but they will also be prepared to act in various ways to bring about  $\Phi$ . As Amir Saemi points out, it is a conceptual truth that an 'agent needs to have some commitment to execute his/her intention' (2015, p. 202). My intention to fly to Hawaii for the summer is not just a belief about what will happen to me in the summer; but rather, it is a series of commitments of mine that relate to my goal of bringing about this state of affairs—namely, flying to Hawaii. Thus, it is somewhat inconsequential that INT will sometimes generate true beliefs about one's intentions. That we often intend to do what we will do is a point that I agree with Byrne on. But this is not enough to provide us with a plausible way of extending TM to intention.

To further illustrate my point here, let us consider a behaviourist version of Byrne's rule INT, that I will call INT-BEHAVIOUR. It can be formulated as follows:

INT-BEHAVIOUR: if it looks like you are  $\Phi$ -ing, or about to  $\Phi$ , believe you intend to  $\Phi$ .

Like Byrne's rule INT, following INT-BEHAVIOUR will sometimes yield true beliefs about one's own intentions. Suppose I am at a party, and I observe that I am searching for my car keys. If I follow INT-BEHAVIOUR, I will have a true belief about the fact that I intend to leave the party. But surely this will not make INT-BEHAVIOUR a good rule. While this rule may sometimes generate true beliefs about one's intentions, it will also be severely limited. Complex intentions, such as my intention in July to attend the New Year's Eve parade may not have any obvious behavioural manifestations to observe. And so, by only focusing on behaviour, it is not clear how I can have knowledge of this intention until I start behaving in such a way that indicates that I intend to pursue this course of action.

One may reply to the objection I have raised by saying that while INT-BEHAVIOUR cannot account for every intention one can have, this is not a problem, as it is reliable enough in practice. This would be an analogous response to the one that Byrne offered to a similar objection that was raised with respect to the rule INT. Both replies, in my view, are unsuccessful because they overlook the fact that intentions are things done for *reasons*. By only attending to what will happen to one, or one's own behaviour, one must take a passive stance

towards one's own intentions. This does not do justice to the concept of an intention. Although following INT will sometimes generate true beliefs about what one intends to do, the fact that the rule cannot do justice to the concept of an intention requires that we attempt to find an alternative approach. INT cannot account for the fact that intentions are judgement-sensitive—meaning that they are actions rational agents do for reasons. This provides us with another example of the passivity objection.

Although I have been critical of Byrne's attempt to extend TM to intention, I agree with him that TM can be extended to intention. In what follows, I will show how a rationalist interpretation of TM can avoid the problems I have raised.

### **8.5.2 The Rationalist Approach to the Transparency Method—Intention**

What is the right 'outward phenomena' that one should attend to, then, in order to know what one intends to do? Like with our analysis of belief, and desire, we must first have a good grasp of what the concept of an intention is, before we can answer the question. We can do so by considering the following from Michael Bratman, who points out that 'an intention to act is a complex form of commitment to action, a commitment revealed in reasoning as well as in action' ([1987]1999, p. 107). This complements Saemi's claim that I mentioned above, which states that an 'agent needs to have some commitment to execute his/her intention' (2015 p. 202). An intention, thus, is an attitude one takes towards a future directed action, rather than just an event in the future that will happen to one.

An alternative explanation of how one can know what one intends by following TM, is given by Sarah Paul, who claims ‘we can come to know what we intend by making a decision about what to do and self-ascribing the content of that decision as our intended action’ (2012, p. 327). This means that one can come to know what one intends to do by asking oneself the question ‘Have I decided to  $\Phi$ ?’ The addition of this deliberative requirement seems to capture something important about what it means to have an intention—something that is absent from Byrne’s account. If I suddenly *decide* to fly to Hawaii for the summer, then it seems to make sense to self-ascribe the *intention* fly to Hawaii. Support for this claim comes from considering how startling the following sentence would sound if uttered by a speaker: ‘I have decided to fly to Hawaii for the summer, but I do not intend to’. There seems to be something ‘absurd’ about this sentence—something analogous to the absurdity found in Moore’s paradox sentences that we looked at when we originally discussed the paradox in chapter 5.

While I think that Paul’s approach gets closer to the conception of an intention than Byrne’s approach, I still do not think it offers us a complete story. Paul claims that ‘deciding to  $\Phi$  is normally sufficient to count as intending to  $\Phi$ ’ (2012, p. 343). While I think that deciding to  $\Phi$  may be necessary, I think that knowledge of what one intends involves more than just a decision to  $\Phi$ . Consider, for example, my intention to enter the annual City to Surf fun run. If I really do have this intention, then I will not only have decided to enter, but I will also be committed to bringing about this state of affairs. I will make sure my schedule is clear for that day, I will be aware of the time that the run commences, I will make sure I have a pair of running shoes, and so on. The commitment I

make to bringing such an event about seems to go beyond just a decision, as is suggested in Paul's account. The *decision* to enter the enter the fun run is only the first step in the process. Perhaps a necessary step, but not a sufficient one.<sup>33</sup>

In order to get closer to the concept of an intention, I propose the following way of characterising TM, for the purposes of acquiring knowledge of one's intentions. In my view, one can know what one intends to do, by attending to the question:

(2) Am I committed to bringing about  $\Phi$ ?

By attending to (2), I mean that one would commit oneself to bringing about  $\Phi$ . Importantly for the rationalist versus empiricist debate, such an activity would not be something that could be easily reducible to self-detection or observation. By a commitment to bringing about  $\Phi$ , one would not only make up one's mind to do  $\Phi$ , but one would be committed to the truth of various other propositions which relate to this future action.

For instance, a rational actor should only intend to do things she believes are possible. Suppose I learn that only under 18-year-olds are eligible to enter City to Surf fun run. Given my background belief that I am not under 18, then, the intention to enter a fun run that is eligible to only under 18-year-olds, should not be something that I am committed to. When I consider the reasons for this

---

<sup>33</sup> Paul may object to my criticism of her view in the following way. She may claim that someone who *decides* to enter the City to Surf fun run, and then never acts or thinks about the fun run ever again, never decided to enter the fun run in the first place. Such a person was simply insincere in their avowal. To decide something, simply means to commit oneself to bringing a certain state of affairs about. If this is the case, then Paul's view and my own would not be in conflict. I do not think, however, that this accurately characterises her view. Paul claims that to have knowledge of one's intention, '[o]ne must merely know *at the moment* of decision what one has decided' (2012, p. 339, my emphasis); and adds 'an intention does not depend for its existence on there being any follow-up' (2012, p. 342).

course of action, I find that I have a good reason not to pursue it. A rational actor, after all, ought not to be committed to an action they believe to be impossible to perform. So, it seems like my commitment involves more than just a solitary decision. It involves not only a commitment to a specific course of action, but also a commitment to the related beliefs that I take to be true. If I hold the belief that a certain state of affairs cannot be brought about, then I should not intend to bring about the state of affairs.

