Self-knowledge, Deliberation, and Memory

Robert Anthony Davies

PhD

University of York

Philosophy

March 2017

Abstract

In this thesis, I argue that the epistemology of memory is a neglected and useful explanatory resource in the philosophical treatment of problems associated with introspection. Our vocabulary of introspection is uniquely confused and unhelpful among accepted modes of knowledge, and a number of options are available to improve matters: we might (i) attempt to make introspection terms more scientifically respectable, or (ii) offload some of introspection's duties onto other modes of knowledge. This latter approach has received a good deal of attention in contemporary self-knowledge literature that aims to explain introspection 'economically'. However, one approach to an economic theory of introspection has largely escaped detailed attention in recent literature. This, broadly, is Ryle's (1949) suggestion that memory can explain much of what we take to be introspection. The aim of this thesis is to gauge the extent to which that suggestion, in general, might be helpful in resolving some intractable problems in the literature on self-knowledge.

To motivate the inquiry, I point to a far-reaching convergence in our thinking about introspective failure and memory failure, and suggest that this convergence extends to introspective success. To test the extent of the convergence, I categorise a range of purported features of introspective thought into a set of desiderata that can be set against a theory to measure its success. I then argue that the epistemology of memory plays an important role in explaining how a prominent theory of self-knowledge meets a number of these desiderata; that memory can explain or contribute to explanations of the three main desiderata; and that a theory of self-knowledge constructed around a standard case of recollection can meet most if not all of the desiderata for a theory of self-knowledge.

Table of Contents

ABSTRACT 2

TABLE OF CONTENTS 3

INTRODUCTION 6

ACKNOWLEDGEMENTS 9

DECLARATION 10

CHAPTER 1: FAILING TO KNOW OUR MINDS 11

INTRODUCTION 11
1. FAILING TO KNOW OUR LEXICON 15
2. IMPROVING OUR LEXICON 22
3. MEMORY AND INTROSPECTIVE FAILURE 25
3.1 COGNITIVE BIAS 28
3.2 CONFABULATION 30
3.3 CHOICE BLINDNESS 31
4. VARIETIES OF SELF-KNOWLEDGE FAILURE 37
5. MEMORY AND INTROSPECTIVE SUCCESS 39
CONCLUSION 41

CHAPTER 2: DESIDERATA FOR A THEORY OF SELF-KNOWLEDGE 43

INTRODUCTION 43
1. PECULIARITY AND INTROSPECTION 47
1.1 INTROSPECTION AND WAYS OF KNOWING 50
1.2 PECULIARITY 55
2. VARIETIES OF IMMEDIACY 58
2.1 EVIDENTIAL AND EXPLANATORY IMMEDIACY 61
3. EPISTEMIC SECURITY 68
3.1 TRADITIONAL NOTIONS OF EPISTEMIC SECURITY 69
3.2 COGITO-LIKE JUDGEMENTS 73
3.3 FALLIBILITY, ONTOLOGICAL DISTANCE, AND PARITY 75
3.4 MODEST APPROACHES TO EPISTEMIC SECURITY 77
4. UNIFORMITY, ECONOMY, AND TRANSPARENCY 80
4.1 UNIFORMITY 80
4.2 ECONOMY 83
4.3 TRANSPARENCY 84
5. ADDITIONAL DESIDERATA 86
5.1 AGNOTIC ACCESS 87
5.2 PRESERVED ACCESS 88
5.3 EVALUATIVE ACCESS 89

5.4 Self-Blindness 90
Conclusion 90

CHAPTER 3: TRANSPARENCY, DELIBERATION, AND MEMORY 93

Introduction 93
1. Transparent self-knowledge 94
2. Transparent deliberation 96
2.1 Objections to the deliberative view 98
3. Transparent inference 101
3.1 The contamination objection 104
4. Mnemic and deliberative schemas 107
4.1 Immediacy and recall 108
4.2 Deliberation and recall 108
4.3 New belief formation 109
5. Transparency and Uniformity 111
5.1 Procedures for belief, desire, and intention 113
6. Transparency and doxastic deliberation 119
7. Memory, evidence, and beliefs about evidence 123
Conclusion 128

CHAPTER 4: MEMORY AND SELF-KNOWLEDGE 130

Introduction 130
1. Kinds of memory 131
1.1 Factual memory 136
1.2 Distinctive features of self-knowledge 138
2. Memory and first-person peculiarity 138
2.1 The intuitive peculiarity of memory 139
2.2 Recalling first-order and second-order beliefs 141
2.3 Peculiarity and the doxastic schema 147
3. The Immediacy thesis 152
3.1 Psychological immediacy and inference 155
3.2 Epistemic immediacy and inference 157
4. Epistemic Security 162
4.1 Epistemically beneficial memory effects 164
4.2 Epistemic security and positive epistemic status 166
4.3 Immunity from error 170
4.4 Deliberation-resistant attitudes 171
Conclusion 173

CHAPTER 5: DOXASTIC RECOLLECTION 175

Introduction 175
1. The Transparency–Transition problem 177
1.1 Inference and reflection 178
1.2 The Transparency–Transition assumption 183
1.3 Deliberation and the TT assumption 186
1.4 Transparency and the TT assumption 189
2. Transparency and self-attribution 191
3. Doxastic self-knowledge as recollection 194

3.1 NON-DELIBERATIVE DOXASTIC SELF-KNOWLEDGE 197
4. MEETING SELF-KNOWLEDGE DESIDERATA 199
4.1 PECULIARITY, IMMEDIACY, AND EPISTEMIC SECURITY 199
4.2 UNIFORMITY, ECONOMY, AND TRANSPARENCY 202
4.3 ADDITIONAL DESIDERATA 203
5. QUERIES AND OBJECTIONS 205
CONCLUSION 208

CONCLUSION AND FURTHER WORK 209

APPENDIX 1: CHOICE BLINDNESS AND INTROSPECTIVE COMPETENCE 212

BIBLIOGRAPHY 251

Introduction

In this thesis, I argue that the epistemology of memory is a neglected and useful explanatory resource in the philosophical treatment of problems associated with introspection. I explore a way of improving matters with regard to our vocabulary of introspection that has been neglected in much contemporary literature. Given a broad-ranging convergence in our thinking about introspective failure and memory failure, I suggest that it might be productive to investigate how far that convergence extends to introspective success. A positive response to that question is likely mean a positive response to the question of whether memory can explain what is thought special or distinctive about self-knowledge.

Broadly, the aim of the thesis can be cast in Ryle's terms: it is to gauge the extent to which memory might 'carry the load of which introspection has been nominated the porter' (Ryle 1949). The correspondingly broad claim is that a good deal can be gained by considering our views of memory when thinking about problems usually associated solely with self-knowledge. The following breakdown marks the main specific claims for which I argue in each chapter.

In chapter one, I argue (i) that memory can play an important role in explaining what we often think of as introspective failure, and (ii) that it is worthwhile investigating whether that convergence in our thinking extends to cases of introspective success. I suggest (iii) that a promising way for the inquiry to proceed is by outlining the desiderata against which the success of a theory of self-knowledge might be measured.

In chapter two, I consider a range of features thought to account for what is special or interesting about knowledge of our own minds and present a list of desiderata—minimal criteria, ideal desiderata, and additional criteria—against which any theory of self-knowledge might be measured. In chapter

three, I set a prominent approach to self-knowledge—the Transparency approach—against a number of these criteria. I argue that a particular view of memory plays an important role in explaining the epistemic desiderata for doxastic self-knowledge on one version of the approach. I conclude (iv) that the epistemology of memory plays an important part in explaining introspective success on such a view, and (v) that this strengthens the case for an inquiry into the extent to which memory might be explanatory in the domain.

Chapter four explores the question of whether memory might explain, or contribute to the explanation, of the three minimal criteria from chapter two. These are: Peculiarity, Immediacy, and Epistemic Security. I argue that there are a surprising number of options available for each of the three desiderata, and suggest (vii) that this merits the construction of a theory that can be tested against the full list of desiderata.

In chapter five, I highlight a problem in the literature on Transparency accounts (e.g. Byrne 2011; Boyle 2011) that appears to negatively affect the ability of a number of accounts to meet a number of specific criteria. The problem, I argue, is an assumption based on too strong a conception of the requirement of Transparency accounts to explain self-ascription. I argue that we should reject the assumption, and that a weaker conception of the requirement (a) better captures the range of data that needs to be explained, and (b) better fits standard conceptions of the important features of Transparency.

With a weaker conception of the self-ascription requirement in place, I outline what I take to be plausible candidate for a standard case of doxastic self-knowledge that can be appropriately described in memory terms. I construct the outline of a theory around this case and set it against the full range of desiderata from chapter two. I show that such a theory can fare well against all three kinds of desiderata. In my final section of the chapter, I offer a range of pre-emptive responses to questions and concerns, some of which will form the basis of further work. The overall conclusion for the thesis is two-fold: (viii) a theory of self-knowledge with the epistemology of memory playing the main

explanatory role can be surprisingly successful when set against desiderata that one might commonly find in the self-knowledge literature—in at least some important cases, self-knowledge can be accurately described as a kind of remembering; (ix) and inquiry into the explanatory capacity of memory with regards to the domain of self-knowledge is able to shed light on a number of intractable problems in that literature.

Acknowledgements

This work is dedicated to the memory of my father R R Davies, to my mother J H Davies, to the rest of my family, and to anyone that seeks, but struggles to secure, a good education.

***

I would first like to thank my supervisor, Professor Tom Stoneham, for his excellent feedback and support (and space and time), without which finishing this thesis would not have been possible.

The support and patience of my family has been a source of stability throughout the period, for which I am very grateful.

I am grateful to Professor Paul Noordhof, and the members of the *Mind and Reason* research group, for valuable opportunities and invaluable feedback.

Much gratitude is due to the editorial staff of *Mind* at the University of York, in particular, Dr Barry Lee; to the administrative staff at the Department, especially Julie Kay and Carol Dixon; and to the Higher Education Academy. I am grateful to all of the students at the Department that I have had the privilege to meet and teach. I would like to thank Dr Anjana Raghavan for her patience, humour, and support.

Finally, I would like to acknowledge the invaluable contribution of all those who provided encouragement, support, affection, music, companionship, and stimulating conversation over the last few years.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as references. A brief overview of the first several sections of Appendix 1 has been published online at the *Imperfect Cognitions* website:
http://imperfectcognitions.blogspot.co.uk/2015/05/refining-our-understanding-of-choice.html. None of the other work has been published elsewhere.

1

Failing to Know Our Minds

Introduction

Why is knowledge of our minds so apparently easy to come by and so hard to explain? On some views, we can barely go wrong when conducting inquiries into our minds. On others we are pitiful, and introspection—or whatever special method we are supposed to deploy—is a kind of comforting illusion. If we are to accept introspection, and its cognate locutions, to represent an authentic way of knowing—that is, the kind of thing to which one can refer as an explanation of putative knowledge[1]—it is likely that neither of these views is correct since neither infallibility, nor pervasive error, help to provide such an explanation. Nevertheless, both positions lie on a broad spectrum of views, still defended in some form, about the reliability of introspection. When compared to other ways of knowing, the diversity of considered opinion with regard to reliability ought to be cause for consternation, especially as we seem prepared to accept a number of intuitions about the kind of access to our minds that introspection affords (see Ch. 2).

Part of the problem, no doubt, is that the term suggests—literally or metaphorically—a quasi-perceptual (e.g. Shoemaker 1994; Byrne 2005) phenomenon for which none of our usual cognitive apparatus is apparently fit. Given that the interior of the human body is not obviously a good place to look, the term is either misleading or we are in need of a dedicated cognitive faculty with a specific remit to extract meaningful data from what is, essentially, *meat*.

If introspection as 'looking inwards' seems implausible, then, perhaps we might see if looking outwards fares better (e.g. Byrne 2005, 2011a).

[1] This follows Cassam (2007a)

Although the suggestion sounds quirky, it retains some key elements of the quasi-perceptual view and has rightfully gained a good deal of traction. (After all, what could be worse than *conoscenza con carne*?) But explaining how one knows one's mind by looking outward brings its own challenges: Is it something that happens quickly and immediately or slowly and deliberately? How is thinking about the world supposed to tell us anything about our minds?

To analyse the problem we need to consider the basic ingredients of introspection and—assuming it is meant to issue in knowledge—these seem straightforward enough. Introspection, at base, must be (or involve) the correct, usually time-sensitive attribution of some fact of the matter about oneself, to oneself. On this description, one might expect our science of introspection to fall within the reasonable limits of any form of knowledge that deals with contingent matters. No-one genuinely claims we cannot get it wrong when it comes to our attempts to learn by hearing, seeing, smelling, or feeling, etc., and no-one—at least no-one offering a theory of perception—supposes we cannot get it right either. But our thinking about introspection does not seem to remain within these limits (or many limits at all). Theories leave subjects infallible, or wholly ignorant; introspection can be quick or slow, immediate or mediate; it can provide conclusions based on evidence, or based on nothing; it is either a form of looking inward, or a form of looking outward. And many of these options cannot be referred to in order to sufficiently explain the body of putative knowledge with which they are concerned.

Perhaps the sensible thing to do when faced with such situations in philosophy is to walk away. But since the problems of introspection intersect with many other areas of philosophy, this will only delay the inevitable. One might, then, look instead at the farthest ends of the spectrum to see which views, if any, can be ruled out. At the optimistic end of the spectrum are views that leave ordinary subjects with almost unimpeachable access to the contents of their minds. Such views are not widespread, but are defended. They are not, however, generally defended in an unrestricted form, and it is not clear that

anyone ever has defended them in that form (Greenough 2012; also Ch. 2). (I return to these below.) Very pessimistic views are not particularly widespread either, but they are not in short supply and tend to claim corroboration by empirical evidence.[2] One particularly bleak view suggests that introspection is a confluence of unreliable processes that we cannot hope to untangle, and allows that almost any number of introspective methods combine to provide an ultimately useless capacity (see Eric Schwitzgebel 2008a, 2009).[3] A different kind of pessimism sees talk of introspection as a kind of simile in which subjects peering into an imaginary realm, when in fact the way we know ourselves is much the same way as we know about others (see Ryle 1949).

The bleak view leaves us with an unusual excess of both good and bad fortune: ostensibly competing epistemic and metaphysical theses turn out to be correct, except in the one respect for which they have been developed—to explain knowledge in the domain; an abundance of lemons, with no prospect of lemonade. The second view leaves us on a par with others when it comes to knowing our minds, and it has persisted well (e.g. Carruthers 2011) despite its success usually relying on 'slight' exaggerations (see Byrne 2012). Without these exaggerations, it looks as though there is some advantage, however modest, to having the mind one is inquiring about.

Over the next few chapters, I explore an idea that has accompanied this latter view but has not persisted nearly as well in the literature, perhaps because it has never received a thoroughgoing treatment, comes in scattered references,[4]

---

[2] The claims against which empirical evidence is effective are a matter of some controversy (see e.g. Stoneham 2004, §6).

[3] It is not made explicit that this is the considered overall view. It is, however, a combination of views expressed over the two cited works.

[4] Alex Byrne (2011a) refers briefly and suggestively to *recalling that p* as being part of the standard case of doxastic self-knowledge (see Ch. 3) (in private correspondence Byrne informed me that chapter of his forthcoming book on self-knowledge will discuss memory in greater detail, but the content—at the time of writing this work—is not ready to be cited); Eric Schwitzgebel (2009) refers to memory both in relation to Transparency accounts and self-scanning mechanisms (see Ch. 4). Psychological literature on memory covers more ground, although it is not always clear which philosophical questions this might answer. For example, Martin Conway (2005) writes about the Self Memory System (SMS). Literature on self-consciousness and memory can be traced back some way, with a discussion of Thomas Reid's (1785) essay unfortunately finding no natural home in this work. More recently, J. L. Bermúdez (2012, 2013, forthcoming) has discussed the inter-dependence of memory and self-consciousness (see Ch. 5), although

or is left implicit. One of its clearest articulations is in Ryle's *Concept of Mind*, where it can be read as two independent thoughts: firstly, that we have given the term 'introspection' much to do that can be explained by our ordinary faculties; and secondly, that much of its work should have been apportioned specifically to the genuine faculty of memory.[5] The first thought has flourished in recent approaches to self-knowledge, even those that afford a good deal of first-person privilege.[6] The second has been largely forgotten.

The aim here is not to analyse the thought as it appears in Ryle, or elsewhere—Ryle was railing against Cartesianism, and a number of other articulations are too brief or suggestive for detailed commentary—although these attempts do provide a helpful point from which to begin. The aim here is to see if there is something to that second thought, in general, that may help shed light on some intractable problems in our thinking about self-knowledge. If there is something in it, a number of philosophers may have been looking down the wrong end of the telescope. If there is not much in it, I may have added to the list of puzzles in the study of this domain, since it is increasingly apparent that our thinking about self-knowledge and our thinking about memory are tightly connected.

My main claim is a positive one: a good deal can be gained from considering our views of memory when thinking about problems usually associated solely with self-knowledge. This is especially pertinent given what has now become a staple assumption when theorising about the domain; that we should first exhaust the explanatory powers of our normal faculties before inventing new ones. I argue that a surprising amount of what we take to be special about self-knowledge can be, at least partially, explained by independently plausible views of memory. Success will either suggest that there

---

discussions of self-consciousness and self-awareness do not always sit naturally beside discussions of self-knowledge as it appears in much of the literature discussed here.

[5] Ryle usually refers to 'retrospection' (e.g. 1949, p. 148) and has a specific use for the term not intended here.

[6] For example, Alex Byrne's view, it has been noted (see Carruthers 2011) leaves us with something close to infallibility.

is an important explanatory role for memory in what we take to be special about self-knowledge, or that self-knowledge is not, strictly speaking, special in those ways, at least as we tend to articulate them.[7]

In this initial chapter, I aim to motivate the inquiry. In §1, I suggest that one way to think about our poor handling of the concept of introspection is to see it as a kind of lexical problem that might be solved by offloading some duties onto knowledge domains with more developed lexica. We have already seen such an attempt in the casting of self-knowledge as either as kind of perception, or a perception-like capacity. Put baldly, since that approach has seen mixed results, we might try thinking of it as a kind of remembering. In §2, I provide initial grounds for suspecting that memory might be a suitable mode of knowledge to investigation by pointing to a connection between our thinking about self-knowledge failure and memory. In section §3, I separate varieties of self-knowledge failure to isolate those that tend to be of philosophical interest, and in section, §4, I consider two examples of self-knowledge failure that are regularly of philosophical interest. In the final section (§5) I outline an appropriate methodology for investigating whether memory can play a role in introspective success as well as failure.