Support for the view that one can know what one intends to do by attending to the outward-directed question ‘Am I committed to bringing about  $\Phi$ ?’ can be given by considering the Moore’s paradox sentences it produces. Recall that on Byrne’s approach to TM, one could follow the INT rule and end up with sentences such as the following: ‘I will fail the exam, but I don’t intend to’; and ‘I will lose my job after my latest indiscretion, but I don’t intend to’. There was nothing absurd about such sentences. I suggested that this shows that the connection between what one *will do* and what one *intends* is not as congruous as it might first seem. On the account that I have proposed, absurd sentences would be produced. Consider the following: ‘I am committed to failing the exam, but I don’t intend to’; and ‘I am committed to getting fired, but I don’t intend to’. The fact that the account I have proposed produces such sentences suggests that it gets closer to the concept of intention.

In addition to having argued that TM can be applied to intention, I have also argued that the self-knowledge of our intentions requires the exercise of reason. If I intend to run the City to Surf fun run, for example, I need to take a *deliberative* stance towards this future action. I have argued that it is hard to see how the empiricist approach to TM could account for this feature.

## 8.6 Extending the Transparency Method Beyond Belief, Desire, and Intention

So far, I have explained how the rationalist version of TM can be applied to several different judgement-sensitive attitudes: belief, desire, and intention. Moreover, I have argued that self-knowledge of such mental states requires the activity of reason. Although belief, desire, and intention are important exemplars of the types of mental states that one can acquire knowledge of by following TM, they represent only a small subset of the total set of mental states that human beings are capable of being in. To get a sense of the size of this set of mental states, that also have intentional content, consider following list that is offered by John Searle:

belief, fear, hope, desire, love, hate, aversion, liking, disliking, doubting, wondering whether, joy, elation, depression, anxiety, pride, remorse, sorrow, grief, guilt, rejoicing, irritation, puzzlement, acceptance, forgiveness, hostility, affection, expectation, anger, admiration, contempt, respect, indignation, intention, wishing, wanting, imagining, fantasy, shame, lust, disgust, animosity, terror, pleasure, abhorrence, aspiration, amusement and disappointment (1983, p. 4).

Clearly, then, there is more to be said about how TM can be extended to these other mental states—more than could be possibly be offered in a single chapter. I will conclude this chapter then, not by attempting to explain as many as I can, but by saying a few *general* things about how such an extension may apply to these other mental states. Although, as we have seen, there are key differences between each type of mental state, there are also some commonalities that I think can be pointed out.

First, let me say something about how rationality could be applicable to some of the mental states that appear this list. Let us take the attitude wishing, for example. Suppose, I *wish* that I had studied physics when I was younger. How can I be said to be required to commit myself to anything here, or use the activity of reason, when there is nothing I can do to bring about a change to the past?<sup>34</sup> Let me first point out that one need not believe something to be possible, in order to *wish* it. For example, I may wish that the law of non-contradiction was false, even though I believe it could not be otherwise. This may be true, but it doesn't answer the question: 'What would one need to commit oneself to, or deliberate about, in order to have self-knowledge of one's wish'?

The answer is that one's attention should be directed towards the intentional content of the wish—namely, that one had studied physics as a child. One should then be prepared to act in various ways, with respect to such content. If I wish that I had studied physics when I was younger, I should also committed to other mental states. For example, I may *believe* that I would more employable as an adult if had learned physics; I may *desire* to converse coherently with my physicist friends; and I may *desire* to know more than I currently do about the natural world. Moreover, I should also be committed to certain counterfactuals, such as 'If I had studied physics as a child, then I would know a great deal more about physics than I do now as an adult'. Given that I desire to know a great deal about physics, my wish to have studied physics when I was younger is, thus, intelligible.<sup>35</sup> Although each mental state in the list above will differ from each other in various ways, the notions of rational consistency and commitment are ones that are applicable to a number of such mental sates.

---

<sup>34</sup> I presume, for the purposes of this discussion, that it is impossible to change the past.

<sup>35</sup> For a more comprehensive discussion of how TM can be extended to wishes, see Wolfgang Barz (2015).

The second main idea that I have defended in this chapter, is the idea that an account of TM should be able to yield Moore's paradox like sentences, just as in the case of belief—e.g., 'it is raining, but I don't believe it is'. If a characterisation of TM cannot produce such sentences, then I think that such a characterisation of TM will have failed to capture the main concept of the attitude it is trying to account for.

### **8.7 Conclusion**

In this chapter, I have offered a comparison between the rationalist and empiricist version of TM. Through a series of examples, I have argued that the empiricist version is beset by an objection that I called the passivity objection. This is the objection that the empiricist account of TM cannot account for the fact that judgement-sensitive attitudes are mental states that are held for reasons. I have argued that in order to avoid this objection, we should accept the rationalist version of TM.

If the arguments given in this chapter are successful, then a response to the generality objection—the objection that TM cannot be extended to other types of self-knowledge apart from just belief—can be given. While there was only space to examine a few different types of mental states, I am optimistic that further work will be able to show that TM can account for the self-knowledge we can have of many other types of mental states.



## Conclusion

I have attempted to characterise the nature and limits of first-person authority, the thesis that our first-person ascriptions about what mental states we are in are more likely to be true compared to the ascriptions that others make about them. I have argued that such first-person authority can be explained by appealing to introspection, in the theory-neutral sense described in chapter 1. It is our ability to introspectively justify our beliefs about what mental states we are in, I contend, that gives rise to our first-person authority.

I began in chapter 1 by offering a novel theory-neutral account of what it takes for a process to count as introspective. I argued against the historically entrenched position that we should consider a process introspective only if it involves inner perception. According to the theory-neutral account of introspection that I proposed, introspection is: (1) the process by which one gains knowledge of one's own, and only one's own, psychological mental states; (2) a process different in kind from alternative ways of acquiring knowledge about the world; and (3) can only give one knowledge of one's current mental states.

In chapter 2 I was concerned with showing how introspection, in this theory-neutral sense, could explain first-person authority. I argued against views which characterise first-person authority in term of infallibility or incorrigibility, as well as views which reject all forms of first-person authority. Instead, I argued that first-person authority should be construed in terms of *introspective justification*. This is the view that we have first-person authority because we possess the ability to justify our beliefs about what mental state we are in, by introspecting that we are in that mental state. I have argued that it is the unique

justification that introspection provides that can account for the fact that our beliefs about our own psychology are more likely to amount to knowledge, compared to the beliefs that others form about our minds.

After arguing that introspection, in the theory-neutral sense, can explain first-person authority, I then addressed the question: ‘What mental states can be introspected?’ In chapter 3, I argued against certain sceptical accounts of self-knowledge which claim that we cannot introspect our propositional attitudes. I first argued that content externalism, the thesis that thought content is determined in part by one’s environment, was compatible with the position that we can introspect our own propositional attitudes. I then considered inferentialism—the view that there is no asymmetry between the way in which we acquire knowledge of our own minds, compared to the way in which we acquire knowledge of the minds of others. After explicating one inferentialist account—namely, Peter Carruthers’ Interpretative Sensory-Access (ISA) theory—I raised some objections to it. In chapter 4, I considered the empirical evidence for the ISA theory and argued that the theory lacks adequate supporting evidence.

After arguing that at least some types of mental states can be introspected, in chapter 5 I addressed the question ‘Can we introspect our propositional attitudes?’ In so doing I defended a view of self-knowledge, which counts as a form of introspection according to my view, called the transparency method (‘TM’). I argued that TM can account for the way in which one can have self-knowledge of one’s beliefs—an important type of propositional attitude. I also argued that in order to do justice to the concept of belief, rational agency must be invoked. I argued that such a result means that we should accept the rationalist version of TM, rather than the empiricist version. In chapter 6, I dealt

with some objections to this position, arguing that these objections are not successful. In chapters 7 and 8, I attempted to extend the application of TM to other categories of mental states, apart from belief. I showed how TM can be applied to desire and intention. I also spoke very generally about how TM might be extended even further. As with the propositional attitude belief, I also argued that rational agency, with respect to the self-knowledge we can have of such mental states, cannot be discounted.

Although my focus on TM was limited—in the sense that only a few types of mental states were discussed—I remain optimistic that the application of TM can be extended even further than has been done here. By this, I mean that I think there are several other categories of mental state—in addition to belief, desire, and intention—where TM is applicable. What I did provide here, however, was a set of desiderata for how such future applications of TM should be implemented. The first desideratum I proposed is that an application of TM should be able to yield Moore’s paradox sentences, analogous to the case of belief. Recall that with the attitude belief, we saw that one can typically know whether one believes that  $P$ , by judging that  $P$  is true. In other words, judging that something is true can give one awareness of what one believes. Support for this came from thinking how absurd it would be for someone to say ‘I judge that  $P$  is true, but I don’t believe it is’. I argued that we should think of an application of TM to a certain category of mental state successful only if it could yield such sentences. Consider the application of TM to intention, for example. Recall that I said that one can know whether one intends to  $\Phi$ , by determining whether one is committed to bringing about  $\Phi$ . I supported such a claim by pointing out how absurd it would be for someone to say: ‘I am committed to

bringing about  $\Phi$ , but I don't intend to  $\Phi$ '. I argued that since this does not arise in rival attempts to extend TM to intention, we should not accept those accounts. We do not, for instance, get such absurdity in the following: 'I will  $\Phi$ , but I don't intend to  $\Phi$ ', as Byrne (2011b) proposes; nor such absurdity in the sentence: 'I ought to  $\Phi$ , but I do not intend to  $\Phi$ ', as Cassam (2014) proposes. Since such accounts do not meet my desideratum, I do not think they are successful applications of TM. When it comes to applying TM to a mental state, whether that be belief, fear, hope, and so on, we should be able to generate such Moore paradox sentences.

The second desideratum that I proposed is that the concept of the mental state one is attempting to apply TM to should be adequately accounted for. Since some mental states, such as judgement-sensitive beliefs, are mental states that are held for reasons, I argued that we need to invoke rationality, in order to explain our self-knowledge of them. When one judges that  $P$  is true, for example, in order to know whether one believes that  $P$ , one needs to commit oneself to the truth of  $P$ . I suggested that it was difficult to account for this in purely empiricist notions—such as detection and observation. So, if we are to do justice to the concept of belief, and other attitudes associated with agency, we need to incorporate rationality into our account of self-knowledge. In arguing that rationality cannot be eliminated from the domain of self-knowledge, I hope to have shown that this need not commit us to an unrealistic conception of human nature. Contrary to what some, such as Cassam (2014), have claimed, I argued that the invocation of rationality, with respect to self-knowledge, does not require us to think of ourselves as perfect epistemic citizens. I suggested that our propensity to hold mental states for reasons is not brought into doubt simply

because we are prone to self-deception, biases, *akrasia*, or other general failures of rationality.

It is crucial to point out that even if I am right that rationality cannot be eschewed from the domain of self-knowledge, there will still be limits to such rationality. I do not wish to give the impression that the position I have defended entails the thesis that rationality needs to be involved in all occurrences of self-knowledge. This was a point I stressed in chapter 7. Furthermore, one should not get the impression that the position that I have advanced entails the thesis that TM is the *only* way in which one can achieve introspective self-knowledge. The question of whether TM is applicable to some other types of mental states, such as emotions, is a question that I have remained agnostic about in this work. Whether we can introspect our emotions, and if so *how*, are issues that must be addressed in future work. What I have claimed is that if it is true that one can introspect one's emotions, then one would have first-person authority with respect to them—just as one would with respect to any other type of self-knowledge one can introspect.

Another important issue related to first-person authority that I did not go into detail about here is the question of *how reliable* our first-person ascriptions are. In chapter 2, I put this question to one side because I said it was not the right question to ask if we wanted to account for the general phenomena of first-person authority, in the theory-neutral sense relating to introspection. This is not to say that the reliability of peoples' first-person ascriptions is irrelevant, however. Eric Schwitzgebel, for instance, thinks that if our psychological self-ascriptions could be shown to be generally unreliable, then 'a methodologically well justified scientific consensus on a theory of consciousness may be beyond

our reach' (2011, p. 115). The question of how often, and under what situations people make mistakes is, therefore, an important one, but is ultimately a separate issue to the one I was addressing here. The account of first-person authority that I proposed aimed to explain the advantage we take ourselves to have with respect to our own psychology—not settle an empirical question about how often our self-ascriptions are true.

In conclusion: the topic of self-knowledge, like some other topics in philosophy such as free will is a personal one. The question 'How do we know our own minds?' is not just one of academic interest. It is a question whose answer has the potential to change the way that we think of ourselves as agents in the world. Most of us feel, after all, that our mental states are not only things we possess, but are phenomena that we have some sort of control over. While recent discoveries in the natural sciences have shown that some of our common-sense intuitions about how the mind works are false, I have tried to show that first-person authority, at least with respect to some types of mental states, is not one of them. There is an important sense in which we have first-person authority with respect to our own mental states. The account of first-person authority I have defended is, moreover, compatible with a naturalism of the sort many contemporary philosophers would accept. This means we can explain the first-person authority we have, with respect to sensory mental states, propositional attitudes, and potentially other mental states, without positing anything ontologically demanding. While the account of first-person authority offered here requires further work, it is my hope that I have done enough here to have shown that there is an important sense in which we have first-person authority with regard to our own minds.