1. Failing to know our lexicon

The casual reader may be forgiven for failing to see the history of philosophical interaction with introspection as a list of glorious triumphs. If one were to take caricatures of traditional positions seriously, it appears that philosophers lost their appetite for questioning assumptions on the matter, and left the ordinary humans with unrestricted and totally reliable access to their minds (Greenough 2012, §1). Of course:

[7] This latter point has a precedent in the argument that the purported *baselessness* of self-knowledge cannot be what is special about knowledge in that domain, since baseless knowledge is either impossible or commonplace (see Cassam 2009).

> It is not easy to see how this ever could have been plausible. In any case, it [is] widely seen as having been refuted by Freud, as well as by recent psychological research of a distinctly non-Freudian character which seems to show both that a vast amount of what goes on in a person's mind is completely inaccessible to that person's introspective consciousness, and, what is equally shocking to Cartesian preconceptions, that when people do report on their own mental operations, these reports are often wrong (Shoemaker 1990, p. 183).

A better explanation of this perplexing error will refer to developments in the perceived targets of introspection—particularly in the twentieth century—making traditional views of our abilities seem usually optimistic, with views aimed at describing the immediate objects of the conscious mind now understood as including deep and sub-conscious processes (see Moran 2001, p. 5). Claims—for example, *omniscience* and *infallibility* with regards to our own mental states—have been confused and conflated (see e.g. Stoneham 2004;[8] Ch. 2); and explicit exceptions and concerns are largely ignored (e.g. both Descartes and Kant express reservations in some cases, see Ch. 2); and items rarely thought to be the objects of especially secure first-person judgements are used as counter-examples to introspective competence in general (e.g. character traits and irretrievably unconscious activity, see e.g. Schwitzgebel 2008a). Meanwhile, the weight of empirical research (e.g. Nisbett and Wilson 1977)—doubtless aided by these factors—has appeared forceful against more optimistic views, even sometimes against the explicit advice of those who carried out the research (see Appendix 1, §1). And because empirical findings about introspection are usually subject to competing conceptual analyses (see e.g. Johansson et al. 2006, p. 675), the philosophical implications of the results are not all that easy to determine.

The problem, in short, has taken on a lexical flavour: our lexicon of introspection terms is inconsistent, inefficient, and unhelpful in explaining the

[8] Stoneham (2004) uses different terms, however, but the point is the same (see Ch. 2).

mode of knowledge it is intended to describe, especially when compared to its counterparts (e.g. in *outer* perception). To put it Ryle's (1949) way, we can 'back up' our assertions by saying we see, hear, feel, smell and taste, but saying that we introspect does not really work as a 'final appeal' (p. 143).[9] Our lexicon for introspection does not simply need a bit of tidying up around the edges, it fails to match even the sparsest lexicon of its cousins in outer perception, that is, in terms of its ability to explain knowledge: its objects are confused, its methods fail to explain general or specific requirements of theories in the domain, and there appear to be no upper or lower limits in terms of supposed asymmetries with knowledge in other domains.

These observations should be concerning if introspection's direct associations with knowledge are to be retained. But assuming for the moment that the term introspection—however currently confused—is being used to pick out a body of putative knowledge that cannot be easily explained otherwise, there are a number of options for reducing the confusion. These include: (i) making its terms more scientifically respectable; and (ii) identifying any duties that can be offloaded onto capacities that explain knowledge in other domains. Both are exercises in tidying up our knowledge lexica more generally, so some brief remarks about comparative lexica will help to elucidate the task.

A number of explanations are available for the difficulties facing our lexicon in the introspection case. The first, which we have already touched upon, is a kind of error theory about introspection: our introspection terms fail to refer, or at least fail to refer to a way of knowing. On this view, introspection differs from other putative ways of knowing in so far as it does not live up to the title. However, our knowledge lexica vary in quality and depth between and within ways of knowing in general. And this goes against the idea that failures to refine the introspection lexicon are down to its failure to be a genuine mode of knowledge. On a second view, introspection is genuine but we have hitherto failed sufficiently to develop a lexicon despite no serious obstacle

[9] Ryle (1949) asks whether being 'conscious' or 'even vividly conscious' should be a final appeal (p. 143).

to success. We have, as it were, somehow neglected to fill out the requisite detail to the appropriate standard. This, I take it, would be the kind of explanation a concerned 'non-error' theorist might offer. But the extent and duration of the problem count against that view: the difficulty stretches back to the Greek treatment of the issue, and the Greek response to the Delphic injunction—*gnóthi seaftón*—varied significantly. Within the space of a single Socratic dialogue it ranges from, 'understanding myself as a whole person', to understanding 'my' *psyche*, and then to understanding '*the*' *psyche*; onto 'an analysis of what each person is persuaded of and why' (Griswold 1996, pp. 3f.). Elsewhere, it includes knowledge of one's own ignorance. Aristotle's thoughts on the matter extend the tally by introducing the suggestion that success for the self-knowing subject comes only by nurturing the right kind of human relationships (see e.g. *NE* 1170b5–7). And despite attempts in some intervening literature to restrict the scope of dialogue to specific classes of judgement or states about which the subject's thinking can be especially privileged or secure (e.g. Burge 1996), the temptation to think that the questions of self-knowledge must be broad enough to incorporate some of the Greek concerns has a tendency to re-surface (see e.g. Cassam 2014; Schwitzgebel 2008a). These factors do not rule out the explanation, but they do suggest a particularly stubborn form of problem.

To account for this stubbornness, a third explanation would see introspection and its cognate terms revealing a genuine way of knowing, but with the relevant facts difficult to code in one or more natural languages. Here we can make use of the term 'ineffability' as it has recently been employed in research on sense vocabularies (Levinson and Majid 2014). The facts relevant to introspection, on this explanation, could be either 'strongly ineffable' or 'weakly ineffable'. Facts 'may be strongly ineffable in the nomological sense that in principle no language can express them, or merely in an empirical sense in that no languages actually do so'; they are *weakly* (or *relatively*) ineffable where they resist:

> codability in language L by any of … three measures (i.e. coding in L is linguistically impossible, inefficient, or inaccurate or a combination thereof), compared either to some other domain in the same language, or the same domain in another language. (Levinson and Majid 2014, pp. 410–12)

There is little reason to suspect that introspection facts are strongly ineffable, and this result would likely rule out introspection as a plausible way of knowing (at least as I have described it so far). But there does appear to be a case for weak ineffability. Coding has been less efficient and accurate than other domains of knowledge in English, and plausibly in other related European languages. So the third option is a good partial explanation of why our lexicon in this domain is persistently feeble. To illustrate, in English, it seems easier to 'linguistically code colors than (non-musical sounds), sounds than tastes, tastes than smells' (p. 415), thus smells might be considered 'relatively ineffable compared to colors in English' (p. 412).[10] Since colour lexicons can vary dramatically across cultures, we might also want to say, for example, that in Yélî Dyne—a language with very few colour terms—colours are relatively ineffable in Yélî Dyne compared to English (p. 412). Introspection terms, in English and other languages, have been certainly less efficient and consistent than the outer senses above, so we might want to suggest that introspection terms are *weakly* or *relatively* ineffable in those languages. What is missing from the explanation is an understanding of *why* introspection facts might be difficult to code. Here, a number of suggestions are already available, and I propose another.

One way of explaining why a domain is relatively ineffable is to focus on cognitive architecture. Facial descriptions are difficult to code in most languages despite our outstanding capacity for facial recognition. One explanation of this in terms of cognitive architecture is that facial recognition is an 'ancient mammalian trait' associated with a specific region of the brain, and is thus a

[10] We can ignore that the fact some of these observations are confessedly based on 'introspection' (e.g. p. 415) since the authors offer a range of examples based on empirical research.

'classic Fodorean encapsulated module' that deals specifically with visual input (p. 417). Thus we can name faces, but not describe them (*Ibid.*). Alternatively, we might consider that the problem is not one of encapsulation, but of 'competition for resources': this will occur when two faculties use the 'similar neural networks', one attempt to explain why the olfaction lexicon is weaker than the colour lexicon (p. 418).[11] Both are perhaps plausible explanations for specific differences in sensory lexicons, but are less plausible in the introspection case. A slightly more promising explanation predicts that only sensations and processes that consciously accessible will be accessible to language. Since the processes of introspection may be among many aspects of mental life not consciously accessible, we should expect our lexicon to be poor. This explanation, however, would fail to account for differences between sense lexica, because many of the processes of seeing, feeling, and hearing, etc., are similarly inaccessible to consciousness—it is usually the results of such processes that are consciously accessible, and this is true both for introspection and those senses.

A better approach suggests where 'under-developed coding of sensory domains may reflect lack of cultural preoccupation',[12] we may 'trade off relative ineffabilities in single sensory fields, with high codabilities of recurrent cross-modal types' (p. 421) depending on need and relevance. So, in the absence of a specific (perhaps industrial, or technological) need, a culture, or cultures are happy to deal in objects as they 'come packaged by nature with their cross-modal properties (a ripe mango has a certain color, taste, texture, shape, etc.)' (*Ibid.*).

Thinking about introspection in broadly this way might, firstly, help to explain why thinking in the domain tends to be muddled and, secondly, point to

[11] The explanation is due to Lorig (1999; in Levinson and Madjid 2014) is that both language and odours 'share complex temporal signatures' (pp. 417–8).

[12] The 'Vatican has maintained a reference set of 30,000 labeled color chips since the 1500s, in order to reproduce mosaics' whereas cultures without colour technology show 'limited abstract color terminology … people without musical instruments (like the Rossel Islanders) may have little use for a metalanguage for tone, cultures with limited cuisine … may not be conducive to elaborate vocabularies of taste and smell' (Levinson and Majid 2014, p. 421).

a potential solution. Unlike many of our genuine perceptual abilities, there is no general cultural requirement for a detailed vocabulary of introspection, and given some basic observations about the limits of natural languages[13] we might speculate that some of our linguistic capacity has been traded off in exchange for capacity in areas of more pressing linguistic need. We do appear to hold a range of assumptions and practices around introspection, and these assumptions and practices are probably *good enough* for us to go about a daily business. But all this leaves us with a range of phenomena with poorly individuated properties in that domain. In the case of the mango, at least the data are cross-modal and all sensuous. In the introspection case, the muddle will likely include both sensuous content (e.g. visual and olfactory information) and non-sensuous content (e.g. memory and memories, various forms of thinking and reasoning, and imagination). If we have, mistakenly or otherwise, joined these various elements together into single capacity or faculty, then it will be no surprise that our thinking can produce vastly different results. This is not to suggest another kind of error theory about introspection. There may be some unique element—on top of the elements listed above—that leaves introspection worthy of a separate name. But it goes some way to explaining why attempts to refine our understanding of introspection, and its cognate terms, may have been frustrated by ignoring the contribution made by some of the main elements.

In short, the particulars of *de se* thinking have been insufficiently or inefficiently coded into a functional introspection vocabulary. In part, this is likely to be because there has been no general cultural requirement for a detailed lexicon in the domain. But this has led us to adopt one that is made up of a variety of sensuous and non-sensuous phenomena. Identifying and untangling some of these elements is the business of forthcoming sections. In particular, I aim to identify and focus upon one element that appears to be bound up in *de se* thinking and suggest that the contribution it makes can

[13] Although language is generative, 'working vocabularies are relatively small (say of the maximum order of 50,000 producible items)' (Levinson and Majid 2014, p. 420).

improve our understanding introspective failure. If it does, it will be worth considering whether it can improve our understanding of introspective success too.

2. Improving our lexicon

In the introduction, I suggested that introspection appears to be an unusual way of knowing. In §1, I offered an explanation for this. Before proceeding, it will help to clear up a matter that is not often not made explicit—namely, what it is for something to be a 'way of knowing'. Earlier, I implied that for knowledge in any domain to be deserving of the title, then it must have something to add in response to questions like, 'How does she know?'. If, in pointing our selected terms, they fail to shed any light on this kind of question, then we ought to ask ourselves whether what we have pinpointed is a route to substantial epistemic success. Having such an assumption made explicit will provide a measure against which it is possible to weed out potential non-starters. Quassim Cassam offers a useful notion:

> **Φ**-ing that P is a way of knowing that P just if it is possible satisfactorily to explain how S knows that P by pointing out that S **Φ**s that P. (Cassam 2007a, p. 339)

Using this as our measure, we might check which of our two options promises the most success.

The first option was to attempt to make introspection terms more scientifically respectable. A number of attempts have been made along these lines. 'Self-scanning' mechanisms are one such attempt that sees subjects with a monitoring mechanism capable of scanning certain kinds of mental state (e.g. Nichols and Stich 2003). Here, I will briefly outline another option in keeping

quasi-perceptual thinking about knowledge in this domain. One can begin by splitting our perceptual apparatus between the outer and the inner senses. 'Inner sense', in this case, is not the variety that meets with Ryle's disapproval, but refers to physiological indicators of internal events and processes. Whereas our external senses can collectively be termed 'exteroception', we can group internal senses under the heading 'interoception'. Interoception is the sense (or senses) 'of the physiological condition of the body' (Tsakiris, Tajadura-Jimanez and Constantini 2011). They proceed by means of stretch receptors, chemoreceptors, and the like, for low-level processes, and are often 'managed preconsciously' (*Ibid.*). However, they also encompass conscious sensations such as the fullness of the bladder, hunger, and nausea (Garfinkel and Critchley 2013, p. 231).

In the philosophical literature, such matters are rarely treated as targets of introspection, although they are clearly the targets of self-knowledge, broadly conceived. And, assuming the proper targets of introspection have underlying physiological changes detectable via such senses, we have a potential strategy for improving matters with regards to this mode of knowledge. Initial steps, largely in the field of cognitive science, and typically with regards to emotions, have been made in this respect. Emotions are a good candidate on which to model this kind of mechanism, especially if one adopts a theory upon which they depend on 'cognitive interpretations of physiological changes' (Seth 2013, p. 565; Garfinkel and Critchley 2013). If one can *infer* from these physiological changes (e.g. Seth 2013), then one has the beginnings of 'scientifically' appealing way to fill out the introspection lexicon that can be referred to in explaining how S knows that *p* (i.e. where *p* is a proposition about S's emotional state). Of course much more detail will be required if such mechanisms are to satisfactorily improve how we see introspection. But the devil, in this case, is more likely to be found in the *scope*. It may be reasonable to suppose such a mechanism is plausible with regards to knowledge of our basic emotions such as

anger—which are thought to have 'universal' physiological correlates[14]—and it is potentially promising for a range of sensations (itches, tickles, and some pains), but things look less promising when one steps outside of that limited range. Mapping physiological changes to their interoceptive is counterparts worryingly complicated for emotions such *disdain* and *contempt*, and finding the physiological–inferential mechanism for beliefs and intentions looks implausible given the potentially infinite number available. (Which receptors, for example, indicate that one intends to go the nearest store and buy the smallest packet of wooden clothes pegs, as opposed to the second nearest store for the third smallest packet of plastic clothes pegs?)

Although this is not the only way one may attempt to refine the introspection lexicon with (broadly) scientifically respectable terms, it demonstrates a difficulty facing any attempt to cast the self-knowledge in terms of physiological events and processes: knowledge of many 'introspective targets' is complicated and fine-grained, and so it is hard to generate plausible explanations in purely physiological, or even quasi-physiological, terms. Because many of our mental states are not as easily mapped onto our physiology, adopting this strategy will likely require either an artificially restricted range of target objects, or the substantive revision of other—namely our folk psychological—vocabularies.

An alternative option was to offload some of introspection's duties onto other ways of knowing—that is, to develop or refine our self-knowledge vocabulary by seeing how much explanatory work can be done elsewhere. This option has precedent in a number of theories of self-knowledge that aim to explain self-knowledge *economically*—that is, only by reference to capacities employed for knowledge in other domains. This approach in general has received a good deal of attention in contemporary literature (see e.g. Shoemaker 1994; Moran 2001; Byrne 2005; Fernández 2014), but here I would

[14] 'Paul Ekman and colleagues … showed that some specific emotions, which they named basic emotions, appear to be expressed in the same way in every human culture where this has been tested. In particular, they found that basic emotions produce the same patterns of changes in the face' (Zamuner 2013, p. 183).

like to focus on one aspect of the approach that has not hitherto received a great deal of attention; namely, by investigating the explanatory role that memory can play for knowledge in this domain. Although it has not received a great deal of attention, there are an increasing number of clues to indicate it is a connection is worthy of investigation. In the next section, I point to an idiosyncrasy in our general thinking about introspective 'failure' that is telling in this respect: we sometimes seem to bundle together cases of memory failure and self-knowledge failure.

3. Memory and introspective failure

In the broadest sense, forgetting one's own age and shoe size are self-knowledge failures, although they are more likely to be of interest to clinicians, psychologists and cobblers than philosophers. One might worry that disregarding such failures and successes plays into a kind of Cartesianism that sees mental life and ordinary physical attributes as somehow separate issues (see e.g. Byrne 2011a, p. 201). However, this need not be the case. One's knowledge of such details can usually be sufficiently explained in exactly the same way in our own case, in the case of others, and the world in general. Despite the protestations of a few (e.g. Ryle 1949; Carruthers 2011), it is not so easy to sufficiently explain how Sarah knows she intends to stop eating meat on New Year's Day. Narrowing the scope of relevant cases, and untangling different kinds of self-knowledge failure will be an important part of reducing confusion in the domain (see §4).

In a more restricted sense, on both commonsense and (some) theoretical views, self-knowledge failure and memory failure tend to converge in a number of cases. Peter stays out for after-dinner drinks later than he expressly intends, although this comes as no surprise to his companions. (He regularly expresses the intention to go home early, and regularly stays out late.) Assuming that

Peter genuinely believes he will go home early on each occasion, making sense of his forming the intention to Φ in the face of clear defeating conditions—such as a high probability that he will ψ instead—poses an interesting conundrum. One explanation is that all or most of the times that Peter ψs when intending to Φ are temporarily unavailable as evidence at the moment he forms the intention. And one way in which they might be unavailable is that they are not retrievable at the crucial intention-forming moment. Certainly other explanations are available: he might recall them and deem them irrelevant, for example. However, the memory explanation chimes anecdotally: often when one challenges a subject to recall the times they managed to Φ rather than ψ in such situations, their conviction that they will Φ weakens notably. It is at least a plausible explanation in cases of this kind, that whenever a subject forms an intention to Φ despite the presence of clear defeating conditions—namely that almost every time he intends to Φ, he ψs instead—that a failure to recognize those defeating conditions is a failure to retrieve the relevant information. And in such cases, were it not for this failure, the subject would not self-ascribe the intention, for Peter cannot intend to do what he knows he will not do. Thus we can describe such cases as memory failures. We might also want to call them introspective failures, since the subject has failed to know something important about himself. But how can this case be contrasted with the shoe size case? All of the data that Peter ought to have used when thinking about his intention is available third-personally. Thus if it is a failure, it is not a failure to discern his intention by means of introspection, but to discern his character. It is introspective failure of a kind, but introspective knowledge of character has rarely been thought to come with any significant degree of security.[15] (We might be decidedly worse at judging our characters than observers in many cases.)