## References

- Albahari, M. (2006) *Analytical Buddhism: The Two-tiered Illusion of Self* (New York: Palgrave Macmillan).
- Albahari, M. (2014) 'Alief or Belief? A Contextual Approach to Belief Ascription', *Philosophical Studies* **167**: 701–720.
- Alston, W. P. (1971) 'Varieties of Privileged Access', *American Philosophical Quarterly* **8**: 223–241.
- Anscombe, G. E. M. ([1957] 2000) *Intention* (Cambridge, MA: Harvard University Press)
- Armstrong, D.M. (1968) *A Materialist Theory of the Mind* (London: Routledge and Kegan Paul).
- Audi, R. ([2001] 2015) 'Doxastic Voluntarism and the Ethics of Belief', in *Rational Belief: Structure, Grounds, and Intellectual Virtue* (Oxford: Oxford University Press).
- Ayer, A.J. (1956) *The Problem of Knowledge* (London: McMillan and Co.).
- Ayer, A.J. (1963) 'Privacy', in *The Concept of a Person and other Essays* (New York: St Martin's Press).
- Barnes, E. (1994) 'Explaining Brute Facts', *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* **1**: 61–68.
- Bar-On, D. (2004) *Speaking My Mind: Expression and Self-knowledge* (Oxford: Oxford University Press).
- Bar-On, D (2009) 'First-Person Authority: Dualism, Constitutivism, and Neo-Expressivism', *Erkenntnis* **71**: 53–71.
- Barz, W. (2015) 'Transparent Introspection of Wishes', *Philosophical Studies* **172**, 1993–2023.
- Bayne, T., and Montague, M. (2011) 'Cognitive Phenomenology: An Introduction', in *Cognitive Phenomenology*, ed. T. Bayne and M. Montague (Oxford: Oxford University Press).
- Bernecker, S. (2011) 'Representationalism, First-Person Authority, and Second-Order Knowledge', in *Self-Knowledge*, ed. A. Hatzimoysis (Oxford: Oxford University Press).
- Bilgrami, A. (2006) *Self-Knowledge and Resentment* (Cambridge, MA: Harvard University Press).
- Block, N. (1990) 'Inverted Earth', *Philosophical Perspectives* **4**: 53–80.
- Block, N. (2011) 'The Higher Order Approach to Consciousness is Defunct', *Analysis* **71**: 419–431

- Boghossian, P. (1989) 'Content and Self-Knowledge', *Philosophical Topics* **17**: 5–26.
- Boghossian, P. (1992) 'Externalism and inference', in *Information, Semantics and Epistemology*, ed. E. Villanueva (Oxford: Basil Blackwell).
- Bourget, D. and Mendelovici, A. (2017) 'Phenomenal Intentionality', *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/phenomenal-intentionality/>
- Boyle, M. (2009) 'Two Kinds of Self-Knowledge', *Philosophy and Phenomenological Research* **78**: 133–164.
- Boyle, M. (2011) 'Transparent Self-knowledge' *Aristotelian Society Supplementary*, **85**: 223–241.
- Braddon-Mitchell, D. and Jackson, F. (2007) *Philosophy of Mind and Cognition: An Introduction*, second edition (Oxford: Blackwell).
- Brandom, R. (2015) *From Empiricism to Expressivism* (Cambridge, MA: Harvard University Press).
- Bratman, M. ([1987] 1999) *Intention, Plans, and Practical Reason* (Stanford: CSLI Publications).
- Brueckner, A. (2011) 'Neo-Expressivism', in *Self-Knowledge*, ed. A. Hatzimoysis (Oxford: Oxford University Press).
- Briñol, P. and Petty, R. E (2003) 'Overt Head Movements and Persuasion: A Self-Validation Analysis', *Journal of Personality and Social Psychology* **84**: 1123–1139.
- Burge, T. (1979) 'Individualism and the Mental', *Midwest Studies in Philosophy* **4**: 73–121.
- Burge, T. (1988) 'Individualism and Self-Knowledge', *Journal of Philosophy* **85**: 649–63.
- Burge, T. (1996) 'Our Entitlement to Self-Knowledge', *Proceedings of the Aristotelian Society* **96**: 91–116.
- Byrne, A. (2005a) 'Introspection', *Philosophical Topics* **33**: 79–104.
- Byrne, A. (2005b) 'Perception and Conceptual Content', in *Contemporary Debates in Epistemology*, ed. E. Sosa and M. Steup (Oxford: Wiley Blackwell).
- Byrne, A. (2011a) 'Knowing that I am Thinking' in *Self-Knowledge*, ed. A. Hatzimoysis (Oxford: Oxford University Press).
- Byrne, A. (2011b) 'Transparency, Belief, Intention', *Proceedings of the Aristotelian Society, Supplementary Volume*, **85**: 201–221.
- Byrne, A. (2012) 'Knowing What I Want', in *Consciousness and the Self: New Essays*, ed. J. Liu and J. Perry (Cambridge: Cambridge University Press).

- Carruthers, P. (2005) *Consciousness: Essays from a Higher-Order Perspective* (Oxford: Oxford University Press).
- Carruthers, P. (2008) ‘Cartesian Epistemology: Is the theory of the Self-Transparent Mind Innate?’, *Journal of Consciousness Studies* **15**: 28–53.
- Carruthers, P. (2009) ‘How We Know Our Own Minds: The Relationship Between Mindreading and Metacognition’, *Behavioral and Brain Sciences* **32**: 121–138.
- Carruthers, P. (2010) ‘Introspection: Divided and Partly Eliminated’, *Philosophy and Phenomenological Research* **80**: 76–111.
- Carruthers, P. (2011) *The Opacity of Mind: An Integrative Theory of Self-Knowledge* (Oxford: Oxford University Press).
- Carruthers, P. and Veillet, B. (2011) ‘The Case Against Cognitive Phenomenology’, in *Cognitive Phenomenology*, ed. T. Bayne and M. Montague (Oxford: Oxford University Press).
- Caruso, G. D. (2012) *Free Will and Consciousness: A Determinist Account of the Illusion of Free Will* (Lanham, MD: Lexington Books).
- Cassam, Q. (2010) ‘Judging, Believing and Thinking’, *Philosophical Issues* **20**: 80–95.
- Cassam, Q. (2014) *Self-knowledge for Humans* (Oxford: Oxford University Press).
- Chalmers, D. (2003) ‘The Content and Epistemology of Phenomenal Belief’ in *Consciousness: New Philosophical Perspectives*, ed. Q. Smith and A. Jokic (Oxford: Oxford University Press)
- Chalmers, D. (2006) ‘The Foundations of Two-Dimensional Semantics,’ in *Two-Dimensional Semantics: Foundations and Applications*, ed. M. Garcia-Carpintero and J. Macia (New York: Oxford University Press)
- Chalmers, D. (2011) ‘Verbal Disputes’, *Philosophical Review* **120**: 515–566.
- Chisholm, R. M. (1957) *Perceiving: A Philosophical Study* (Ithica: Cornell University Press)
- Churchland, P. M. (1981) ‘Eliminative Materialism and the Propositional Attitudes’, *Journal of Philosophy* **78**: 67–90.
- Churchland, P. M. (1988) *Matter and Consciousness* (Cambridge, MA: MIT Press).
- Churchland, P.S. (1986) *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. (Cambridge, MA: MIT Press).
- Coleman, A. M. (2015) *A Dictionary of Psychology*, fourth edition (Oxford: Oxford University Press).
- Coventry, A., and Kriegel, U. (2008) ‘Locke on Consciousness’, *History of Philosophy Quarterly* **25**: 221–242.
- Crane, T. (2009) ‘Is Perception a Propositional Attitude?’ *The Philosophical Quarterly* **59**: 452–469