[15] Gertler (2011b) has noted that what passes for privileged knowledge of character, might well be access to one's intention, for example, to be courageous.

Let us amend the example. It is possible, having formed and self-ascribed the intention to Φ, that one gets oneself in a position to ψ instead. Sometimes one does this not due to a change of heart, or a form of akrasia: Φ–ing is what one has a mind to do, and despite putting oneself in a position to ψ, when one is *re*minded or queried upon what one intends, one realizes one's error. For example, Susan intends to go home, but is enjoying after dinner drinks and rolls with the feeling. After a while she is *re*minded, and corrects course. Again it seems natural to describe such events in terms of memory—Susan intended to go home early, and forgot for some time. It is also possible to describe it as introspective failure:

> At t_1, S believes that S is in mental state M (and is in M); at t_2, S does not judge that S is in M (and is in M); and at t_3, S judges that S is in M (and is in M)

Again, of course, other descriptions are available. One might instead, suggest that Susan's intentions first changed and then changed back, although it is possible to reconfigure examples for the current reading for duration. And in many respects, this is not a surprising form of introspective failure: having climbed the stairs with a specific purpose in mind, one sometimes finds oneself nonplussed—that is, until one is back down again (I take it in such cases one cannot self-attribute the specific intention at each point, even if one attempts to); complicated intentions[16]—for instance with regards to one's life goals—often require a good deal of effort to keep in mind; and if, *per impossibile*, one had to constantly be mindful of all one's mental states, it would be difficult to get anything else done. We are not omniscient with regards our mental states,

[16] Lewin (1951) distinguishes between two 'concepts of forgetting' with regard to intentions. The first is 'the usual conception of memory' and the second relates to cases of intention that are 'not carried out'. The two are independent, although he recognizes they may be connected (p. 106). I am referring to the 'memory' variety, rather than the 'not carrying out' variety.

and we are forgetful. We can call this diachronic introspective failure for mental states.

Accepting the analyses of these examples for the sake of argument, we might suppose that a number of experiences can be appropriately called both memory failure, and self-knowledge (or introspective) failure (although this is not to suggest that the descriptions will always be equivalent). What unites the two kinds of case is the passage of time. In one case, the requisite data for a successful judgment is accumulated over time; in the other, the requisite data for consistent judgement concerns a state that persists over time. We might venture, at this point, to make a descriptive claim:

> (DSK) in diachronic cases, our thinking about self-knowledge and memory sometimes converges such that a failure or gap in one domain either is, or can be partially explained by, pointing to failure in the other.

One might think this is a quirk of the examples I have selected, or that instances of DSK are rare, so we ought to see if the DSK idiosyncrasy holds elsewhere.

3.1 Cognitive bias

Lists of human cognitive biases can be impressive reminders of the depth and range of our failings.[17] A 'good' list can be disconcerting enough for some to conclude that humans cannot be close relatives of the rational beings they are assumed to be.[18] Many cognitive biases are—or can be readily described in terms of—self-knowledge failure. A fairly harmless way to describe them that way, at least for the present purposes is to say that if a cognitive bias results in a subject making provably false statements about her mental states, events, or

[17] See e.g. the *Cognitive Bias Codex 2016* (Benson 2016).
[18] See e.g. Cassam (2014). Whether these are the right kind of examples to put pressure on claims to a subject's first-person privilege is a matter for elsewhere.

processes, then the subject has prima facie failed, introspectively speaking.[19] Lists of cognitive biases also provide us with a way to gauge the degree to which our thinking about memory failure and self-knowledge failure converge: an abundance of cognitive biases are, straightforwardly, memory effects. We can call these the *simple cases*. A small sample of simple cases of cognitive biases that are explicitly referred to as memory effects might include: confirmation bias; consistency bias; crypto-amnesia; hindsight bias; humour effect; memory inhibition; misinformation effect; mood congruent memory bias; peak–end rule; placement bias; rosy retrospection; source confusion; suggestibility; telescoping effect; the verbatim effect … and so on.[20]

Cognitive biases and memory effects serve as a kind of circumstantial evidence in support of (DSK), showing additionally that the convergence of thinking between the two domains goes beyond pre-theoretical thinking and into a significant range of empirical work in the area. Beyond that, the conclusions one can draw from the prevalence of memory phenomena in cognitive bias research are limited, unless one can conduct a meta-analysis of the literature, or an analysis on a case-by-case basis. The former may well be a worthwhile exercise, but it is not something that can be done here. Since the range of cognitive phenomena covered in cognitive bias research is broad, it is unclear how engaging in the latter would advance the main thesis under discussion. Instead, we might take this circumstantial evidence to be enough to move on to a related, but more interesting, question: whether memory failure can play a role in explaining self-knowledge failure for cases that are not recognised memory phenomena. In the following two sub-sections, I will outline two examples that indicate a positive response to that question.

[19] The prima facie rider allows for the elimination of cases that we would be disinclined to accept as a genuine self-knowledge failure, but as the case is purely illustrative, I will not attempt to fill out the details here.

[20] See e.g. the *Cognitive Bias Codex*, 2016.

3.2 Confabulation

In addition to the simple cases—in which cognitive biases are already recognised as memory effects—a number of cognitive biases that are not recognised as memory effects, are best explained as involving a failure, distortion or gap in memory. Confabulation is one such effect. As with a number of terms primarily studied in clinical cases—such as delusion (see Berlyne 1972)—definitions of confabulation tend to be constructed around the common source of data (see Appendix 1, §7), and so confabulations are often defined as 'false narratives or statements about the world and/or self that unintentionally arise due to some underlying pathological condition' (McVittie et al. 2014). Call this the Pathology view. Defining confabulation according to the Pathology view is a mistake. Empirical research tends to suggest that confabulation—or something very much like it—is widespread, and may even be the norm (see e.g. Hall et al. 2012).[21] In one heavily cited example (Nisbett and Wilson 1977), non-clinical participants over-selected items in arrays of identical nylon stockings due to position, while later denying the possibility that position might have affected choice when asked for their reasons (pp. 243f.). A key assumption behind the research is that participants ought either to recognise the role of position in their selection, or admit ignorance. This is diachronic self-knowledge failure with regards to one's reasons (or other decisive factors in one's decision-making).

Assuming the results are good, it would be foolish to reason from this finding about non-clinical participants—on the basis of the Pathology view—to the conclusion that cognitive pathology is widespread or the norm (even if it turns out to be true independently). We need not follow the definition too far down that particular rabbit hole. Perhaps providing a definition is too 'thorny' an issue (Sullivan-Bissett 2015), and we can work around the problem by pointing to a range of common, but neither necessary nor sufficient, features (p.

[21] In Appendix 1, I argue that this is a needlessly pessimistic conclusion given the evidence, however, the conclusion—however implausible it may currently seem—may be shown to be true elsewhere.

4). But, pace those who espouse such a view, there is little need to be timid in defining such phenomena either. The basis of a useful definition can be traced back over a century (see Bonhoeffer 1901), and recognises a link between a failure or gap in memory and a tendency to confabulate (see Appendix 1, §7). From this view, we might describe confabulations as 'statements or actions that involve unintentional but obvious distortions of memory' (Moscovitch and Melo 1997, p. 1018; following Berlyne 1972), a description still in currency. The view can accommodate clinical cases, because the cause of the memory failure can either be pathological or non-pathological, but it allows for non-clinical cases to be explained without resorting to mass attribution of morbidity (and it still de-stigmatises the phenomenon). (See Appendix 1, §7 for further discussion.)

3.3 Choice blindness

In addition to simple cases, and memory related phenomena such as confabulation, more complicated examples of introspective failure can also be best explained as involving a failure or gap in memory. One example is Choice Blindness. Choice Blindness sees subjects fail to know their minds in the following respect: when they make a choice, and offer reasons for that choice, they offer reasons that simply could not be their reasons for that choice (see Appendix 1, §1). In this respect it is similar to, and draws upon, the frequently cited, and sometimes misunderstood,[22] work by Nisbett and Wilson (1977). It combines that research with Change Blindness[23] research and, conjuring tricks—or 'magic' (Hall et al. 2010)—to reverse a subject's selection (in manipulated trials) and present them with the object they rejected as if it were their choice (see e.g. Johansson et al. 2008). Subjects are then asked to offer their reasons for the selection.

[22]See e.g. Schwitzgebel (2006)

[23] Change Blindness is an effect in which participants 'fail to detect changes in a scene when the change is accompanied by some other visual disturbance' (Johansson et al. 2008, p. 142).

In a study in which subjects were asked to select one of two faces on the basis of attractiveness, and then received their rejected choice as their own, the number of participants that detected the manipulation was low—'no more than 30% of all manipulated trials were detected' (p. 144)—even with unlimited time to explain their preference. A large majority of explanations were clearly confabulatory, since subjects explained their selection by referring to features 'not possessed by the initially chosen face' (Lopes 2014).

Numerous revisions have been made to address the methodological issues, and the effect appears to be robust. More recent versions of the studies suggest that the effect extends far beyond matters for which caprice might be excusable (e.g. simple matters of taste with no obvious repercussions), and to cases where we are generally held to be rationally criticisable, such as moral judgements (see e.g. Hall et al. 2012). For instance, subjects not only failed to notice when statements about moral positions were 'reversed' (p. 1), but argued 'unequivocally for the opposite of their original attitude' (p. 4). (See Appendix 1, §§2–3 for further details.)

The implications of this kind of research for our status as introspectively competent rational decision makers (see Appendix 1; Davies 2015) can be initially unsettling. If the research is sound, then not only do we most often fail to notice our preferences and attitudes have been manipulated, but all-too-easily we offer explanations, and even argue, for choices that were not—and are even directly opposed to—our own. Standard explanations of the results tend to corroborate the feeling that the research is damaging. Dominic Lopes (2014, p. 29f.) suggests we can choose one of two hypotheses to explain the effect:

(1) We do not choose for reasons; we choose and provide reasons. The manipulation merely brings this out by setting up an unusual situation where the reasons miss their target.
(2) Reasons offered for the choices do not 'target [participants'] initial choice and preference'. The belief that they chose x 'determines

their preference' and so the reasons offered accord with their eventual preference.

Both of these hypotheses do damage to our 'conception of rational decision making' (p. 30; Appendix 1, §4) because either our 'choices are not based on the reasons we give' or our attitudes are 'fickle' (Lopes 2014, p. 29f.). Either way our 'reasoning about decisions is post hoc' (p. 30).

Whatever the merits of the two hypotheses, a number of concerns count against them (see Appendix 1, §5). Not least among these is the concern that the conclusions reach significantly beyond what the data suggests. We might reasonably take the data to show that in many cases, non-clinical participants—and perhaps, therefore, the population more broadly—are willing to provide demonstrably false statements about reasons for their selections, in response to inquiries, and when they have failed to notice that their choices have been manipulated (*Ibid.*). But it is notable that a sizable proportion of participants detect the manipulation, and some offer statements that are true of their original choice but not true of the manipulated choice. These data are not irrelevant to an explanation of the phenomena. It is not irrelevant that some participants *do* perform the way one might expect of an introspectively competent decision maker, even if we are willing to accept that this happens somewhat less frequently than we might have thought. The two hypotheses on offer struggle to explain this feature of the data.

An alternative explanation relies upon the assumption that some of our cognitive processes are such that, from the first-person perspective, transitions between those processes can sometimes go undetected (see Appendix 1, §8). While such an assumption is not wholly uncontroversial, it is quite safe: it is allowable even on some traditionally optimistic views of our introspective capacities (see Ch. 2). If one is willing to accept this assumption, then an explanation of the effect can proceed via a search for the appropriate processes. Promising candidates are processes involving factual memory and those, at least

partially, involving deliberation. At least in the conditions of Choice Blindness research, both are activities that that manifest some recognition of a question, and are typically Transparent to factual inquiry (see Ch. 3; Appendix 1). The main difference between them is that deliberation also aims at resolving an issue, whereas in cases of factual retrieval one has already resolved it. The respective epistemologies of deliberation and factual memory already provide good clues as to why it would be that a subject might transition between them without detection, especially in response to inquiries into what the subject thinks.

Inquiries of the variety 'Do you think that *p*?' can be understood in more than one way—as an invitation to make up one's mind, or an inquiry into what one already thinks. But keeping these matters apart is not straightforward, since the questions 'Do I think that *p*?' and 'Is it the case that *p*?' are either first-personally indistinguishable (Edgley 1969, p. 90; in Moran 2001), or have a tendency to elide—with the former giving way to the latter (see e.g. Shah and Velleman 2005). The conditions under which one might successfully divine the attitudes one already has, therefore, can be expected to be limited at best. Whatever they are, they ought not include any making up of one's mind, which would risk contaminating the result of the inquiry 'by possibly altering the state one is trying to assay' (Shah and Velleman 2005, p. 507).

One way in which the participants of Choice Blindness research might have access to what they *already* think would be to pay attention to their 'brute' or 'spontaneous' responses.[24] But this can only be part of the explanation, at best, since many sounds we spontaneously make with our mouths are not good indicators of what we think.[25] There is something helpful in the thought, however: whatever form our responses take, if they are to be a good way of knowing what we already think, then they must be non-deliberative, since

[24] This is Shah and Velleman's (2005) concern about Transparent reasoning as a way of knowing one's mind. Moran (2012) argues that their view (a form of Neo-Expressivism) cannot explain self-knowledge.

[25] Moran (2011) makes this point against Shah and Velleman's (2005) form of Neo-Expressivism. (See Ch. 3 for a more in-depth discussion of what can be taken from this position.)

deliberation will risk contaminating the response whenever there is an internal change in the subject, or a change in her environment.

The remaining part of the explanation can be found in our most plausible epistemologies of factual memory. These epistemologies of factual memory have the following three features: (i) a prima facie epistemic authority for a subject to continue believing in the absence of defeating conditions (e.g. Owens 1999, p. 319); (ii) which allows a subject to relinquish her reasons after the attitude has been formed (p. 317); and (iii) phenomenological paucity (Teroni 2015; see also Ch. 4; Appendix 1, §10 for a more complete discussion of these features).

How does this help to explain Choice Blindness? A subject in Choice Blindness conditions forms or recalls an attitude when presented with a selection.[26] In either case, the subject can rationally retain that attitude while relinquishing her reasons for it, and it may even be the norm to do so (see Owens 1999, p. 317). Presenting the participant with something other than her original choice is sufficient to disrupt the prima facie authority of the initial attitude in any case, but asking for reasons will likely reveal a gap in the memory of an unguarded participant. Thus, at least in part, the subject begins to assess the features in her environment with a view to resolving an issue—namely, by considering the features that go in favour of one selection over another, rather than a process which aims at retrieving factual information (no longer a viable source of the reasons one is required to offer). This transition can occur undetected, and so considering the matter afresh risks contaminating the self-knowledge procedure as long as there has been a change in the individual or her environment. We know from the methodology of the research that manipulating the selections relies upon a change in the subject's environment, and so whenever a subject considers factors in favour of one choice over another, she faces the very real risk of reporting upon features that could not be

[26] It will depend upon the specifics of the study. In the study involving faces, for instance it is more plausible that the attitude is formed upon seeing the faces. In the moral attitudes case, it is at least possible that the subject recalls a prior stance rather than forming a new attitude.

the decisive reasons for her original choice. On the other hand, the case in which the subject 'remembers well' is the one in which she will behave much as we would expect of a rational decision maker.

This simple explanation of the data acknowledges the vulnerability of subjects to a specific variety of self-knowledge failure without accepting the catastrophic implications for our status as introspectively competent rational decision makers that are implied by the other two hypotheses.

So far, we have seen that our thinking about introspective failure and memory converges in commonsense cases and a significant range of simple cognitive biases. Thinking about the role that memory plays also helps to improve explanations of phenomena such as confabulation, and complicated cases of introspective failure such as Choice Blindness. Were it not for an extensive literature that mostly ignores memory in the discussing self-knowledge, it would be tempting to say that it should form an essential part of theorising in this domain.

Someone opposed to such a temptation is likely to object in the following ways: (i) commonsense examples are subject to a broad range of competing analyses; (ii) examples such as failure to know one's character are beside the point since judgements about character are not supposed especially secure, and many of the cognitive bias cases are closer in nature to character judgements than they are to cases usually found in the philosophical literature on self-knowledge; and (iii) all of the other cases (e.g. those involving confabulatory reporting of reasons) concern diachronic self-knowledge and so will be susceptible to memory effects.

For the purposes of argument, I am happy to concede points (i) and (ii). The use of these examples is mainly to illustrate the degree to which our vocabulary and, to a significant degree, our theorising about phenomena in these two domains coincides when we have not taken pains to keep them apart. The third issue is more substantive, for it effectively suggests that pointing to cases

of diachronic self-knowledge failure begs the question by allowing memory to come into play, and thus it deserves a more complete response (perhaps more complete than can be afforded here). However, it should be noted that truly synchronic self-knowledge is an illusory phenomenon on all but a small number of self-knowledge theories (i.e. those which suppose an immediate metaphysical acquaintance with mental states). Scanning (Nichols and Stich 2003), deliberation (Moran 2001; Boyle 2011), reflection (Boyle 2009), and inference (Byrne 2011a), are all diachronic in some respect, and the latter two appear to have well-defined temporal components (I return to this point in Ch. 5). The spirit of the objection, I take it, is that there are importantly different varieties of self-knowledge failure, and for the point about memory to hit home, the examples used must be of a specific variety traded within a certain kind of literature. With regards to this issue, I am sympathetic. In the next section, I outline some varieties of self-knowledge failure that should be kept apart.