- Dainton, B. (2008) *The Phenomenal Self* (Oxford: Oxford University Press)
- Damasio, A. (2010) *Self Comes to Mind: Constructing the Conscious Brain* (London: Vintage Books).
- Davidson, D. ([1963] 2006) 'Actions, Reasons and Causes', in *The Essential Davidson*, ed. K. Ludwig and E. Lepore (New York: Oxford University Press).
- Davidson, D. (1982) 'Rational Animals', *Dialectica* **36**: 317–327.
- Davidson, D. (1984) 'First-Person Authority', *Dialectica* **38**: 101–111
- Davidson, D. (1987) 'Knowing One's Own Mind', *Proceedings and Addresses of the American Philosophical Association* **60**: 441–458.
- DeLuca, J. (2009) 'Confabulation in Anterior Communicating Artery Syndrome' in *Confabulation: Views from Neuroscience, Psychiatry, Psychology and Philosophy*, ed. W. Hirstein (New York: Oxford University Press).
- Dennett, D. C. (1991) *Consciousness Explained* (Boston: Little, Brown and Co.).
- Dennett, D. C. (2002) 'How Could I Be Wrong? How Wrong Could I Be?', *Journal of Consciousness Studies* **9**: 13–16.
- Dennett, D. C. (2017) *From Bacteria to Bach and Back: The Evolution of Minds* (New York: W.W. Norton and Company).
- Descartes, R., ([1641] 1984) 'Meditations on First Philosophy', in *The Philosophical Writings of Descartes*, volume II, ed. and trans. J. Cottingham, J. Stoothoff. and J. Murdoch (Cambridge: Cambridge University Press).
- de Sousa, R. (2013) 'Emotion', *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/emotion/>
- de Vignemont, F. (2015) 'Bodily Awareness', *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/bodily-awareness/>
- Deweese-Boyd, I. (2016) 'Self-Deception', *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/self-deception/>
- Dominus, S. (2011, May 25) 'Could Conjoined Twins Share a Mind', *The New York Times*. Retrieved from: <http://www.nytimes.com/2011/05/29/magazine/could-conjoined-twins-share-a-mind.html>
- Dretske, F. (2003) 'How Do You Know You Are Not a Zombie?' in *Privileged Access: Philosophical Accounts of Self-Knowledge*, ed. B. Gertler (Burlington, VT: Ashgate Publishing Company).
- Egan, L. C., Santos, L. R., and Bloom, P. (2007) 'The origins of cognitive dissonance: Evidence from children and monkeys', *Psychological Science* **18**: 978–983.
- Einstein, A. ([1950] 1954) 'On the Generalized Theory of Gravitation' in *Ideas and Opinions* (New York: Three Rivers Press).

- Evans, G. (1982) *The Varieties of Reference* (Oxford: Oxford University Press).
- Falvey, K. and J. Owens (1994). 'Externalism, Self-Knowledge, and Skepticism'.  
*Philosophical Review* **103**: 107–37.
- Fernández, J. (2013) *Transparent Minds: A Study of Self-Knowledge* (Oxford: Oxford University Press).
- Fiala, B. and Nichols, S. (2009) 'Confabulation, Confidence, and Introspection' (Commentary on Peter Carruthers) *Behavioral and Brain Sciences* **32**: 144–145.
- Finkelstein, D.H. (2003) *Expression and the Inner* (Cambridge, MA: Harvard University Press).
- Finkelstein, D.H. (2012) 'From Transparency to Expressivism', in *Rethinking Epistemology* volume 2, ed. G. Abel and J. Conant (Berlin: De Gruyter).
- Frankfurt, H. (1971) 'Freedom of the Will and the Concept of a Person', *Journal of Philosophy* **68**: 5–20.
- Freud, S. ([1916–1917] 1966) *Introductory Lectures on Psychoanalysis* (1916–1917), in *The Standard Edition of the Complete Psychological Works of Sigmund Freud*, volume 5, ed. and trans. J. Strachey in collaboration with A. Freud (London: Hogarth Press).
- Gallois, A. (1996) *The World Without, the Mind Within: An Essay on First-Person Authority* (Cambridge: Cambridge University Press).
- Gallois, A. (2011) 'Deflationary Self-Knowledge', in *Self-Knowledge*, ed. A. Hatzimoysis (Oxford: Oxford University Press).
- Gazzaniga, M. S. (1995) 'Consciousness and the Cerebral Hemispheres', in *The Cognitive Neurosciences*, ed. M. S. Gazzaniga (Cambridge, MA: MIT Press).
- Gazzaniga, M.S. (1998) *The Mind's Past* (Berkeley: California University Press).
- Gazzaniga, M.S. (2000) 'Cerebral Specialization and Inter-Hemispheric Communication: Does the Corpus Callosum Enable the Human Condition?', *Brain* **123**: 1293–1326.
- Gendler, T. S. (2008) 'Alief and Belief', *Journal of Philosophy* **105**: 634–663.
- Gertler, B. (2003) *Privileged Access: Philosophical Accounts of Self-knowledge* (Burlington, VT: Ashgate Publishing Company).
- Gertler, B. (2011a) *Self-Knowledge* (London: Routledge).
- Gertler, B. (2011b) 'Self-Knowledge and the Transparency of Belief', in *Self-Knowledge*, ed. A. Hatzimoysis (Oxford: Oxford University Press).
- Gertler, B. (2012a) 'Renewed Acquaintance', in *Introspection and Consciousness*, ed. D. Smithies and D. Stoljar (Oxford: Oxford University Press).
- Gertler, B. (2012b) 'Understanding the Internalism-Externalism Debate: What is the Boundary of the Thinker?', *Philosophical Perspectives* **26**: 51–75