4. Varieties of self-knowledge failure

On the basis of what has been considered so far, we might conclude that there is, in general, already an implicit association between self-knowledge failure and memory. Not only are the boundaries unclear (e.g. some memory failure is considered self-knowledge failure), but some clear cases of self-knowledge failure are best explained in terms of memory (i.e. considering the role of memory in these cases offers a better explanation than alternative explanations). (DSK) seems not only plausible, but fairly commonplace. The test of whether an investigation into memory's role will help shed light on some of the more intractable problems of self-knowledge will be, in the first instance, whether it can shed light on those failures that tend to be the focus of philosophical debate. As we have seen, there is not one view in this respect: Greek Philosophy was concerned with a range of matters including soul and character; Early-Modern

Philosophy was largely concerned with the objects immediately accessible to the conscious mind; and contemporary philosophy has mainly focused upon intentional states, such as beliefs and desires, but has entertained a much broader range. One way to narrow the scope of an investigation is to isolate those items in our mental lives for which our judgements are thought to carry some special weight or privilege.[27] Self-knowledge failure when it comes to those items is usually the stock and trade of philosophical discourse. With this in mind, we can differentiate between the following kinds of self-knowledge and their corresponding failures:

(1) Introspectively unavailable (e.g. a blemish on the back of one's head)
(2) Conflux (e.g. character traits, motives, emotions)
(3) Interoceptive (e.g. hunger, thirst)
(4) Process (e.g. decision-making, reasoning)
(5) Intentional states (e.g. beliefs, desires, intentions)
(6) Phenomenal character of experiences (e.g. seeing colours or objects)

Varieties (1), (2) and (3) will not form a significant part of the discussion. There is no perceived privilege for objects of variety (1); at best, perceived privilege for objects of variety (2) is minimal, and while judgements concerning (3) are likely to carry some privilege,[28] they are not generally discussed in the literature. Since first-personal judgements about processes (4) and intentional states (5), and the phenomenal character of experience (6), are thought, at least sometimes, to carry a distinct weight or privilege, these will be the focus of the discussion. So far I have suggested that memory is taken to provide all or part of the explanation for many examples of introspective failure. For others, memory has no explicitly cited role, but it can still aid in improving our understanding of where things go wrong. Despite the explanatory value of memory for thinking

[27] See Gertler (2011b, Ch. 3) for a helpful discussion of narrowing the scope of self-knowledge inquiry.
[28] I assume here that privilege can be contingent, such that without the correct advances in medical science, or the correct apparatus, the subjects experiencing these phenomena is in a position of privilege.

about introspective failure, not much time has been given to the question of whether the epistemology of memory can help to explain introspective success. A positive response to that question, I take it, would require that the epistemology of memory can help to explain what we take to be special about self-knowledge, and this might seem initially implausible. In the final section, of this chapter, I want to explore how we might go about finding a positive answer to that question.

5. Memory and introspective success

In considering memory's relevance to a range of introspective phenomena, we might consider two theses that appear in some guise the literature:

> (i) Memory plays a ubiquitous and indispensable role in human cognition.
> (ii) At least some features of memory and first-person thought coincide, or are importantly related.

The two theses can be taken as starting points for distinct approaches to answering the question of whether the epistemology of memory can help to explain introspective success. The first often goes unchallenged in philosophical literature: memory is thought to be a necessary, ubiquitous, and largely involuntary feature of human cognitive operations (see e.g. Owens 1999):

> Our memory is not one more informational device which we can use or not as we please: it is fundamental to all cognitive transactions, including any that would be involved in establishing the reliability of memory itself. (Owens 1999, p. 313)

'Working' memory is required for inference, and for the 'stream of consciousness' by which, on some accounts, we access our ongoing thoughts (see e.g. Carruthers 2015). 'Procedural' memory allows me to type this sentence without looking at the keyboard. And memory has a role in sensory perception—allowing for 'persistence of vision' in events that would otherwise seem broken, segmented, or static: a series of 'frozen images interspersed with brief periods of darkness' is seen as a 'continuous scene' at the cinema; a solitary glowing ember moved around in the dark can be seen as shapes, patterns, or letters (see Baddely 1999, p. 11). On a number of views, 'almost everything is memory' (Teroni 2005, p. 7), and because memory plays a fundamental and ubiquitous role in human cognition, a full suspension of one's reliance on memory (if possible) would result in a suspension of one's capacity for intellectual change (cf. Owens 1999).

Striking clinical cases—demonstrating, for instance, the effects of chronic failure in short-term/long-term memory transfer—hint at just how alien human cognition would be without reliance on memory. And even in those cases many elements of the memory system (short-term, procedural, and sensory memory) remain in working order. Memory is not only fundamental to 'normal' cognition, but generally features in 'abnormal' cognition too.

Given its fundamental role in our cognitive operations, it may be tempting to reason as follows: (a) if memory is required for all 'cognitive transactions', it must be required for transactions that issue in self-knowledge; and so, sans memory (i.e. via a failure, gap, or distortion) self-knowledge involving any cognitive transaction would be impossible; (b) we can predict, for any given attempt at self-knowledge, that memory failure of the appropriate variety will result in, or contribute to, a commensurate failure in that attempt.

Both of these statements may turn out to be true. But they do not promise a great deal in terms of strategy for the current investigation. The first amounts to a transcendental defence of a thesis not especially under attack, and they both fail to provide insight into the potentially interesting convergence

between memory and the features of self-knowledge that occupy philosophical discourse in the field—they are more concerned with whether memory is a genuine enabling factor in self-knowledge than with the particular explanatory contribution it might make. In contrast, the second thesis (ii) is a promising starting point since it allows for a understanding of how memory might be involved in first-personal thought rather than whether it must be.

In order to pursue that line of thought, the first step is to evaluate and prioritise a range of approaches, assumptions, and specific theses that have accumulated in the literature on self-knowledge and introspection with a view to producing a list of desiderata against which the success of a theory of self-knowledge can measured.

Conclusion

I have suggested that our vocabulary of introspection is uniquely confused among accepted modes of knowledge. Part of the problem is that introspection terms pick out a mixed bag of phenomena including a range of sensuous and non-sensuous content. Among those elements frequently conflated with introspection discourse are memory phenomena. The extents to which memory phenomena occur in such discourse suggest that there is a convergence in our pre-theoretical thinking about memory and self-knowledge. The prevalence of memory effects in cognitive bias research suggests that this convergence stretches to theoretical thinking too. While one might object that the convergence bespeaks a lack of conceptual clarity on the issue, one might also inquire as to whether this convergence is explanatorily useful. To demonstrate that it is useful, I provided two examples of diachronic self-knowledge failure that are better explained when the role of memory is made explicit.

Despite the explanatory value of memory for thinking about introspective failure, not much time has been given to the question of whether

the epistemology of memory can help to explain introspective success. A positive response to that question would require that the epistemology of memory is able to help explain what we take to be special about self-knowledge. I have argued that such an investigation may be fruitful in shedding light on some of the more intractable problems in our theorising about self-knowledge. The first step in such an inquiry is to describe what features we take to be special about knowledge in this domain in such a way that the success of a theory of self-knowledge can be measured against them. This is the business of the next chapter.

2

Desiderata for a Theory of Self-Knowledge

Introduction

Philosophical interest in self-knowledge tends to focus on a number of asymmetries between first-personal and third-personal attribution, and on features of the first-person case that are purportedly distinctive of the domain.[29] The purpose of this chapter is to isolate these features and asymmetries, and to construct a set of desiderata that can be set against any theory of self-knowledge as a measure of its success. I divide the discussion of the desiderata into three varieties: 'minimal criteria', 'ideal desiderata', and considerations coherence and compatibility with cognitive phenomena[30] ('additional desiderata'). My primary concern is with characteristics that are thought distinctive in that they are not exhibited in other knowledge cases (i.e. of others' mental states, and the environment). These will need to be explained—or explained away—by any theory of self-knowledge to be considered successful.

A background assumption, and several discrete theses can be found in a number of prominent views about self-knowledge. The assumption is that we *can* have knowledge of our mental states, and I will leave this assumption largely unchallenged (see Ch. 1). The discrete theses concern how it is that knowing our mental states differs from other cases of knowledge.[31] I will refer to three of

[29] See e.g. Gertler (2011a, 2011b, Ch.1); Byrne (2011a); Fernández (2013).

[30] The terms are loosely due to Douglas (2013), although Douglas refers to 'coherence' issues as divided into 'internal' and 'external' concerns. A more in-depth treatment of the desiderata than is possible here might make better use of the internal/external distinction. However, for the present purposes, I believe 'coherence and compatibility' are sufficiently fine-grained.

[31] See e.g. Byrne (2011a). This latter carries with it another assumption: that knowledge of others' minds, and knowledge of the environment are broadly of the same kind. This assumption does deserve to be

these discrete theses as the PIE theses, or PIE conditions for a theory. They are: (P) *Peculiarity*; (I) *Immediacy*; and (E) *Epistemic Security*. These will serve as 'minimal criteria' for a theory of self-knowledge.

Each comes with associated questions: Must *Peculiarity* be explained through some special introspective faculty? If not, how it can be explained by reference only to those faculties required for knowledge in other domains? Is *Immediacy* to be read as psychological, epistemological, or explanatory? What are the implications of each? Is *Epistemic Security* a matter of degree or of kind? How secure is self-knowledge compared to knowledge in other domains? A discussion of issues arising from these questions makes up the first three sections of the chapter.

There are also questions of whether a theory ought to explain access to all mental states, occurrences, and processes in the same way (it has generally been assumed that it should);[32] whether this explanation ought to restrict itself to epistemic capacities deployed in knowledge of other kinds, or needs additional resources (it is generally assumed that metaphysical extravagance is to be avoided);[33] and whether a theory is compatible with the transparency of first-person thought (a recent pre-occupation in the literature). Call the first concerns about *Uniformity*, the second concerns about *Economy*, and the third about *Transparency*. Sometimes a theory is, or is not, *Uniform*, *Economical*, and compatible with *Transparency* as a straightforward consequence its structure (e.g. theories make use of a 'special faculty' will not be Economical), but this is not always the case.[34]

These considerations will not determine the success of a theory in the same way that minimal criteria will: there are good reasons to question whether self-knowledge is a unitary phenomenon admitting of genuinely uniform

challenged, although it will not be the main focus of the investigation here and will be touched upon relatively briefly.

[32] See Boyle (2011) for a list of approaches that have subscribed to this assumption, implicitly or explicitly.

[33] See e.g. Byrne (2011a) for a helpful discussion.

[34] Contrast Cartesianism, for instance (see Ch. 1) with Transparency approaches (see Ch. 3).

explanation;[35] some ostensibly good reasons for rejecting the notion of a special faculty (e.g. rejection of Cartesianism) do not count against all 'special-faculty accounts' of self-knowledge;[36] and views of Transparency and its role in self-knowledge vary considerably.[37] Thus Economy, Uniformity, and Transparency are *ideal desiderata*. Whether there are good reasons to reject any of these desiderata is explored below.

Coherence and compatibility considerations will also bear upon our assessment of a theory.[38] A theory must be internally coherent, of course, but it must also cohere with what we take to be right more generally (in the absence of principled grounds for revision). For the domain in question, massive introspective failure and/or incompetence, for instance, look likely to make impossible some epistemic capacities that we take ourselves to have (e.g. 'critical reasoning'),[39] and may bring into question whether we are the rational decision-makers we take ourselves to be.[40] We can accept that we probably over-estimate our competence in both of these respects, but we would need an especially robust reason to relinquish our claim to them completely. So, a theory must cohere with established cognitive phenomena unless it provides independent, principled reasons for rejecting them.

I address four of these considerations in section five (§5). The first is a capacity to recognize that which we do not believe or know—that is, a subject's capacity to identify her own ignorance or lack of belief.[41] Call it *Agnotic Access*. The second—purportedly an implication Epistemic Security (E)—suggests a subject should have uncontaminated access to a mental state in place prior to the initiation of a self-knowledge procedure.[42] I take this capacity to be a

[35] See e.g. Boyle (2011)
[36] Modern proponents of 'faculty accounts' include David Armstrong's version of the 'Inner Sense' theory (see e.g. 1981).
[37] For markedly different interpretations of Transparency, compare Byrne (2011a) and Boyle (2011).
[38] There is not enough space here to engage in a useful discussion on what makes a theory a good theory. See Douglas (2013) for a discussion about cognitive virtues in theory construction.
[39] See e.g. Burge (1994)
[40] This appears to be a consequence of some interpretations of empirical research such as Choice Blindness, although I argue against these views elsewhere (see Appendix 1)
[41] See e.g. Fernández (2013) for a discussion of the desideratum
[42] This 'implication' of Epistemic Security is discussed by Brie Gertler (2011a).

requirement for critical reasoning (although I do not argue for that conclusion here). Call this *Preserved Access*. The third, *Evaluative Access,* affords the subject the ability to reflect upon or assess her current mental states. The fourth and final consideration—whether or not a theory has damaging implications for our status as rationality creatures—I will call simply *Self-Blindness*.

Following a discussion of each of these three varieties of desiderata, I conclude the chapter with a summary of specific formulations of these desiderata against which the success of any theory of self-knowledge can be measured. (In the following chapter, I set a recent theoretical approach to self-knowledge against these desiderata.)

I have referred to three PIE theses, the explanation (or principled rejection) of which will serve as a main measure for the success of a theory of self-knowledge. These three minimal criteria—Peculiarity (P); Immediacy (I); Epistemic Security (E)—can be summarized as follows:

> (P) Self-knowledge is sometimes acquired via first-personally peculiar means
>
> (I) Self-knowledge is sometimes non-inferential or (in some other sense) immediate
>
> (E) Self-knowledge is epistemically secure compared to knowledge in other domains[43]

The main aim of the next three sections (1–3) is to arrive at a formulation of each thesis against which the success of a theory of self-knowledge can be measured. In each case I argue for a conception that is a suitable measure of a theory's success.

[43] The list is comparable e.g. to Gertler's (2011b, p. 60), although diverges in the inclusion of immediacy.

1. Peculiarity and introspection

In this section, I argue for a conception of Peculiarity on which it is a method or procedure by pointing to which it is possible, satisfactorily to explain how S knows her mental states, and that cannot be used satisfactorily to explain how S comes to know the mental states of others.

A prominent candidate for asymmetry between first-personal and third-personal access to mental states is the proposal that the former ordinarily proceeds via a different means to the latter (see e.g. Byrne 2011a; Gertler 2011b). While our methods of coming to know, for instance, one's character traits (e.g. courage) may be more or less the same in the first-person case and third-person case (Ch. 1), it is not easy to see how the contents of our thoughts and daydreams, for example, could be (see Byrne 2012). Generally, any perceived asymmetry goes in favour of the first-person, resulting in a kind of epistemic advantage, but while often related, the two issues they are independent in the sense that 'neither entails the other' (Byrne 2011a; also §1.3).

The method by which we come to know our own mental states is commonly—within certain strands of analytic philosophy—labeled simply 'introspection', although there is surprising diversity of opinion about what that means (Ch. 1).[44] The Cartesian view sees introspection as a kind of 'inward reflection', a view that has survived in some form to the early twentieth-century psychological methodology,[45] and no doubt still has a hold in commonsense psychology: there is a reason we are offered a penny for our thoughts (Gertler 2011b); and the daydreamer's unfocused gaze suggests she sees something we cannot see, etc. Despite its commonsense appeal, anti-Cartesian sentiment about self-knowledge has almost become dogma.[46] It is sometimes associated with an implausibly high degree of reliable access to a subject's own mind, and

[44] Compare 'Inner Sense' accounts (e.g. Armstrong 1981) with Evans-inspired Transparency accounts of introspection (e.g. Byrne 2005).
[45] See e.g. the work of Edward Titchener (1867–1927).
[46] Stoneham (2004) highlights some weak points in common cases against Cartesianism.

an implication that the subject is gazing into a, surely imaginary, 'second world' (see Ryle 1949).

Philosophical opinion has shifted fairly quickly from seeing the mind as 'totally open to introspection' to doubts about 'the very reality of introspection' (Moran 2001, p. 5); and from an association of mental events with those 'immediately present to consciousness' (see Freud 1915; in Moran 2001) to doubt about whether any *process* or *activity* of the mind is conscious (see e.g. Dennett 1969). Gilbert Ryle saw introspection-talk as a kind of 'simile' based upon a Cartesian misconception of mind:

> the fact that we generally know what we are about does not entail our coming across any happenings of ghostly status … there are no such happenings; there are no occurrences taking place in a second-status world, since there is no such status and no such world and consequently no need for special modes of acquainting ourselves with the denizens of such a world. (Ryle 1949, p. 143)

Ryle's (1949) view suggests no difference (in kind) between first-person and third-person access to the mind is required to explain self-knowledge. Combined with the view that there is also no principled difference in Epistemic Security in the first-person case, we can call this the Parity Thesis (PT):

> (PT) There are no differences in kind between first-person and third-person access to the mind, and first-person access affords no substantive epistemic advantage.

The thesis has contemporary support in the work of Quassim Cassam (2014) and Peter Carruthers (2011) and amounts to a rejection of both Peculiarity (P) and Epistemic Security theses (E). However, with regards to the former, it must at least be a 'slight exaggeration' (Byrne 2012) since proponents of (PT) frequently depend on 'silent soliloquy' (Carruthers 2011; Ryle 1949), 'retrospection' (Ryle 1949), or 'occurrent conscious propositional attitudes'

(Cassam, forthcoming) to explain our access to our own minds, and none of these is easily explicable in terms of third-person only access.

The implications of doubts about access to conscious processes and activities are also questionable: they rely on the identification of the objects of first-person awareness with sub-personal, neurological, or computational processes; not the objects most commonly associated with first-person access to the mind:

> For the object of first-person awareness (on any account of it) is not all of psychological life, but primarily the states of mind identified under the categories of what is sometimes called "folk psychology": the hopes and fears, pains and experiences we relate to each other in daily life, and not states or processes defined either neurologically or computationally. (Moran 2001, p. 7)

Failure to know these processes is self-knowledge failure of a kind, but does not obviously indicate general, or further specific, incompetence. Such processes can be viewed as 'irretrievably beyond the individual's control or consciousness' (Burge 2011, p. 325) even if one thinks that self-knowledge of mental states is robust and reliable (see e.g. Nisbett and Wilson 1977; Schwitzgebel 2006; Ch. 1, and Appendix 1). By contrast, the inhabitants of Freudian unconscious *are*, in principle, retrievable to the conscious mind (see Moran 2001, p. 7). So, neither sub-conscious processes nor Freudian unconscious present an immediate threat to robust self-knowledge of mental states achieved via some peculiar means and, notably, Parity theorists who are wary of introspection-talk, sometimes move its duties onto a more familiar cognitive faculty, and in so doing leave traces of unexplained self-knowledge.[47]

[47] See Alex Byrne's (2012) discussion of the problem as it applies to the thoughts of Ryle and Carruthers, and here in chapter one.

The challenge, then, is not to eliminate Peculiarity completely, but to formulate it in a way that makes room for both metaphysically austere and extravagant conceptions, allowing the formulation to function as a minimum criterion that can accommodate a range of approaches. It is worth taking a brief look at the diversity of candidates that such a formulation would need to accommodate.