- Gertler, B. (2016) 'Self-Knowledge and Rational Agency: A Defense of Empiricism', *Philosophy and Phenomenological Research* **95**: 1–19
- Gettier, E. (1963) 'Is Justified True Belief Knowledge?' *Analysis* **23**: 121–123
- Green, M., and Williams, J. (2007) 'Introduction' in *Moore's Paradox: New Essays on Belief, Rationality, and the First Person*, ed. M. Green and J. Williams (Oxford: Oxford University Press)
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E., and Banaji, M.R. (2009) 'Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity', *Journal of Personality and Social Psychology* **97**: 17–41
- Goldman, A. (2006) *Simulating Minds* (Oxford: Oxford University Press)
- Gopnik, A. (1993) 'How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality', *Behavioral and Brain Sciences* **16**: 1–14.
- Gordon, R.M. (1996) "'Radical" Simulationism', in *Theories of Theories of Mind*, ed. P. Carruthers P. and P. Smith (Cambridge: Cambridge University Press).
- Gordon, R.M. (2007) 'Ascent Routines for Propositional Attitudes', *Synthese* **159**: 151–165.
- Hacker, P. M. S. (2013) *The Intellectual Powers: A Study of Human Nature* (Oxford: Wiley Blackwell).
- Hájek, A. (2007) 'My Philosophical Position Says "p" and I Don't Believe "p"', in *Moore's Paradox: New Essays on Belief, Rationality, and the First Person*, ed. M. Green and J. Williams (Oxford: Oxford University Press).
- Hall L., Johansson P., Tärning B., Sikström S., Deutgen, T. (2010) 'Magic at the Marketplace: Choice Blindness for the Taste of Jam and the Smell of Tea', *Cognition* **117**: 54–61.
- Hardin, C. L. (1988) *Colour for Philosophers: Unweaving the Rainbow* (Indiana: Hackett Publishing Company).
- Harman, G. (2009) 'Skepticism about Character Traits', *The Journal of Ethics* **13**: 235–242.
- Harper, S. 'Introspection' in *Online Etymology Dictionary*. Retrieved from: <http://www.etymonline.com/word/introspection>
- Haslanger, S. (2008) 'Changing the Ideology and Culture of Philosophy: Not by Reason (Alone)', *Hypatia* **23**: 210–223.
- Heil, J. (1992) *The Nature of True Minds* (Cambridge, MA: MIT Press)
- Hirstein, W. (2005) *Brain Fiction: Self-Deception and the Riddle of Confabulation* (Cambridge, MA: MIT Press).

- Hirstein, W. (2009) 'Introduction: What is Confabulation?', in *Confabulation: Views from Neuroscience, Psychiatry, Psychology and Philosophy* (New York: Oxford University Press).
- Hobbes, T. ([1651] 1996) *Leviathan*, ed. J.C. A. Gaskin (Oxford: Oxford University Press).
- Horgan, T., and Kriegel, U. (2007) 'Phenomenal epistemology: What is Consciousness that we May Know it so Well?', *Philosophical Issues* **17**: 123–144.
- Hume, D. ([1748–51] 1975) *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, ed. L. A. Selby Bigge and P. H. Nidditch. (Oxford: Oxford University Press).
- Hume, D. ([1748] 1999) *An Enquiry Concerning Human Understanding*, ed. T. L. Beauchamp (Oxford: Oxford University Press).
- Hume, D. ([1739–40] 1978) *A Treatise of Human Nature*, ed. L. A. Selby-Bigge and P. H. Nidditch (Oxford: Oxford University Press).
- Hume, D. ([1739–40] 2000) *A Treatise of Human Nature*, ed. D. F. Norton and M. J. Norton. (Oxford: Oxford University Press).
- Hurlburt, R. T., and Schwitzgebel, E. (2007) *Describing Inner Experience? Proponent Meets Skeptic* (Cambridge, MA: MIT).
- Hunter, D. (2009) 'Belief, Alienation, and Intention', unpublished manuscript.
- Jackson, F. (1973) 'Is There a Good Argument Against the Incorrigeability Thesis?', *Australasian Journal of Philosophy* **51**: 51–62.
- Jackson, F. (1998) *From Metaphysics to Ethics: A Defence of Conceptual Analysis* (Oxford: Oxford University Press).
- James, W. ([1890] 1981) *The Principles of Psychology*, volume 1 (Cambridge: Harvard University Press).
- James, W. ([1903–1904] 1988) 'The Many and the One', in *Manuscript Lectures*, ed. F.H. Burkhardt and F. Bowers (Cambridge, MA: Harvard University Press).
- Johansson, P., Hall, L., Sikström, S., and Olsson, A. (2005) 'Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task', *Science* **310**: 116–119.
- Kahneman, D. (2011) *Thinking, Fast and Slow* (New York: Macmillan).
- Kant, I. ([1781] 1998) *Critique of Pure Reason*, ed. and trans. P. Guyer and A. W. Wood (Cambridge: Cambridge University Press).
- Kim, J. (2010) *Essays in the Metaphysics of Mind* (Oxford: Oxford University Press).
- Klein, S. B. (2014) *The Two Selves: Their Metaphysical Commitments and Functional Independence* (Oxford: Oxford University Press)
- Kriegel, U. (2008) 'Real Narrow Content', *Mind and Language* **23**: 304–328.

- Lau, J., and Deutsch, M. (2014) 'Externalism About Mental Content', *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/content-externalism/>
- Lawlor, K. (2009) 'Knowing What One Wants', *Philosophy and Phenomenological Research* **79**: 47–75.
- Levy, N. (2014) *Consciousness and Moral Responsibility* (Oxford: Oxford University Press).
- Levy, N. (2017) 'Implicit Bias and Moral Responsibility: *Philosophy and Phenomenological Research* **94**: 3–26.
- Lewis, D. (1973) *Counterfactuals* (Oxford: Basil Blackwell).
- Locke, John ([1690] 1975) *An Essay Concerning Human Understanding*, ed. P. H. Nidditch (Oxford: Oxford University Press).
- Ludlow, P. (1995) 'Externalism, Self-Knowledge, and the Prevalence of Slow-Switching', *Analysis*, **55**: 45–49.
- Lycan, W. G. (1996). *Consciousness and Experience* (Cambridge, MA: MIT Press).
- Lycan, W. G. (2004) 'The Superiority of HOP to HOT' in *Higher-Order Theories of Consciousness*, ed. R. Gennaro (Amsterdam: John Benjamins).
- McDowell, J. (1994) *Mind and World* (Cambridge, MA: Harvard University Press).
- McGeer, V., and Pettit, P. (2002) 'The Self-Regulating Mind', *Language and Communication*, **22**: 281–299.
- McKinsey, M. (1991) 'Anti-Individualism and Privileged Access', *Analysis* **51**: 9–16.
- Mele, A. R. (2001) *Self-Deception Unmasked* (Princeton: Princeton University Press).
- Mele, A. R. (2009a) 'Delusional Confabulations and Self-Deception', in *Confabulation: Views from Neuroscience, Psychiatry, Psychology and Philosophy*, ed. W. Hirstein (New York: Oxford University Press).
- Mele, A. R. (2009b) *Effective Intentions: The Power of Conscious Will* (New York: Oxford University Press).
- Mele, A. R. (2010) 'Weakness of Will and Akrasia', *Philosophical Studies* **150**: 391–404.
- Mellor, D. (1977) 'Natural Kinds', *British Journal for the Philosophy of Science* **28**: 299–312.
- Moore, G. E. (1903) 'The Refutation of Idealism', *Mind* **12**: 433–453.
- Moore, G.E. (1942) 'A Reply to My Critics', in *The Philosophy of G.E. Moore*, ed. P.A. Schilpp (Evanston: Northwestern University Press).
- Moran, R. (2001) *Authority and Estrangement: An Essay on Self-Knowledge* (Princeton: Princeton University Press)
- Moran, R. (2004) 'Replies to Heal, Reginster, Wilson, and Lear', *Philosophy and Phenomenological Research* **69**: 455–472