1.1 Introspection and ways of knowing

The range of positions that could need accommodating by a formulation of Peculiarity (P) includes: *Acquaintance*, *Inner Sense*, *Rationalism*, *Transparency* approaches,[48] and *Simple* theory. In what follows I outline the main features of these views, and comment briefly upon their viability as (i) peculiar methods, and (ii) methods that allow us to retain the assumption (Ch. 1) that judgements about our mental states can lead to knowledge of those states.

On the *Acquaintance* view, introspection is conceived of as a direct, unmediated awareness of, or *metaphysical acquaintance* with, our mental states. Because contact with the states is unmediated, it is unlike (outer) perception in that there is no room for a causal process—such as light reflecting onto the retina—that is, the room that sceptical scenarios tend to exploit.[49] On this view, knowledge of our mental states is especially secure. It need not result in indubitability or infallibility (though some views do stress these qualities), but because it is knowledge of a *different kind*, at least some of the ways that knowing about the world may be subject to error do not apply when it comes to knowing our minds (see Gertler 2011b, Chs. 1, 4; Russell 1912).

A challenge for the approach is that the metaphysical contact that eliminates some kinds of error also appears to make the explanation of *how S knows*—our adopted hallmark of knowledge—quite difficult. Whatever direct metaphysical contact might explain, it does not sufficiently explain of how S

[48] The grouping is broad and loose. As Brie Gertler remarks that none 'of these theories is monolithic. Each admits of multiple versions, which differ in some details' (2011b, p. 4).
[49] See e.g. Schwitzgebel's (2008a) 'alien neuroscientist' that supplies us with misleading phenomenology.

knows about her mental states. So while it is easy to see how the method is first-personally distinctive, it is not easy to see how we might accept it as a genuine route to epistemic success.

Rather than focusing on the difference between our knowledge of our mental states and our knowledge of the world, the Inner Sense view, broadly speaking, point to the similarities (thus rejecting a central claim of Acquaintance theory). Introspection remains, ostensibly, a distinctly first-person method, since a subject has unique view of her own mind. But, in an important respect, there is no difference in kind between our knowledge of our own minds and knowledge of the environment. While a number of initial proponents of the approach held our access to our minds to be almost unimpeachable, if the mechanism is akin to 'outer perception' one can be wrong about one's mind in the same kind of ways that one can be wrong about objects in one's environment[50] (Gertler 2011b, Chs. 1, 5). Thus, even though the view appears to provide a plausible first-person means of knowing our minds, its ability to provide any special security is open to challenge: first-personal *privilege* does not follow from a dedicated 'inner' sense alone.

The Inner Sense view faces an additional challenge: in casting the subject's relation to her mental states as one of an observer observing independent objects, it leaves open the possibility of 'self-blindness'—that is, the inability of a creature to recognise its own thoughts and sensations. It has been argued that (Shoemaker 1994) self-blindness is not possible for rational creatures, and assuming the argument is successful,[51] the Inner Sense approach may provide a plausible model of peculiar first-person *access to minds*, but not for the kinds of creatures that we happen to be.

Inner sense theory leaves open another possibility: because the peculiarity is strictly contingent, one could, in principle, have the same access to

[50] A point that allows more contemporary versions of the approach to enjoy plausibility in the face of concerns about high degrees of privileged access to our minds (see e.g. Armstrong 1993).
[51] Whether the argument poses a direct threat to Inner Sense theory has been brought into question by Amy Kind (2003) and Brie Gerter (2011b).

others' mental states as one does to one's own mental states (see Armstrong 1993; Gertler 2015). (This need not concern us too much here, since we can restrict the inquiry to humans of this world.)

Inner Sense emphasises the role of the subject as 'observing' or 'detecting' her mental states—presenting the subject as standing in a passive relation to those states. This is a questionable fit for mental states such as pain and belief. Mere observation, especially on the perceptual analogy, fails to account for the apparent difference between my awareness that I see glass of water to the right of my computer screen, on the one hand, and my awareness of a dull ache in my right shoulder, or my deciding upon the best course of action, on the other (see Gertler 2011b; Schwitzgebel 2008b).

Rationalist approaches, in contrast, emphasise the subject's role as *agent* rather than *patient* (for at least some mental states). On this approach, we are authoritative about our mental states precisely because we are responsible for shaping them (Gertler 2011b). The processes leading to some mental states are *activities*—things that we do rather than things that happen to us (see Boyle 2009), and our authority, at least in some cases, arises from a process of 'making up one's mind', or deliberating (Moran 2001; Boyle 2009). But where the Inner Sense approach fails to account for an apparent *agency* when it comes to knowledge of some states, Rationalism fails to account for an apparent *patiency* for others: beliefs, desires, and intentions seem a good fit, but 'experiences', sensations, and sometimes daydreams (see Boyle 2009, 2011; Gertler 2011b) are prima facie unresponsive to reason: 'to be a thinker and an agent is to be capable of a kind of activity that stands in contrast to the passivity of sensation' (Boyle 2009).

Contemporary rationalist accounts (e.g. Moran 2001, 2012; Boyle 2009, 2011—usually give a central role to the '"transparency" of one's own thinking' in highlighting what is distinctive about self-knowledge:[52]

[52] The term is attributed by Moran to Roy Edgley (1969).

> Ordinarily, if a person asks himself the question "Do I believe that P?", he will treat this much as he would a corresponding question that does not refer to him at all, namely the question "Is P true?". And this is not how he will normally relate himself to the question of what someone else believes. (Moran 2001, p. 60)

The notion of 'introspection', here, takes a turn away from an 'inward' glance or reflection. The first-person and third-person cases remain different, because in the first-person case one's attitudes are immediately responsive to facts about the world, whereas facts about the world have no immediate relevance when we try to ascertain what someone else thinks. (In some cases, this difference is thought to allow for an epistemic distinctiveness or security.) Remarks from Gareth Evans (1982)[53] make the extent of the turn explicit:

> [In] making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward—upon the world. (Evans 1982, p. 255)

Transparency approaches vary widely: not all focus on the Rationalist's concerns; acknowledge, or seek to explain, an epistemic asymmetry; and not all aim at a *uniform* explanation of self-knowledge.[54] Of course, a solely outward-directed inquiry does not provide a sufficient explanation of self-knowledge, since the conclusion of such an inquiry refers to the *world*, not the *self*. So it is usually conceded that there must be some transition from the former conclusion to the latter. A pressing question for Transparency approaches is, therefore, how a conclusion about the world, *p*, can bring about a (rational) self-ascription: *I believe that p*. (See Chs. 3, 5.)

One Transparency approach (see Byrne 2005, 2011a) sees introspection as partially constituted by a world-to-mind inference. A conclusion about the

[53] See also Alex Byrne (2005, 2011a)
[54] Though this appears to be the aspiration of Evans (1982); Moran (2001); Byrne (2011a).

world, *p*, can form the basis for a self-attribution of a belief—*I believe that p*—if one follows a rule or schema that takes one from the former to the latter. (Whether or not this kind of Transparency approach meets other desiderata such as Uniformity is a matter for elsewhere.)[55] A clear challenge for this inferential Transparency view is whether a world-to-mind inference is plausibly a way of *knowing* since it is 'neither deductively valid or inductively sound' (Byrne 2011a). Boyle (2011) describes the inference as 'mad'.[56] (See Ch. 3 for a discussion of Transparency accounts.)

Shah and Velleman (2005) suggest one can know one's beliefs by 'posing a question *whether p* and seeing what one is spontaneously inclined to answer'; a process in which the question serves as a 'brute stimulus': 'One comes to know what one already thinks by seeing what one says—that is, what one says in response to the question *whether p* (p. 506).[57] But observation of one's response to brute stimuli alone looks insufficient for knowledgeable self-attribution of a belief unless one is also happy to accept sneezes (Moran 2012, p. 221) and other spontaneous oral noises to 'give voice' to such states.

'Simple theories' of introspection aim to offer an alternative to both 'observation' and 'inference' by suggesting that one can know one is in a conscious state by 'forming a belief on the basis of that very conscious state' (Peacocke 1998; Smithies, forthcoming). Being in pain alone, for instance—that is, the 'experience of pain'—can be 'a thinker's reason for judging that he is in pain' (Peacocke 1998, p. 72). An attempt (Smithies, forthcoming) to extend the scope of the approach to a range of states can be stated as follows:

[55] See Byrne (2011a) in favour, and further discussion in Ch. 3.

[56] Boyle contrasts this to a 'reflective' approach to explaining Transparency (Boyle 2012) on which the subject takes different sort of step: 'from believing P to *reflectively judging* (i.e., consciously thinking to himself): I believe P'. This step is not a 'transition between *contents*' but 'a coming to explicit acknowledgement of a *condition* of which one is already tacitly aware' (p. 5).

[57] The view is attributed to Dorit Bar-on (2004) although Bar-on's position here is not with traditional epistemic concerns, but with 'giving voice' (p. 318) to states in much the same way as one might say let out yelp in response to a painful stimulus.

> you know by means of introspection that you are in some mental state M when you believe that you are in M on the basis of a reason that is constituted by the fact that you are in M. (Smithies, forthcoming)

Thus, the explanation can be argued to extend to beliefs, desires, and intentions, and may offer an alternative answer to the Transparency question: one does not move from one state (a belief about the world) to another state (a belief about one's state of mind). Rather, one's being in the former state is our reason for believing that we are in that state.

1.2 Peculiarity

The foregoing discussion reveals, firstly, that use of the term 'introspection' for views as diverse as Acquaintance and inferential Transparency views suggest it has become shorthand for whatever method is used to assay one's mental states (see also Smithies, forthcoming); and secondly, highlighted the fact that a number of accounts appear to fall short of sufficiently explaining knowledge in the domain. With these two things in mind, characterising Peculiarity in light of this should (i) avoid prohibiting, at the outset, the possibility that any of these approaches could be correct given (ii) that just how they sufficiently explain of knowledge in the domain is given due attention.[58]

A sufficiently broad notion of Peculiarity is offered in Byrne's (e.g. 2005, 2011a) 'peculiar access':

> one has peculiar access to one's mental states if 'one has a way of knowing about one's mental states that one cannot use to come to know about the mental states of others' (Byrne 2011a, p. 202)

[58] We might add a third point that emphasises that the inquiry relates to humans of this world (thus allowing for contingent forms of Peculiarity). Cassam (2014) dedicates a good deal of space to emphasising the dangers of theorising on the mistaken basis that we are some other, more competent, species.

The majority of approaches outlined will be captured by this notion of Peculiarity as long as they turn out to be bone fide ways of knowing (although we have seen that there is doubt in some cases that aim to explain knowledge,[59] let alone those for whom traditional epistemic pressures are not a central concern).[60] Since Byrne's formulation captures a good range of theoretical approaches, and contains the knowledge requirement, the formulation has a good deal going for it. However, it offers little guidance on how we should think of the process as a way of knowing. Since whether a procedure can issue in knowledge is relevant to the characterisation of Peculiarity as I have cast it, some explicit guidance on this matter will be useful. To this end we can supplement Byrne's formulation with our explanatory view of knowledge (Ch. 1):

> **Φ**-ing that P is a way of knowing that P just if it is possible satisfactorily to explain how S knows that P by pointing out that S **Φ**s that P. (Cassam 2007, p. 2)

Combining these two features leaves us with a more comprehensive guide to what is required for a subject's access to her own mind to be peculiar:

> ***Peculiarity***—a method or procedure by pointing to which it is possible, satisfactorily, to explain how S comes to knows her mental states, and

[59] See Ch. 3 for a discussion of whether Alex Byrne's self-knowledge procedure is genuinely peculiar in this sense. Byrne describes a challenge to his approach that suggests it leads to a 'paradox': the procedure itself is a prima facie plausible route to knowledge and the Transparency of belief is 'obvious once pointed out', but the inference 'could hardly be worse, and so the second-order beliefs it yields will not be knowledge' (Byrne 2011a, p. 204).

[60] Some versions of Expressivism—for example, the Simple Expressivism attributed to Wittgenstein—deny 'that utterances like "I'm in pain" are even truth-apt, let alone reflect knowledge of one's mental states' (Gertler 2015, §3.8). Acquaintance approaches, on the other hand, leave no obvious room for the kind of cognitive achievement associated with knowledge, and thereby do not explain how knowledge by such means is possible. By analogy, I have direct contact with all 206 bones in my body, but this mere fact goes no distance at all towards explaining how it is that I come to know that number. Beliefs formed through such a method would be 'too close to their objects to qualify as genuine, substantive knowledge' (Gertler 2015).

that cannot be used satisfactorily to explain how one S comes to know the mental states of others.

To see whether this is a useful arrangement, we can set it against a number of the examples covered above.

I suggested above that one concern about Acquaintance theory is that direct metaphysical contact alone looks insufficient for knowledge. (We can assume the challenge has gone unanswered for the purposes of this discussion.) If such an acquaintance holds, it would evince a difference between first-person and third-person *access* to one's mental states, and this meets the second half of our formulation, thus establishing a first-person/third-person asymmetry. However, it would fail to meet the first half of the formulation because the contact in question does not, as yet, satisfactorily explain how S comes to know her mental states (pointing to metaphysical acquaintance alone is insufficient explanation).

One might think that Byrne's formulation of Peculiar Access already covers such cases—it stipulates that a method should be a way of knowing, and if Acquaintance is not, then it does not meet the standard. In this case, my amendment has added no value. An example that may help to demonstrate its value, however, is Byrne's own attempt to explain Epistemic Security. Byrne's rule takes the subject from *p* to *I believe that p* and is meant to be knowledge-conducive in part because it is strongly self-verifying (see Byrne 2011a) in the first-person case. Peculiarity is meant to be explained because following the rule in the third-person case 'will often lead us astray' (Ibid.). However, to say that that a method will often lead us astray is not the same thing as saying it is a method that *cannot be used*, only that it is a markedly less successful method. Playing badminton on one leg might lead to a largely poor run of results, but, unlike trying to play with no racket, it is still a way of playing badminton. So, at the very least, it is not clear that Byrne's account of the self-knowledge

procedure meets his own formulation of Peculiar Access.[61] However, it does look reasonably promising against my amended formulation. Assuming for the sake of argument that the world-to-mind inference is, in fact, knowledge conducive, one could satisfactorily explain how S comes to know about her mental states by pointing to the rule. On the other hand, one cannot not satisfactorily explain how S comes to know the mental states of others solely by pointing to the same rule.[62] The amendment I have introduced is modest, but it will produce different results than Byrne's formulation (including when it comes to Byrne's view), and it is one way to make the knowledge requirement explicit. From this point, by 'Peculiarity', I mean the formulation above.

2. Varieties of Immediacy

The Immediacy criterion requires that a theory of self-knowledge explain the thesis that self-knowledge is distinctive in that it is *immediate* or non-inferential. In this section, I briefly point to a number of ways the thesis has been articulated in the literature, and refer to three possible readings of the underlying intuition: explanatory, epistemic, and psychological immediacy. I argue that while one reading—*explanatory immediacy*—should be rejected,[63] *psychological immediacy* should remain a minimum criterion for a theory, and *epistemic immediacy* should be retained as an ideal desideratum.

The thesis I wish to capture, for the purposes of compiling a list of desiderata, is the underlying intuition behind the claim that self-knowledge is *immediate, groundless*, *baseless*, *non-evidential*, or *non-inferential*. For the purposes of the argument, I take these terms as attempts to describe common phenomena: in some core cases, it is (i) difficult for a subject to make sense of

[61] In private correspondence, Byrne (in 2016) has suggested that he perhaps should have worded this differently. Certainly, however, following the rule in the third-person case 'isn't a good idea'.
[62] One may also need to point, for instance, that medium-sized dry goods are predictable and stable, and form a good part of what preoccupies the average person, and that together with a largely reliable set of shared senses (and so on) this helps to ensure any reasonable level of success that could be achieved.
[63] Following Cassam (2009)

requests for reasons or evidence for a self-ascription, and she cannot always provide them; and (ii) the fact that a subject cannot always provide reasons or evidence does not count against her being knowledgeable about such states.

One might first try to understand the phemomena in terms of commonsense psychology (see e.g. Roessler's 2013 discussion), and the case, psychologically speaking, for immediacy is fairly strong. Take standard cases of our knowledge of our own abilities:[64]

> A character in a P. G. Woodhouse novel is asked whether she can speak Spanish and replies 'I don't know: I've never tried'. The point here is that in order to know that you can speak Spanish you don't need to have tried and failed to speak Spanish … You 'just know' that you can't speak Spanish (Casssam 2014, pp. 34–5)

Alternatively, take self-attribution of intentions. A natural response to inquiries into how one can tell that one intends to go to the cinema might be to insist that going to the cinema is *just what one is minded to do* (Roessler 2013, p. 42). In describing our day-to-day self-ascriptions of intentions, there does not appear to be a great deal more to say. The question for our current purposes is whether there is a great deal more to say about the significance of immediacy, epistemically speaking.

One way to explain these phenomena is to suggest that that knowledge in this domain, unlike knowledge in general, does not typically have, or require, the support of reasons or corroborating evidence. Crispin Wright's (2000) remarks[65] on *groundlessness* in the case of knowing that one is in pain are a helpful expression of this thought:

[64] Though not 'level of ability', for which evidence tends to show we are susceptible to all sorts of biases. See e.g. the literature on Depressive Realism (e.g. Alloy and Abramson 1988).
[65] See Cassam (2009)

> The demand that someone produce reasons or corroborating evidence for such claims is always inappropriate. There is nothing (in that sense) upon which the claims are based. (Wright 2000, p. 14)

Whereas it may be natural to think of knowledge of others' mental states as being based upon interpretation (i.e. inferences from behaviour), such an explanation is not available in the first-person case, or at least it cannot be the basic case (Wright 2000, p. 16) because there are limits to what can be explained that way:

> When Emma interprets her reaction to Harriet's declaration as evidence that she herself loves Knightley, there is an avowable ground—something like 'I am disconcerted by her love for that man and, more so, by the thought that it might be returned'—which is a datum for, rather than a product of, self-interpretation. (Wright 2000, p. 16)

The example is meant to show is that successful self-interpretation in some cases relies upon data that is drawn from 'non-inferential knowledge of a basic range of attitudes and intentionally characterized responses' (p. 16), and so self-interpretation cannot be all there is to self-knowledge, even if interpretation is all there is to knowledge of others' minds. If we accept that knowledge of others' minds is inferential, based on interpretation, there is an asymmetry between first-person and third-person knowledge of mental states that is a candidate for a distinctive epistemic mark of self-knowledge.