- Moran, R. (2012) 'Self-Knowledge, "Transparency", and the Forms of Activity', in *Introspection and Consciousness*, ed. D. Smithies and D. Stoljar (Oxford: Oxford University Press)
- Nagel, J. (2013) 'Knowledge as a Mental State' in *Oxford Studies in Epistemology*, volume 4, ed. T.S. Gendler and J. Hawthorne (Oxford: Oxford University Press).
- Nagel, T. ([1970] 1978) *The Possibility of Altruism* (Princeton: Princeton University Press).
- Nagel, T. (1974) 'What Is It Like to Be a Bat?', *The Philosophical Review* **83**: 435–450.
- Neta, R. (2011) 'The Nature and Reach of Privileged Access' in *Self-Knowledge*, ed. A. Hatzimoysis (Oxford: Oxford University Press).
- Nietzsche, F. ([1887] 1998) *On the Genealogy of Morals: A Polemic*, ed. and trans D. Smith (Oxford: Oxford University Press).
- Nisbett, R. and Wilson, T. (1977) 'Telling More than we Can Know' *Psychological Review*, **84**: 231–295.
- Nguyen, A. M. (2004) 'Davidson on First-Person Authority', *Journal of Value Inquiry* **38**, 457–472.
- Nugent, P. M.S. (2013) 'Belief Perseverance', in *PsychologyDictionary.org*. Retrieved from <https://psychologydictionary.org/belief-perseverance/>.
- Onishi, K. H., and Baillargeon, R. (2005) 'Do 15-month-old Infants Understand False Beliefs?', *Science* **308**: 255–258.
- Parent, T. (2017) 'Externalism and Self-Knowledge', *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/self-knowledge-externalism/>
- Paul, S. K. (2012) 'How we Know What We Intend', *Philosophical Studies* **161**: 327–346.
- Payne, B. K. (2006) 'Weapon Bias: Split-Second Decisions and Unintended Stereotyping', *Current Directions in Psychological Science* **15**: 287–291.
- Petitmengin, C., Remillieux, A., Cahour, B., and Carter-Thomas, S. (2013) 'A Gap in Nisbett and Wilson's Findings? A First-Person Access to Our Cognitive Processes', *Consciousness and Cognition* **22**: 654–669.
- Pitt, D. (2004) 'The Phenomenology of Cognition, or What is it Like to Think that P?' *Philosophy and Phenomenological Research* **69**: 1–36.
- Plato (1997) 'Phaedrus' in *Plato: Complete Works*, ed. J.M. Cooper (Indianapolis: Hackett).
- Prinz, J. (2012) *The Conscious Brain: How Attention Engenders Experience* (New York: Oxford University Press).

- Putnam, H. (1975) 'The Meaning of "Meaning"' in *Philosophical Papers*, volume 2, (Cambridge: Cambridge University Press).
- Quine, W.V.O. (1966) 'On Simple Theories of a Complex World', in *The Ways of Paradox* (New York: Random House).
- Reed, B. (2010) 'Self-Knowledge and Rationality', *Philosophy and Phenomenological Research* **80**: 164–181.
- Rey, G. (2008) '(Even Higher-Order) Intentionality Without Consciousness', *Revue Internationale de Philosophie* **62**: 51–78.
- Rey, G. (2013) 'We are not all "Self-Blind": A Defense of a Modest Introspectionism', *Mind and Language* **28**: 259–285.
- Rorty, R. (1970) 'Incorrigibility as the Mark of the Mental', *The Journal of Philosophy* **67**: 399–424.
- Rosenberg, A. (2016, July 18) 'Why You Don't Know Your Own Mind' in *The New York Times*. Retrieved from <https://www.nytimes.com/2016/07/18/opinion/why-you-dont-know-your-own-mind.html>
- Rosenthal, D. M. (1986) 'Two Concepts of Consciousness', *Philosophical Studies* **49**: 329–359.
- Rosenthal, D. M. (1997) 'A Theory of Consciousness' in *The Nature of Consciousness: Philosophical Debates*, ed. N. Block, O. Flanagan, and G. Güzüldere (Cambridge, MA: MIT Press).
- Rousseau, J. J. ([1755] 2004) *Discourse on the Origin of Inequality*. (Minola, New York: Dover Publications).
- Rowbottom, D. P. (2007) "'In-Between Believing" and Degrees of Belief', *Teorema* **26**: 131–137.
- Ryle, G. (1949) *The Concept of Mind* (New York: Barnes and Noble).
- Saemi, A. (2015) 'Aiming at the Good', *Canadian Journal of Philosophy* **45**: 197–219.
- Sauret, W. and Lycan, W. G. (2014) 'Attention and Internal Monitoring: A Farewell to HOP', *Analysis* **74**: 363–370.
- Scaife, R. (2014) 'A Problem for Self-Knowledge: The Implications of Taking Confabulation Seriously', *Acta Analytica* **29**: 469–485.
- Scanlon, T. (1998) *What We Owe to Each Other* (Cambridge: Harvard University Press).
- Schnall, S., Haidt, J., Clore, G. L., and Jordan, A. H. (2008) 'Disgust as Embodied Moral Judgment', *Personality & Social Psychology Bulletin* **34**: 1096–1109.
- Schooler, J. W. and Schreiber, C. A. (2004) 'Experience, Meta-Consciousness, and the Paradox of Introspection', *Journal of Consciousness Studies* **11**:17–39.