I take it that the argument does enough to invite a response from any Parity theorist, and more specifically those who defend the view that self-knowledge *is* self-interpretation. Available responses include a rejection of the conclusion that self-interpretation relies on data that is non-inferential, for example, by suggesting that an inference is present but undetected; [66] or accepting the thesis that self-knowledge is non-inferential, but rejecting the

[66] Inferentialism of this (interpretive) variety has support from e.g. Carruthers (2011) and Cassam (2014).

assumption that knowledge of other minds is (essentially) inferential.[67] I will briefly touch upon both in attempting to explore whether Immediacy can be characterized as describing an interesting epistemic feature of self-knowledge.

2.1 Evidential and explanatory immediacy

We have allowed that there are excellent grounds for the commonsense psychological view that self-knowledge is immediate. This in itself neither supports nor eliminates a further epistemic reading of immediacy, and in this section I will explore two possible versions of such a reading courtesy of Cassam (2009). Cassam (2009) suggests that there are two plausible (epistemic) notions of the idea that there is nothing upon which a subject's claims about some of her mental states are based (see §1.2 above): *explanatory* baselessness and *evidential* baselessness. Cassam argues that if self-knowledge is baseless at all, it cannot be baseless in the explanatory sense, and so it must be baseless in an *evidential* sense. But since 'evidential baselessness' is 'relatively commonplace' (in particular, it occurs outside of the domain of self-knowledge), it is not the notion of evidential baselessness that captures what is 'epistemologically distinctive' about self-knowledge. Being a relatively 'commonplace' phenomenon, evidential baselessness does not require any 'special explanation' (p. 3). It is worthwhile considering some of the main points in Cassam's argument.

2.1.1 Evidential Immediacy

To state that one's awareness of *x* is evidentially baseless is to state, roughly speaking, that one's awareness of *x* is not 'inferred from observational evidence'; 'evidentially baseless knowledge is knowledge that is not evidence-mediated … not inferred from observational evidence' (Cassam 2009, p. 6), or 'not inferred from anything epistemically more basic' (Moran 2001, p. 10;

[67] The prospect of non-inferential knowledge of the mental states of others is also not particularly popular, but has been argued for in some depth e.g. by Will McNeill (2012).

Ibid.). A few points of clarification: knowing by 'inference from observational evidence' is not the same as knowing by observation (pp. 8–9); and, arguably, a cognitive transition can be 'evidence mediated' without involving an inference (much depends on one's view of 'inference').[68]

The former case is illustrated by Austin's (1962) remarks that a pig coming into view does not 'provide me with evidence that it's a pig, I can now just *see* that it is, the question is settled' (p. 115).[69] Evidence of 'porcine presence' might include buckets of pig food:

> Yet pig food can be present without any pig being present. In contrast, the visible presence of a pig isn't an *indication* of its presence and does not leave open that there is no pig in the vicinity. (Cassam 2009, p. 8)

Thus, when one knows that a pig is present by seeing that there is a pig present, one's knowledge 'is evidentially baseless even if it is based on observation' (p. 8). And so, in perception, we have a plausible example of evidentially baseless knowledge that is not self-knowledge. Other candidates include testimonial knowledge (p. 7) and—often taken to be the standard contrast case to self-knowledge—knowledge of other minds, for example, our knowledge that someone is in pain (e.g. McDowell 1998, pp. 304–305) or angry (e.g. McNeill 2012):

> In such cases, it will be inappropriate to describe one's knowledge ass 'mediated' by awareness of outer manifestations. Moreover, if one can literally see that someone is in pain … then the resulting knowledge is evidentially baseless in the sense in which a lot of ordinary perceptual

[68] See Boghossian (2014), who suggests that philosophers have allowed themselves to use the term freely, without explaining what it means. Space prohibits in-depth discussion of the term in this chapter, although the issue does receive further attention in chapter four of this volume.
[69] Also in Cassam (2009)

> knowledge is evidentially baseless. That is … when one sees that someone else is in pain that settles the question.[70] (Cassam 2009, p. 14)

So we can perfectly well make sense of how self-knowledge could be immediate (*baseless*) in the evidential sense. However, the fact that this kind of knowledge can be found in a number of examples outside of the target domain counts against its ability to shed much light on 'what makes self-knowledge special' (p. 4). (At the end of the section I remark upon whether this is enough to see Immediacy removed from our list of desiderata.)

2.1.2 Explanatory Immediacy

An alternative notion of baselessness is *explanatory* (Cassam 2009). We saw, in Austin's example, that knowledge can be observational without being inferred from observational evidence. In such cases, knowledge is explained by one's *seeing* that there is a pig present. That is, there is something that one can say in response to inquiries about how it is that one knows there is a pig present: one sees, and thereby knows; or one knows by seeing (e.g. Cassam 2009, p. 8). In contrast, a case of explanatory baselessness would see the subject with 'nothing illuminating' that 'can be said in answer to the question "How do you know?"' (p. 6):

> On this account, to know that *P baselessly* would be to know that *P* without their being any substantive explanation of one's knowledge that *P* (Cassam 2009, p. 6)

We have seen this is not the case in Austin's pig example, and arguably it is not the case for testimonial knowledge: I may be presently unable to tell you that testimony was the source of my knowing that oxygen is released from boiling

[70] Cassam lists three conditions here: '(1) when one sees that someone else is in pain that settles the question, (2) seeing that another person is in pain entails that he is in pain, and (3) when one knows that another person is in pain by seeing that he is one does not infer or conclude that he is in pain' (2009, p. 14).

water, but my current inability to provide that information does not detract from it being a good explanation of how I came to know it (p. 8). The difficulty with the *explanatory* notion of baselessness is that it is hard to square with the intuition that if one knows that *p*, there must be a 'specific way' in which one knows that *p* (p. 12).[71]

A promising example in favour of the 'explanatory' notion of immediacy is pain, which we can accept for the sake of argument is *evidentially* baseless (p. 11):

> when I know that I am in pain, there is no answer to the question 'How do you know?' … this is one of those occasions on which the question 'How do you know?' is 'at least absurd, and perhaps unintelligible as a question' (Cassam 2009, p. 10)[72]

A quick response might suggest the worry is down to a confusion between conversational impropriety and a form of epistemic achievement: a natural candidate for the explanatory basis of one's knowledge that one is in pain is that one *feels*—or *feels that one is in* (p. 11)—and thereby knows, that one is in pain. But we can strengthen the case by considering a particular view of pain that would bar this response. On some accounts, feeling pain and being in pain are the same thing (see e.g. Shoemaker 1994, p. 128; here in Cassam 2009, p. 10) and since there is no 'ontological distance' between the two, *feeling* pain, cannot really be a way of knowing that one is in pain, since, 'one cannot know that *P* simply by its being the case that *P*' (Cassam 2009, p. 10).[73]

However, although '*S* feels pain' cannot be a way of explaining S's knowledge that she is in pain, the propositional attitude 'S can feel that she is in pain', which requires the concept *pain* is a more 'advanced cognitive achievement' and bypasses a 'sensible ontological distance' requirement for

[71] Whether one accepts Cassam's (2007a) explanatory view of knowledge, of course, will affect one's view of such a point. I have taken his view to be independently plausible (Ch. 1).
[72] Cassam quotes Hampshire (1979, pp. 282–3) in the passage.
[73] See the Acquaintance view of self-knowledge for an analogous concern (§1.1).

knowledge (p. 11).[74] Thus, regardless of whether *feeling* pain and being in pain are the same thing, knowing that one is in pain by feeling that one is in pain is a genuine explanation of how one knows that one is in pain, and so it does not 'come out' as explanatorily baseless (*Ibid*.).

2.1.3 Immediacy as a desideratum

Cassam (2009) draws two conclusions: (i) that no self-knowledge is explanatorily baseless; and (ii) that some self-knowledge is evidentially baseless, but since evidential baselessness is a commonplace phenomenon it requires no special explanation. Those conclusions suggest we should reject at least two of the three notions of immediacy we have considered: explanatory baselessness because it is not a *genuine* 'mode of epistemic access', and evidential baselessness because it is not a *unique* mode. Since the commonsense psychological notion of immediacy with which we started is also plausibly commonplace, we might consider dropping Immediacy from our list of desiderata altogether. This would be a mistake.

Firstly, Cassam's argument does not aim at eliminating all notions of baselessness, only two plausible epistemic notions. Perhaps epistemic immediacy in the self-knowledge case is different to immediacy in other cases, or self-knowledge might be both (a) immediate and (b) not another listed kind of knowledge (Cassam 2009). In the first case, we would need to stipulate the difference between immediacy in one case and immediacy in the others (i.e. highlight some epistemic feature or property that distinguishes the self-knowledge case), and in the second another risk becomes apparent:

> Perhaps then what makes self-knowledge special is that much of it is non-inferential, non-perceptual *and* not based on testimony. At this point, however, the claim that self-knowledge is epistemologically distinctive is in danger of reducing to the claim that self-knowledge is self-knowledge and not

[74] Cassam (2009) points to two other possible responses: one could reject that feeling pain and being in pain are one and the same thing; or one could reject the ontological requirement (p. 11).

> some other kind of knowledge. Everything is what it is and not something else. (Cassam 2009, p. 15)

Two things should be said about Cassam's (2009) dialectical position: (i) the notion of baselessness is being evaluated as a candidate for the sole defining or distinctive *epistemic* mark or feature of self-knowledge; and (ii) Cassam is responding to a very particular kind of question: 'How is it possible that our knowledge of our inner lives is baseless?' (p. 16). The difficulty with this approach is that once one realizes there are no obvious obstacles to the possibility of baseless self-knowledge there is little force to the 'How possible?' question. But we need not agree that baselessness is meant to be the sole explanandum of self-knowledge.

One reason for thinking that baselessness is meant to be the only epistemically distinctive feature of self-knowledge is that other candidate features have been discredited:

> more traditional accounts of what makes self-knowledge special have focused on its alleged infallibility or incorrigibility. Yet the suggestion that these are the epistemic privileges that make self-knowledge special faces some serious challenges, the main one being that much less of our self-knowledge in infallible or incorrigible than has traditionally been supposed (Cassam 2009, pp. 12f.)

However, infallibility and incorrigibility are not the only forms of epistemic privilege worth considering (see §3) and in §1, it became apparent that the Peculiarity thesis is more robust than some Parity theorists would have us believe.[75] So there is support for the thesis that Immediacy is one among a number of criteria rather than the only one. Thus we might amend Cassam's 'how possible' question to: 'How is it possible that our knowledge of our inner

[75] Note that, as a 'way of knowing', some notions of Immediacy would fulfil the Peculiarity criterion. Explanatory baselessness, if it were an epistemic phenomenon, looks like one.

lives is peculiar, immediate, and affords a first-person epistemic advantage?'. Or, more pertinent to the present task, 'How does a theory of self-knowledge explain these features?'

Now the issue is whether Cassam's concerns are enough for us to abandon Immediacy from our list of desiderata. A number of considerations suggest not: (a) this alternative version of the question has philosophical force that Cassam's version lacks, because in the absence of examples that demonstrate knowledge with these characteristics occurs outside of the target domain, we might think this *combination* of features *is* unique or special; (b) this way of structuring the question does not face the same reduction problem. Regardless of the number of criteria that enter the list, it could turn out that self-knowledge is not the only form of knowledge that meets them; (c) the concerns do not (nor are they intended) to bring into question the target phenomena that give rise to the commonsense psychological view that self-knowledge is immediate; and (d) at least some cases of epistemic immediacy are controversial (see e.g. Cassam 2009, p. 7) and/or rely on very specific notions of inference (e.g. McNeill 2012).[76] Because, 'we may fancy that we see and feel what in reality we infer' (Mill 1882/1990, p. 20), the matter of deciding whether a cognition is epistemically or psychologically immediate sometimes appears to be speculation.

The upshot is that we have not seen enough to remove Immediacy from our list of desiderata altogether. In particular, a theory of self-knowledge should at least explain (or explain away) the commonsense psychological view that self-knowledge is immediate based upon the initial phenomena, that is: (i) it is sometimes difficult for a subject to make sense of requests for reasons or evidence for a self-ascription, and she cannot always provide them; and (ii) sometimes, the fact that a subject cannot always provide reasons or evidence does not count against her being knowledgeable about such states.

[76] McNeill (2012) appears to draw a distinction between 'epistemic inference' and inference more generally.

> ***Psychological Immediacy***—a subject (S) can be knowledgeable about her current mental state (C) without being able to provide her reasons or evidence for self-ascribing mental state (C).

In the event that we find good reason to 'elevate it to the status of epistemology' (Velleman 1989; in Roessler 2013) we might also retain an epistemic version of the notion as an ideal desideratum:

> ***Epistemic Immediacy***—a subject (S) can be knowledgeable about her current mental state (C) without inferring that she is in (C) from reasons or evidence that she is in (C).

For the forthcoming chapters, when I refer to 'Immediacy' (I), I am referring to the commonsense version of the thesis. Whenever I am referring to the epistemic version of the thesis I will make that explicit.

3. Epistemic Security

The aim of this section is to see whether there is anything worth retaining in the claim that self-knowledge places the first-person in an epistemically privileged position when it comes to knowledge of her mental states. I argue for a conception of Epistemic Security offered by Byrne (2011a), [77] on which 'beliefs about one's mental states are more likely to amount to knowledge than one's corresponding beliefs about others' mental states' (Byrne 2011a, p. 202). In doing so, it will be helpful to highlight some common misconceptions of traditional views before moving onto more contemporary versions of the thesis, which propose a modest form of epistemic privilege.

[77] See also Byrne (2005), although there are minor differences between the two formulations.

3.1 Traditional notions of epistemic security

A common foil for arguments against a certain approach to self-knowledge is Cartesianism. Loosely described, the position suggests that our access to our minds is complete and (almost) unimpeachable, although this is not the only thing that commentators have found objectionable.[78] One way of attempting to capture the position is to say that it must be committed to both *infallibility* and *omniscience* about the mind:

> One is *infallible* about one's own mental states if, and only if, one cannot have a false belief to the effect that one is in a certain mental state. (In other words, one's belief that one is in a particular mental state entails that one is in that mental state.) One is *omniscient* about one's own states if, and only if, being in a mental state suffices for knowing that one is in that state. (In other words, one's being in a particular mental state entails that one knows that one is in that state.) (Gertler 2011a, pp. 61–2)

Some philosophers call this combination of views Cartesian 'transparency' (see e.g. Carruthers 2011). To avoid confusion with the Transparency approach that is the focus of forthcoming chapters, I will use the term 'Transpicuity': one's access to one's mind is Transpicuous if and only if one is both infallible and omniscient with regards to its contents.

Something like the Transpicuous access view has been attributed, among others, to Descartes, and Kant,[79] and it has been argued (Carruthers 2011, Ch. 2) that such Cartesian assumptions are not only prevalent in Western philosophy, but are universal among humans, sometimes as 'tacit assumptions' rather than 'explicit beliefs' (p. 31). Whatever wide endorsement (p. 33) it may have enjoyed within and without philosophy, the view is now almost universally rejected as a serious attempt to characterize our access to our minds (see Ch. 1).

[78] See e.g. Ryle (1949,Ch. 6) for a sustained critique of the position as related to self-knowledge.

[79] Cassam (2003), for instance, suggests that Kant's explanation of how transcendentally necessary conditions are meant to be known a priori (e.g. A13/B26) 'relies on the somewhat Cartesian-sounding premise that what is internal to us is also transparent to us' (p. 198).

In light of the fact that Transpicuity looks implausible in the face of supposed counterexamples, [80] including an increasing body of empirical evidence,[81] it is appropriate to question whether anyone ever genuinely held—or at least defended—such a view.[82] Certainly Descartes's preparations for the *Meditations* suggest that he was aware that self-knowledge does not always come easily (Gertler 2011b), and elsewhere he recognizes a number of difficult cases. In *The Passions of the Soul* Descartes suggests that those 'most strongly stirred by their passions aren't the ones who know them best' (*Passions* I, 28). In the same passage he writes that the potential for confusion when it comes to our passions is due to 'the soul's close alliance with the body'. And later (*Passions* II, 147) states that since 'commotions of the soul are often joined with passions that resemble them, they frequently occur with other passions, and they may even come from passions that are their opposites':

> A husband mourns his dead wife, though he would be sorry to see her brought to life. Perhaps his heart is oppressed by the sadness aroused in him by the funeral display and by the absence of a person to whose company he has become accustomed. And perhaps some remnants of love or of pity occur in this imagination and draw genuine tears from his eyes. And yet, despite all this he feels a secret joy in the innermost depths of his soul. (*The Passions of the Soul* II, 147)

Descartes, then, is unlikely to have taken ordinary thinkers to have Transpicuous access, at least with regards to the passions. A revealing passage in

[80] For example: 'Kate trusts a friend's insights into her own psychology, and so she believes the friend when he tells her that she wants to live in the country. But the friend is mistaken—Kate really wants an urban life, though she hasn't reflected on her desires enough to realize this. Hence, Kate has a false belief about her own desires' (Gertler 2011c). Note, the self-knowledge failure in this case also 'undercuts the claim of omniscience: in the case described, Kate is unaware of her real desire, which is to live in the city' (Gertler 2011c).

[81] Nisbett and Wilson (1977) and research on Choice Blindness (e.g. Hall, Johansson, and Sikström 2008) are sometimes taken to be clear examples of how poor our access to our minds can be. However, the examples can be misleading if one does not keep apart different varieties of self-knowledge failure (see Nisbett and Wilson 1977, p. 255; Schwitzgebel 2006; Ch. 1, and Appendix 1).

[82] Patrick Greenough (2012) e.g. suggests that it is not clear that anyone ever did.

the *Discourse on Method*[83] suggests that they do not enjoy Transpicuous access to their own beliefs either:

> I thought that in order to discover what opinions they really held I had to attend to what they did rather than what they said. For with our declining standards of behaviour, few people are willing to say everything they believe; and besides, many people do not know what they believe, since believing something and knowing that one believes it are different acts of thinking, and the one often occurs without the other. (*Discourse on Method*, AT VI 23)

The passage shows that we can at rule out 'omniscience' from Descartes' view since one's first-order belief (B*p*) can fail to issue in a corresponding higher-order belief (BB*p*). It may also imply that we can go wrong about our minds when we think we do not believe something (i.e. ruling out infallibility as expressed above), and although it leaves open the possibility that Descartes thought that whenever we have a higher order belief (BB*p*) we will have the corresponding first-order belief (B*p*) (see Stoneham 2004), this more restricted position suggests that many intended counterexamples, including the Freudian unconscious miss their mark.[84]

Immanuel Kant's assertion that 'It must be possible for the "I think" to accompany all my representations' (B131–2) is sometimes taken as one example that betrays Cartesian sympathies (see Carruthers 2011, p. 27). Although the interpretation of this phrase is controversial, Kant elsewhere suggests that *some* aspects, or operations, of the mind 'have not to be sought for without' and 'cannot remain hidden from us' (A13/B26). However, it is also clear that Kant is sensitive to the possibility that parts of our minds remain hidden:

> In fact it is absolutely impossible by means of experience to make out with complete certainty a single case in which the maxim of an action otherwise in

[83] Stoneham (2004, p. 259) and Wilson (2014. p. 669, n. 3) also point to the passage.
[84] See Stoneham (2004) for a detailed treatment of this point. He refers to the position as 'incorrigibility' (see pp. 559–661).

> conformity with duty rested simply on moral grounds and on the representation of one's duty. It is indeed sometimes the case that with the keenest self-examination we find nothing besides the moral ground of duty that could have been powerful enough to move us to this or that action and to so great a sacrifice; but from this it cannot be inferred with certainty that no covert impulse of self-love, under the mere pretense of that idea, was not actually the real determining cause of the will; for we like to flatter ourselves by falsely attributing to ourselves a nobler motive, whereas in fact we can never, even by the most strenuous self-examination, get entirely behind our covert incentives since, when moral worth is at issue, what counts is not actions, which one sees, but those inner principles of actions that one does not see. (Kant 1785/1997, §2, 4: 407, pp. 19–20)

The precise implications for Kant's views of self-knowledge are open to debate. On one view the passage suggests that one cannot tell 'which of our transparently-introspectable impulsions causes one to act on a given occasion' rather than 'a doubt about the accessibility of those impulses (Carruthers 2011, p. 27). However, the text is compatible with a reading upon which not all of one's incentives (or impulses)[85] are introspectively available. And in either case the possibility of error is evident. In the first case, the belief that one acted for ignoble reasons may be unavailable to, or obscured for one, precisely because one is inclined to think oneself noble. This puts the first person and third person in more-or-less the same position with regard to incentives, and this goes against the general Cartesian intuition that the first-person is at a distinct advantage. If some incentives are not introspectively available, then omniscience, on Kant's view, must be false.[86] And access to inefficacious incentives has no obvious epistemic benefits. So, Kant did not support an unrestricted version of Transpicuity either.

[85] Alternative translations could be part of the issue here. Carruthers (2011) focuses on a key sentence that reads 'get to the bottom of our secret impulsions', rather than 'get entirely behind our covert incentives'. Although it is difficult to see that the possibility that in either case some the impulsions or incentives themselves may be hidden from us has been eliminated.

[86] Much will rest on the view of incentives, here, but space does not permit a more detailed discussion.

Whatever its origins, unrestricted Transpicuity is a mistaken view, but more restricted versions enjoy some contemporary support. Chisholm (1981), for instance, suggests that for a restricted class of states, anyone who is in such a state knows that she is in it such a state. Stoneham (1998) defends a form of *incorrigibility* via a 'Containment Claim'—'If someone believes that he believes that *p*, then he believes that *p*' (p. 128).[87] And support for the security of a limited class of mental state—cogito-like, and self-verifying judgements more generally—is still in currency (Burge 1996; Brown 2000; Byrne 2011a). Since some of these remnants will feature in the following chapters, it is worth expanding briefly.

3.2 Cogito-like judgements

Cogito-like judgements, such as those employed in Descartes's anti-sceptical arguments, are a paradigm case of a class of judgements that are both epistemically special and environmentally neutral (Burge 1996, p. 91). Take 'I am thinking that there are physical entities', for example (p. 92). In order to be true, it is only required that 'I am engaged in some thought whose content is that there are physical entities' (*Ibid.*). Given the way that such judgements have been used in anti-sceptical arguments, it may be tempting to class among their qualities that they are beyond doubt. However, this would be wrong:

> The scope for human perversity is very wide. One could be so far gone as to think to oneself: 'I do not know whether I am now thinking or not; maybe I am dead or unconscious; my mantra may have finally made me blissfully free of thought'. Such mistaken doubt would evince cognitive pathology, but I think it possible. (Burge 1996, p. 92)

[87] Stoneham (1998) takes the belief claim (BBp → Bp) to be an instance of a broader claim ''BAp → Ap': if someone has a belief that he now holds some attitude to p, then he does. 'BDp → Dp' (for desire) and 'BIp → Ip' (for indifference) are also instances' (p. 128).

It would also be largely irrelevant to their contemporary use in discourse about self-knowledge. A property that is of relevance is that the judgements are contextual self-verifying: 'once one makes the judgement, or indeed just engages in the thought, one makes it true … One cannot err if one does not think it, and if one does think it one cannot err. In this sense, such thinkings are infallible' (p. 92). On a certain view (Burge 1996; Brown 2000) this property plays an important role in underpinning a variety of epistemic capacities such as *critical reasoning*. If we are critical reasoners, and critical reasoning requires reliable judgements about our own beliefs, desires, and intentions, then we must be competent self-knowers in these respects.[88] This appears to block a pernicious form of scepticism that suggests access to our minds is subject to 'massive and pervasive' error (e.g. Schwitzgebel 2008a).

Two concerns arise about how such an approach can help to answer questions central to this chapter: (i) is how far such an approach can get us with regards to a theory of self-knowledge, and specifically its epistemic desideratum; and (ii) is how any abundance of true beliefs guaranteed by such a means would be explanatory of knowledge.

With regards to (i), this kind of argument provides one good reason for thinking that judgements about one's own mental states must be in some sense secure or reliable, without stipulating the precise method by which we come to know them. This preserves the independence of the Peculiarity and Epistemic Security theses. However, it appears to be silent on the degree of security afforded, and so our formulation of Epistemic Security criterion will need to be flexible on that matter. Chapter three examines (ii) in greater detail, but concerns in the literature include that self-verifying judgements are *cheap* and *beside the point* (Schwitzgebel 2008a), a mere 'philosophical curiosity' (Burge

[88] This kind of strategy is in currency at the time of writing. Maja Spener employs a like-minded defence of our introspective abilities in 'Introspection and Abilities' delivered at the *First-Personal Data* conference, University of Bergen, 28–29 August, 2014. (It is vulnerable to a similar set of concerns.) For an interesting current use of self-verifying judgements in this literature, see Byrne (e.g. 2011a).

1996),[89] and there are questions over whether they are knowledge-conducive (Byrne 2011a).

3.3 Fallibility, ontological distance, and parity

There are difficulties in characterizing Epistemic Security at the high end of the spectrum. Not only do positions such as Transpicuity look implausible, but as judgements about one's mind become extremely secure, they begin to look less like cases of knowledge on our accepted view (see e.g. Acquaintance theory §1.1; and pain in §2.1.2). A theory of self-knowledge needs to leave room for explanation. This requires, plausibly, a degree of 'ontological distance' (Cassam 2009), or room for cognitive achievement (e.g. Fernández 2013, pp. 33, 103), (*Ibid.*). This places an *upper limit* on degree of privilege that a theory can provide. Before saying more about characterisations of Epistemic Security that stay within that upper limit, I will briefly consider whether there is a good case for thinking there is *lower limit*.

Burge-style (1996) arguments suggest that we must have reliable access to our mental states, given that we do have certain epistemic capacities. We have seen that this access is likely to 'sub-Transpicuous', and needs to leave room for explanation (i.e. ontological distance, or cognitive achievement). But the Parity theorist may claim that all three conditions can be met without additional first-person privilege. The epistemic claim of the Parity Thesis is that there is no first-person epistemic advantage:

> Knowledge of what there is to be known about other people is restored to approximate parity with Self-Knowledge … residual difference in the supplies of requisite data makes some differences in degree between what I can know about myself and what I can know about you, but these differences are not all in favour of Self-Knowledge (Ryle 1949, pp. 137–8)

[89] Burge is reporting, here, rather than endorsing.

Echoing Ryle, Carruthers (2011) allows that there may be 'more evidence available for interpretation in the first person than in the third' but suggests this 'doesn't always entail an increase in reliability' because 'sometimes the presence of more data doesn't lead to more reliable conclusions' (p. 24).[90] Nevertheless, humans are exceptional mind-readers, and reliable first-person access to mental states can be explained by us 'turning our mindreading capacities on ourselves' (p. 5). Thus we have reliable access to our mental states that is fallible, and allows for explanation (the latter two because the process is interpretive).

There are a number of concerns with this view. It seems patently false, for instance, to suggest that to know which dessert I most desire from the menu, that I would first need to observe which one I pick; or to know that I am in pain, I would need to observe and interpret the relevant pain behaviour. So the theory does not look promising for all states. Secondly (Wright 2000), it has been argued that self-interpretive work relies upon more direct, non-interpretive access to a basic range of attitudes; or in Carruthers's (2011) words: 'the mindreading system needs to have access to the agent's own beliefs in order to do its interpretive work' (pp. 236–7).[91] And thirdly, one might urge that no rational being could be (or at least *is*) 'self-blind' (Shoemaker 1994)—that is, having third-person-only access to her mind. This latter concern goes as follows: if the objects of self-knowledge are thought of as being mere objects of observation, they will be thought of as independent of the subject (Speaks 2004).[92] But if they are conceived of as independent of the subject, then the ('self-blind') subject would appear alien: she would (a) fall into errors such as asserting Moore's Paradoxical statements (e.g. 'it is raining, but I don't believe that it is'); (b) would be unable to share her beliefs with others, and would thus be unable to engage in co-operative endeavours'; (c) would be unable to engage in higher-order deliberation on lower-order states, and would thus be devoid of

[90] See Byrne's (2012) review of Carruthers (2011) for a helpful discussion of the similarities between the two positions.
[91] Carruthers is reporting on, rather than endorsing the objection.
[92] See: http://www3.nd.edu/~jspeaks/courses/mcgill/519-self-knowledge/shoemaker-self-knowledge.pdf.

agency as we ordinarily see it; and (d) would regard herself as a 'stranger', for instance, in 'observing [her] own pain-avoidance behavior without grasping her own pain' (Gertler 2011c).[93] The upshot is that while self-blindness may be a possibility, it is not an 'actual condition'; 'there are no individuals who have only third person access to their mental lives, with spared rational and other epistemic capacities' (Byrne 2011, p. 213). Third-person-only access to our minds is, therefore, implausible (at least implausible enough to retain Epistemic Security among our desiderata).

To summarise, whatever formulation of (E) we rest upon should allow for the following: (i) fallibility; (ii) explanation; and (iii) reliable access to mental states; that (iv) affords some—however modest—advantage in first-person cases. In the rest of this section, I briefly compare two 'modest' versions of the thesis, by which the success of a theory of self-knowledge might be measured.

3.4 Modest approaches to Epistemic Security

However we decide to articulate (iv) in the criterion that reflects (E) it will implicitly or explicitly refer to knowledge or knowledge-conduciveness. Ultimately, the first person must end up in a favourable position with regards to knowledge of her mind. But because we are aiming at a formulation that captures a diverse range of candidate theories, the *mode* of that privilege should remain open. In this section I contrast two approaches to explaining (E) that allow for first-person privilege to be relatively modest and argue that only is a suitable candidate for the purposes of this inquiry.[94]

Jordi Fernández (2013) takes the desiderata for self-knowledge to be two features: 'Special Access' and 'Strong Access' (pp. 4–6.). Special Access relates to the source of justification for a subject's beliefs about her mental states. Suppose

[93] Shoemaker's conclusions have been the source of much discussion (see e.g. Gertler 2015). Space prohibits detailed treatment here.

[94] While the formulations allow for modest privilege, they sometimes deliver a substantive advantage. It has been commented that Alex Byrne's (2005, 2011a) account, for instance, provides something close to infallibility (Carruthers 2011).

Jim believes that Sarah wants milk in her coffee. For Jim's belief to be justified, it must be based on behavioural evidence and reasoning (pp. 4f.); perhaps Jim notices that Sarah chooses milk whenever the option is available, etc. Jim's evidence is gathered by observing Sarah's behaviour. By contrast, Sarah seems to know at least some of her mental states without the need to observe her behavior (it would seem odd if this were not the case at least sometimes) Fernández (2013) takes this apparent asymmetry to be a difference in the 'source of justification'—Jim observes and reasons or infers, and Sarah does not—that can be captured in the following principle (p. 5):

> For any S and S*, propositional attitude A and proposition P: Normally, if both S and S* are justified in believing that S has A towards P, then
>
> (1) S*'s justification for believing that S has A towards P relies on reasoning and behavioral evidence.
>
> (2) S's justification for believing that she has A towards P relies on neither reasoning nor behavioural evidence.

Fernández, in (2), takes the apparent lack of observation and reasoning at face value. (While we may resort to self-interpretation in some cases, we do not 'normally' need to.)

Strong Access concerns the strength of justification. Normally, when Sarah claims to want milk in her coffee, we defer to her 'opinion on the matter': 'By default, we seem to think that each of us knows best what is in our own minds, which suggests that we take it that our beliefs about our mental states are more strongly justified that anybody else's beliefs about them' (p. 6). This asymmetry can be captured by the 'Strong Access' principle (*Ibid.*):

> For any subjects S and S*, propositional attitude A and proposition P: Normally, if both S* and S are justified in believing that S has A towards P,

> then S is more strongly justified in believing that she has A towards P than S* is.

Fernández (2013) takes our deferral not only as a description of social and linguistic practices, but to reflect some genuine epistemic asymmetry.

Several aspects of these principles are worthy of note. Firstly, Special Access assumes without argument something that we have not yet established: that *epistemic* immediacy is a genuine feature of self-knowledge (§2). This is relevant to our inquiry because, in doing so, it might eliminate a number of theories worthy of consideration (e.g. it has been argued that the Transparency approach is a poor fit for an epistemic notion of Immediacy, see Cassam 2014; Ch. 3). Secondly, Strong Access assumes without argument that the advantage must be explained in terms of justification. Not only will this prematurely restrict our list of candidate theories but, for our purposes, would do so by begging the question as to how an individual ends up in a favourable position by fixing the mode of privilege.[95]

By contrast, Alex Byrne (2011a) offers a formulation of the Epistemic Security thesis that does not, ostensibly, eliminate any specific mode of privilege:

> one has privileged access to one's mental states if 'beliefs about one's mental states are more likely to amount to knowledge than one's corresponding beliefs about others' mental states'. (Byrne 2011a, p. 202)

There is, however, one clear concern: the formulation frames the epistemic advantage in Reliabilist terms. However, 'more likely', here, need not create the same kind of problem that justification-focused criteria do, and so this alone does not merit a reformulation. If we accept that 'more likely' can be read as

[95] This is not to suggest Fernández (2013) begs the question in for his own project, in which he explicitly states that he wishes to see how far we can get given a specific (i.e. Internalist) notion of epistemic justification (pp. 41–5)

meaning that in the first person case one is 'in a more favourable position with regards to knowledge acquisition', then Byrne's formulation meets our requirements in that it: (i) can be fulfilled without infallibility, (ii) allows room for the view of knowledge adopted here, (iii) allows access to be reliable, (iv) builds in a first-person advantage. For the chapters that follow, by Epistemic Security criterion I will, following Byrne (2011a), mean:

> ***Epistemic Security***—'beliefs about one's mental states are more likely to amount to knowledge than one's corresponding beliefs about others' mental states'. (Byrne 2011a, p. 202)

4. Uniformity, Economy, and Transparency

With the minimal criteria in place, I now turn to concerns that have been brought into specific focus by developments in the literature that discuss Transparency accounts of self-knowledge (e.g. in Moran 2001; Byrne 2005; Boyle 2011). The first is whether a theory ought to explain access to all mental states, occurrences, and processes in the same way (*Uniformity*); the second, whether a theory ought to restrict itself to epistemic capacities deployed in knowledge of other kinds or needs additional resources (*Economy*); and the third, whether a theory is compatible with the transparency of first-person thinking (*Transparency*). In this section, I remark briefly on why we might wish to retain these as ideal desiderata—that is, while Uniformity, Economy, and Transparency will not be decisive in the success of a theory, they are ceteris paribus factors that count towards a theory's success.

4.1 Uniformity

Theories of self-knowledge tend to aim at explaining first-person access to all mental states. A number that do not explicitly attempt this, proceed by

explaining a core case (e.g. belief), and suggest that the explanation for the target state is a good candidate to be rolled out to others. Gareth Evans's (1982) remarks, for example, were directed towards 'ways we have of knowing what we believe and what we experience', but he expressed the aspiration that the approach would provide a 'good model of self-knowledge' in general' (p. 255). The thought, often implicit, appears to be condition upon the success of an account of self-knowledge is that it can explain all cases of first-person authority in the same way (Boyle 2009, p. 141). Moran's (2001) account appears to have fallen foul of the condition because it focuses on *deliberative* cases; ill-suited to self-knowledge of sensation, perception (see Byrne 2011a), and any state that does not 'seem to be subject to our active control' (Boyle 2009, p. 135). A surfeit of criticism suggests the assumption has been accepted, often 'uncritically', by many writing in the field (p. 135), and Boyle (2009) suggests that we reject the assumption. However, it is actively embraced by Byrne (e.g. 2011a), who explicitly aims to explain first-person authority for 'judgement-sensitive' attitudes (Boyle 2009), sensation, perception, inner speech, and mental imagery with the same basic approach. Byrne's (2011a) defends the assumption as follows:

> If the epistemology of mental states is not broadly uniform, then dissociations are to be expected. One might find, for instance, someone who knows what she believes like the rest of us, but whose independent mechanism for discovering her desires is disabled, leaving her with only a 'third-person' way of knowing what she wants. Such dissociations do not seem to occur however. (Byrne 2011a, p. 213)

But these examples do not go against Boyle's suggestion that a division ought to be made along Kantian lines:

> we must distinguish between an active aspect of self-knowledge that is knowledge of ourselves as spontaneous beings, and a passive aspect grounded in our power of sensible receptivity. (Boyle 2011, p. 2)

Both beliefs and desires fall under self-knowledge of the 'active' or 'spontaneous' variety, so the problem of dissociation is not a live one for Boyle's account. However, we may briefly consider whether clinical evidence might point to the kind of dissociations that Byrne has in mind.

A promising example is somatoparaphrenia: a delusion in which a part of one's body belongs to someone else (see Fotopoulou et al. 2011). In an intriguing piece of research, Fotopoulou et al. (2011) found that alternating first-person and third-person perspective—that is, 'direct view' and 'mirror view'—could result in rapidly alternating judgements of ownership for the same limb. The study suggests that 'limb disownership' can be 'altered using self-observation in a mirror, and in turn suggests dissociation between first- and third-person perspectives on the body' (p. 3946). Moreover, because reinstated judgements of ownership were not permanent, it suggests that 'the subjective sense of body ownership remained dominated by an impaired first-person representation of the body that could not be updated' (*Ibid.*).