- Schroeder, T. (2015) 'Desire', *The Stanford Encyclopedia of Philosophy*.  
<https://plato.stanford.edu/entries/desire/>
- Schroeter L. (2017) 'Two-Dimensional Semantics', *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/two-dimensional-semantics/>
- Schwitzgebel, E. (2008) 'The Unreliability of Naive Introspection', *The Philosophical Review* **117**: 245–273.
- Schwitzgebel, E. (2010) 'Acting Contrary to Our Professed Beliefs, or The Gulf Between Occurrent Judgment and Dispositional Belief', *Pacific Philosophical Quarterly* **91**: 531–553.
- Schwitzgebel, E. (2011) *Perplexities of Consciousness*. (Cambridge, MA: MIT Press).
- Schwitzgebel, E. (2012a) 'Introspection, What?' in *Introspection and Consciousness*, ed. D. Smithies and D. Stoljar (Oxford: Oxford University Press).
- Schwitzgebel, E. (2012b) 'Self-Ignorance' in *Consciousness and the Self: New Essays*, ed. J. Liu and J. Perry (Cambridge: Cambridge University Press).
- Schwitzgebel, E. (2014) 'Introspection', *Stanford Encyclopedia of Philosophy*.  
<https://plato.stanford.edu/entries/introspection/>
- Schwitzgebel, E. (2015) 'Belief', *The Stanford Encyclopedia of Philosophy*.  
<https://plato.stanford.edu/entries/belief/>
- Searle, J. R. (1983) *Intentionality: An Essay in the Philosophy of Mind* (Cambridge, UK: Cambridge University Press).
- Searle, J. R. (2001). *Rationality in Action*. (Cambridge, MA: MIT Press)
- Searle, J. R. (2015). *Seeing Things as They Are: A Theory of Perception* (Oxford: Oxford University Press).
- Shah, N. and Velleman, J. D. (2005) 'Doxastic deliberation' *The Philosophical Review*, **114**: 497–534.
- Shoemaker, S., (1963) *Self-Knowledge and Self-Identity* (Ithaca: Cornell University Press).
- Shoemaker, S., (1994) 'Self-Knowledge and "Inner Sense"', *Philosophy and Phenomenological Research* **54**: 249–314.
- Shoemaker, S. (1996) *The First-Person Perspective and Other Essays* (Cambridge, Cambridge University Press)
- Shoemaker, S. (2003) 'Moran on Self-Knowledge', *European Journal of Philosophy* **11**: 391–401.
- Siewert, C. (2012) 'On the Phenomenology of Introspection', in *Introspection and Consciousness*, ed. D. Smithies and D. Stoljar (Oxford: Oxford University Press).

- Silins, N. (2012) 'Judgment as a Guide to Belief' in *Introspection and Consciousness*, ed. D. Smithies and D. Stoljar (Oxford: Oxford University Press).
- Silins, N. (2013) 'Introspection and Inference', *Philosophical Studies* **163**: 291–315.
- Simons, D. J., and Rensink, R. A. (2005) 'Change Blindness: Past, Present, and Future', *Trends in Cognitive Sciences* **9**: 16–20.
- Sinhababu, N. (2017a) *Humean Nature: How Desire Explains Action, Thought, and Feeling* (Oxford: Oxford University Press).
- Sinhababu, N. (2017b) 'Desire and Aesthetic Pleasure', *Australasian Philosophical Review* **1**: 95–99.
- Smithies, D. (2012) 'A Simple Theory of Introspection', in *Introspection and Consciousness*, ed. D. Smithies and D. Stoljar (Oxford: Oxford University Press).
- Smithies, D. (2013) 'On the Unreliability of Introspection', *Philosophical Studies* **165**: 1177–1186.
- Smithies, D. (2014) 'Can Foundationalism Solve the Regress Problem?' in *Current Controversies in Epistemology*, ed. R. Neta (New York: Routledge).
- Smithies, D. and Stoljar, D. (2012) *Introspection and Consciousness* (Oxford: Oxford University Press).
- Smithies, D. and Stoljar, D. (2012) 'Overview: *Introspection and Consciousness*' in *Introspection and Consciousness*, ed. D. Smithies and D. Stoljar (Oxford: Oxford University Press).
- Stalnaker, R. (1981) 'Indexical Belief', *Synthese*, **49**: 129–151.
- Stalnaker, R. (1999) *Context and Content* (Oxford: Oxford University Press).
- Stich, S. (1983) *From Folk Psychology to Cognitive Science* (Cambridge, MA: MIT Press).
- Strawson, G. (1994) *Mental Reality* (Cambridge, MA: MIT Press).
- Strawson, G. (2009) *Selves: An Essay on Revisionary Metaphysics* (Oxford: Oxford University Press).
- Tenenbaum, S. (2010) 'Akrasia and Irrationality' in *A Companion to the Philosophy of Action*, ed. T. O'Connor and C. Sandis (Chichester: Wiley Blackwell).
- Trivers, R. (2011) *The Folly of Fools: The Logic of Deceit and Self-deception in Human Life* (New York: Basic Books)
- Tye, M. (1995) *Ten Problems of Consciousness* (Cambridge, MA: MIT Press).
- Tye, M. (2009) *Consciousness Revisited: Materialism without Phenomenal Concepts* (Cambridge, MA: MIT Press).
- Valaris, M. (2014) 'Self-Knowledge and the Phenomenological Transparency of Belief', *Philosophers' Imprint* **14**: 1–17.

- Vargas, M. (2005) 'The Revisionist's Guide to Responsibility', *Philosophical Studies*, **125**: 399–429.
- Velleman, J. D. (2000) *The Possibility of Practical Reason* (Oxford: Oxford University Press).
- Warfield, T. (1992) 'Privileged Self-Knowledge and Externalism are Compatible', *Analysis*, **52**: 232–237.
- Wegner, D. M. (2002) *The Illusion of Conscious Will*: (Cambridge MA, Cambridge University Press)
- Wells, G. L. and Petty, R. E. (1980) 'The Effects of Overt Head Movements On Persuasion: Compatibility and Incompatibility of Responses', *Basic and Applied Social Psychology* **1**: 219–230.
- Wheatley, T. (2009) 'Everyday Confabulation', in *Confabulation: Views from Neuroscience, Psychiatry, Psychology and Philosophy*, ed. W. Hirstein (New York: Oxford University Press).
- Whiting, D. (2012) 'Does Belief Aim (Only) at the Truth?', *Pacific Philosophical Quarterly*, **93**: 279–300.
- Williams, B. (1973) 'Deciding to Believe', in *Problems of the Self: Philosophical Papers 1956–1972* (Cambridge, UK: Cambridge University Press).
- Williamson, T. (2000) *Knowledge and its Limits* (Oxford: Oxford University Press).
- Wilson, T. D. (2002) *Strangers to Ourselves: Discovering the Adaptive Unconscious* (Cambridge: Harvard University Press).
- Wittgenstein, L. ([1967] 1981) *Zettel*, second edition, ed. G.E.M Anscombe and G.M. von Wright, trans. G.E.M Anscombe (Oxford: Basil Blackwell).
- Zahavi, D. (2005) *Subjectivity and Selfhood: Investigating the First-Person Perspective* (Cambridge, MA: MIT Press).
- Zahavi, D. (2011) 'Empathy and Direct Social Perception: A Phenomenological Proposal', *Review of Philosophy and Psychology* **2**: 541–558.
- Zimmerman, A. (2007) 'The Nature of Belief', *Journal of Consciousness Studies* **14**: 61–82.
- Zimmerman, A. (2008) 'Self-Knowledge: Rationalism vs. Empiricism' *Philosophy Compass* **3**: 325–352.