On its face, this is precisely the kind of dissociation that Byrne predicts for cases where the epistemology of mental states is not broadly uniform. However, the effect is far too local to demonstrate dissociation of that kind. The impaired first-person representation usually affects one limb as opposed to the whole body and so cannot be taken to suggest that an independent mechanism for this variety of self-knowledge has been disabled,[96] and we should consider whether other cognitive impairments might better explain the condition.[97]

[96] An exploration of what exactly it does tell us is worthy of more space than can be afforded here.
[97] Byrne footnotes a similar point about supposed examples of self-blindness in schizophrenia patients in Nichols and Stich (see 2011b, fn. 4)

Other cases are worthy of attention,[98] but for the moment, we can accept that, ceteris paribus, Uniformity Assumption provides an important guideline for theory construction in this domain.

> ***Uniformity***—a satisfactory account of self-knowledge should be fundamentally uniform, explaining all cases of "first-person authority" … in the same basic way' (see Boyle 2009, p. 141).

The ceteris paribus clause allows for principled divisions along Boyle's lines to be treated on a case-by-case basis (i.e. rather than built into the formulation), and in light of their performance with regard to the other desiderata.

4.2 Economy

Coherent arguments against ontological parsimony (or *for* metaphysical extravagance) are difficult to come across.[99] When it comes to the domain in question, metaphysical 'extravagance' (see e.g. Byrne 2011a) sees us employing a dedicated faculty or capacities solely to explain self-knowledge, as opposed to faculties or capacities employed in other forms of knowledge. This kind of extravagance can be found, for instance, in contemporary, materialist descendants of Inner Sense views, such as Nichols and Stich's (2003) 'monitoring mechanism' (see also Armstrong 1981).

Economical theories suggest that self-knowledge and knowledge of other minds use the same cognitive apparatus (e.g. Ryle 1949; Carruthers 2011), that what is special about self-knowledge can be explained by 'normal intelligence, rationality, and conceptual capacity' (Shoemaker 1994; in Byrne 2011b), or that we are dealing with 'merely a special deployment of powers possessed by anyone who can draw inferences about any topic whatsoever' (Boyle 2009, p.

[98] Congenital analgesia is worth considering as a case a counter-example to Byrne's argument that goes in favour of Boyle's distinction.

[99] A few dotted examples that tug the intuition are can be found in the history of science. Francis Crick apparently thought it was ill-suited to biology, 'where things get very messy'. For this and other examples, see Ball (2016).

1).[100] While extravagance is not a bar to success on its own, other things being equal, an Economical theory of self-knowledge will be favoured over an extravagant theory.

> ***Economy***—a theory that explains the distinctive features of self-knowledge without recourse to capacities not employed in other domains of Knowledge (see also Byrne e.g. 2011a).

4.3 Transparency

Remarks about the Transparency of first-person thought and experience predate a recent Transparency 'turn' in the self-knowledge literature considerably. The transparency of experience as related to introspection can be traced back at least to G. E. Moore's observation that experience is diaphanous: 'when we try to introspect the sensation of blue, all we can see is the blue' (1922). The thought is reflected in Ryle's (1949) observation that some cognitive activities are found to be 'oddly elusive' (p. 134). Along with my inferring, deducing, concluding, and hearing: 'my seeing of the hawk seems to be a queerly transparent sort of process' (*Ibid.*).[101]

A number philosophers have adopted the view that this transparency as applied to first-person thought means that one can answer a question about one's own mental state merely by attending to a corresponding question about the 'topic' of that state (see Moran 2012, p. 212), which is typically world-directed rather than self-directed. Gareth Evans (1982) remarks, for instance, suggest that 'in making a self-ascription of belief, one's eyes are … directed outward' (p. 225). The general idea is that I can come to know my mind by attending to 'the world at large' rather than something 'inner' or 'psychological'.

[100] Boyle, here, is commenting on the ambitious nature of Alex Byrne's approach.
[101] Ryle thinks the problem arises because 'these verbs are of the wrong type to complete the phrase 'catch myself … ' (1949, p. 134)

Versions of the thought appear in the work of Edgley (1969), Evans (1982), Moran (2001), Shah and Velleman (2005), Boyle (e.g. 2009, 2011), Byrne (e.g. 2005, 2011a) and Fernández (2013), although accounts diverge dramatically (cf. Moran 2012, fn. 2). A common point of divergence is how one gets from the end of a world-directed inquiry that issues in a result about that world, to a conclusion about the self. Evans (1982) and Moran (2001) offer few clues as to the nature of that transition;[102] Shah and Velleman (2005) adopt a variety of Expressivism (see 2005, fn. 29); Byrne (e.g. 2011a) suggests that there is a world-to-mind inference; Boyle proposes a 'Reflective' approach (2009); and Fernández (2013) offers a version of the 'simple theory' of introspection[103] that he calls the 'Bypass' view. In some cases, the transparency of first-person thought plays a central role explaining how we know our mental states (Evans 1982; Moran 2001); and in others the nature of the cognitive transition to self-ascription provides the explanatory work: Byrne's world-to-mind inference is intended to explain what is special about self-knowledge; Boyle's reflection reveals only what we, tacitly, already know; and for Fernández transparency is one of numerous desiderata for a theory—that it 'should explain why mental states are transparent when we attribute them' (Fernández 2013, p. 38).

The task of constructing a desideratum in this case is to find the elements of the Transparency thesis that are common to competing theories. Matthew Boyle's attempt to do this is as follows:

> 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. Indeed, it seems that, for various sorts of mental states, there is in the normal case no other way to attend to them: all there is for me to contemplate in my sensation of blue is

[102] Evans appears to think that one is 'automatically' in a position to self-ascribe; Moran (2001) suggests that we do so immediately.
[103] See Declan Smithies (forthcoming)

> the (apparent) blueness of some worldly thing, and all there is for me to attend to in my belief that P is the (apparent) fact that P. (Boyle 2011, p. 3).

A formulation that is common to competing theories can make use of Boyle's analysis:

> ***Transparency***—a theory of self-knowledge should allow that there is nothing more to attending to some of one's mental states than to attending to their concomitant features or facts.

Given that it has become orthodox to insist upon some form of transition from the resulting from the activity, this will require further discussion (see Chs. 3 and 5). However, unless otherwise stated, whenever I refer to Transparency as a desideratum, I refer to the formulation above.

5. Additional desiderata

A theory of self-knowledge should leave in place established cognitive phenomena unless it provides independent, principled reasons to reject them. In other words, a theory should not be excessively revisionary of our cognitive capacities. In this section I deal with four concerns: (§5.1) a theory should allow for knowledge of an absence of belief (*Agnotic Access*); (§5.2) the target of a self-knowledge procedure ought not be altered by that procedure (*Preserved Access*); which is a requirement for, but independent of another kind of ability, that is (§5.3) to assess or reflect upon one's current (pre-existing) attitudes (*Evaluative Access*); and finally, a theory of self-knowledge should—without excellent reasons to the contrary—preserve our status as rational creatures (§5.4 *Self-Blindness*). I will briefly touch upon some reasons that we may wish to list these among our desiderata.

5.1 Agnotic Access

It is a common assumption that in addition to knowing one believes that *p*, one can know that we do not believe *p* (or indeed have no attitude towards *p* whatsoever). A similar but more exacting expectation can be found in Plato's *Apology*:

> So I withdrew and thought to myself: "I am wiser than this man; it is likely that neither of us knows anything worthwhile, but he thinks he knows something when he does not, whereas when I do not know neither do I think I know" (*Apology* 21d)

It is a reasonable expectation of a theory that I can be knowledgeable that *I do not believe that p* (or *do not know that p*) in the broadly the same way that I am knowledgeable that *I believe that q* (or *I know that q*) (i.e. I know it by the same method that I employ for self-knowledge in other cases). This becomes pressing if one accepts the Uniformity Assumption (see §4.1). It is also a reasonable expectation that if a theory of self-knowledge bestows some first-person epistemic security with regards doxastic self-knowledge, it should afford that same security with knowing that one does not believe (cf. Fernández 2013, p. 71),[104] especially if it has been stipulated, or is implicit, that one can be knowledgeable about both cases in the same way. So, if we take the assumption seriously, it places constraints both on Peculiarity and Epistemic Security:

> ***Agnotic Access***—A theory of self-knowledge should adequately explain how one could know that one does not have attitude (A) towards (p)

Since not all theories of self-knowledge can comfortably explain Agnotic Access—it has been suggested that some Transparency accounts might

[104] Fernández calls this principle AB (Absense of belief) (*Ibid*.).

struggle[105] in this respect, for example—it will be a useful test of a theory's success.

5.2 Preserved Access

In §3, I suggested that we adopt Byrne's (e.g. 2011a) conception of Epistemic Security. It has been argued (see Gertler 2011a) that there must be at least one important additional constraint on this feature of self-knowledge, namely that:

> If I have no belief that *p* (at *t1*) but consider whether I have a belief that *p* (at *t2*) I will not self-attribute a belief that *p* without creating a new belief. (Gertler 2011b, p. 5)

The intuition that Gertler's comment attempts to capture is that for a self-knowledge procedure to be reliable, it must successfully identify states in place prior to onset of the initiation of the procedure, or at least not misidentify states formed during (or as a result of) the procedure as states in place prior to the initiation of the procedure. Preserving the first part of this intuition would prove exacting on a theory of self-knowledge because the creation of a new belief itself need not constitute a failure in self-knowledge. But there is a clear case in which the latter part of this intuition can be preserved even if a self-knowledge procedure risks or results in the creation of a new belief—that is, just as long as one does not take that belief at t_2 to be in place at t_1 (i.e. one does not take the new belief to be a belief in place prior to the initiation of the procedure). The adjustment required is a minor one and, for brevity, the point could be left as Gertler puts it, but a more complete formulation of the constraint suggests as an important desideratum of a theory of self-knowledge, *Preserved Access*:

[105] This looks like a consequence of Gerter's (2011a) concern (see Ch. 3).

> ***Preserved Access***—If (at t_1) I do not believe that p, and I consider whether I believe that p, I will not (at t_2) self-attribute a newly formed belief that p as the belief I held (at t_1) or form a new belief that prevents access to a belief I held (at t_1).

This desideratum places an additional constraint upon the Epistemic Security criterion (E), by fixing the target of procedure as the belief in place prior to the initiation of whatever procedure is in use. If a procedure fails to meet this desideratum it will have failed also to meet (E).

5.3 Evaluative Access

Preserved Access is, I want to suggest, an enabling condition for something close to a commonsense view of critical reasoning, where the latter is a subject's ability to assess or reflect upon her current attitudes. This separate capacity may follow from the mechanism or procedure that a theory of self-knowledge deploys, though it need not. Going in favour of its addition to the list of desiderata is that it is either assumed or explicitly argued for in a good deal of the literature (e.g. Burge 1996; Brown 2000; Crane 2014). However, this is not always the case, and some philosophers (explicitly or implicitly) doubt either that we ordinarily have the capacity as commonly conceived (Nisbett and Wilson 1977; Dennett 1969), or that it is valuable in the way we take it to be (Kornblith 2012).[106] Here is what I mean by Evaluative Access:

> ***Evaluative Access***—a subject has evaluative access if that access allows her to assess or reflect upon her current (pre-existing) attitudes in light of her available evidence and the norms she accepts.

An expectation that *Evaluative Access* should be explained by a theory of self-

[106] Kornblith (2012) talks of 'Reflection' rather than 'Critical Reasoning' but a good deal of the activity that I am trying to capture here is covered in his discussion.

knowledge is too strict a condition. However, if a theory of self-knowledge assumes that Evaluative Access is possible, that theory ought to have explained how it is, or at least not make it impossible by its own lights.

5.4 Self-Blindness

At the end of §3, we saw that some views of self-knowledge allow for the possibility of self-blindness. We could expect the self-blind subject to: (a) fall into errors such as asserting Moore's Paradoxical statements (e.g. 'it is raining, but I don't believe that it is'); (b) be unable to share her beliefs with others, and would thus be unable to engage in co-operative endeavours'; (c) be unable to engage in higher-order deliberation on lower-order states, and would thus be devoid of agency as we ordinarily see it; and (d) regard herself as a 'stranger', for instance, in 'observing [her] own pain-avoidance behavior without grasping her own pain' (Gertler 2011c).

Since there are no real-life cases of individuals who suffer from self-blindness, 'with spared rational and other capacities' (Byrne 2011a), a theory will be in a mess if it produces individuals who would suffer these symptoms. And so self-blindness produces a plausible negative constraint on a theory not to produce subjects that could be expected to suffer the symptoms listed above.

Conclusion

The forgoing discussion has provided us with a list of desiderata that can plausibly be applied to any theory of self-knowledge. The full list of desiderata is listed and numbered as follows:

Minimal criteria:

(I) ***Peculiarity***—a method or procedure by pointing to which it is possible, satisfactorily, to explain how S comes to knows S's mental

states, and that cannot be used satisfactorily to explain how one S comes to know the mental states of others.

(II) ***Immediacy***—sometimes, a subject (S) can be knowledgeable about her current mental state (C) without being able to provide her reasons or evidence for self-ascribing mental state (C).

(III) ***Epistemic Security***—one has privileged access to one's mental states if 'beliefs about one's mental states are more likely to amount to knowledge than one's corresponding beliefs about others' mental states'. (Byrne 2011a, p. 202)

Ideal desiderata:

(IV) ***Epistemic Immediacy***—sometimes, a subject (S) can be knowledgeable about her current mental state (C) without inferring that she is in (C) from reasons or evidence that she is in (C).

(V) ***Uniformity***—a satisfactory account of self-knowledge should be fundamentally uniform, explaining all cases of "first-person authority" … in the same basic way' (see Boyle 2009, p. 141).

(VI) ***Economy***—a theory that explains the distinctive features of self-knowledge without recourse to capacities not employed in other domains of Knowledge (see also Byrne e.g. 2011a).

(VII) ***Transparency***—a theory of self-knowledge should allow that there is nothing more to attending to some of one's mental states than to attending to their concomitant features or facts.

Additional desiderata:

(VIII) ***Agnotic Access***—A theory of self-knowledge should adequately explain how one can know that one does not have attitude (A) towards (p)

(IX) ***Preserved Access***—If (at *t1*) I do not believe that *p*, and I consider whether I believe that *p*, I will not (at *t2*) self-attribute a newly formed belief that *p* as the belief I held (at *t1*)

(X) ***Evaluative Access***—a subject has evaluative access if that access allows her to assess or reflect upon her current (pre-existing) attitudes in light of her available evidence and the norms she accepts.

(XI) ***Self-Blindness***—a theory of self-knowledge should not leave subjects with third-person only access to their mental states.

In the next chapter, I will apply some of these measures to an approach to self-knowledge that has been the focus of much recent discussion in the literature.

3

Transparency, Deliberation, and Memory

Introduction

In chapter one, I pointed to a number of ways in which our thinking about introspection and memory converge. I concluded that to gauge the extent of that convergence one might see whether memory can explain some of what is taken to be special about self-knowledge. In the last chapter I outlined some features of first-person thinking that tend to shape theories of self-knowledge, and suggested formulations of these against which one might measure the success of a theory. In this chapter I look at an approach to self-knowledge that has gained prominence in the literature—the Transparency approach—and argue that, on one version of the approach, the epistemology of memory will play an important role in explaining how the account meets a number of desiderata.

The Transparency approach suggests that there is something in the way we go about responding to questions about mental states that can explain what is interesting or special about knowledge in the domain. But while a number of authors endorse a broad version of the Transparency thesis (e.g. Moran 2001; Byrne 2005; Boyle 2011a), or recognise its relevance to the epistemology of self-knowledge (Fernández 2013), notions of transparency are diverse, and the role those notions play in theories of self-knowledge varies.

The chapter proceeds as follows. In §1, I outline the Transparency approach and highlight its core features. In §2, I discuss a version of the approach that emphasises cases of 'making up one's mind', and point to a number of objections to that approach in line with desiderata from chapter two. In §3, I discuss an alternative view that aims to explain first-person privilege for

all mental states. I argue that the account constitutes progress for the approach with regards to several main desiderata. It also highlights an important contribution for our current inquiry: factual memory is likely to come into play if a Transparent and Economic account is to explain standard cases of doxastic self-knowledge. However, the view still faces a style of objection common to Transparency accounts. In §4, I highlight a version of that objection, and outline how the use of factual recall presents an initially promising response. In §5, I discuss the attempt to make the inferential Transparency account *uniform*. I suggest that this attempt makes the identification of the epistemology of memory being assumed a pressing matter for the account. In §6, I outline a view of the account's success, including one plausible view of factual memory, and contrast that view with the account's requirements for self-knowledge of intention. In §7, I explore three candidate views of memory that might explain how the account can meet the epistemic criteria outlined in chapter two, and reject all three as plausible views of memory given the account's assumptions. I conclude that the account has the potential to meet the main epistemic desiderata for doxastic self-knowledge, if it makes use of an appropriate view of factual memory. Insofar as the account is likely to succeed in its attempts to explain bouletic self-knowledge, that success is also likely to be due to filling in the detail of the assumed view of memory.

1. Transparent self-knowledge

Experience has a diaphanous or transparent quality. In trying to introspect about one's experience of 'blue' when observing a blue object (see Moore 1922), for instance, one comes up with 'nothing but the blue' (Byrne 2011a). Ryle (1949) echoes the thought:

> If I descry a hawk, I find the hawk but I do not find my seeing of the hawk. My

> seeing of the hawk seems to be a queerly transparent sort of process, transparent in that while a hawk is detected, nothing else is detected answering to the verb in 'see a hawk'. (Ryle 1949, p. 134)

He extends the observation beyond sensuous experience to a range of activities including deducing and concluding (*Ibid*.). More recently, it has been suggested (see e.g. Evans 1982; Moran 2001) that the Transparency of our reasoning or attitudes may be the key to understanding what is special or distinctive about first-person attribution of beliefs. At its core, the suggestion is that questions about, for example, whether *I believe that p,* need not—and typically do not—require the subject to focus on psychological facts, but are treated as—and are uniquely sensitive to—the non-psychological question, 'Is p true?'. By contrast, this is not typically how the question is treated when it relates to another person (see Edgley 1969; in Moran 2001, p. 60). This core thought has been remarked upon widely in self-knowledge literature since the turn of the twenty-first century:

> the idea that our standpoint on our own mental lives is in some sense "transparent" to our standpoint on the world at large has played an increasingly prominent role in philosophical discussions of self-knowledge. This idea has inspired important work on how we know ourselves to hold "judgment-sensitive" attitudes such as belief, desire, and intention, and it has also provided the impetus for a reconsideration of our knowledge of what our sensory and perceptual experiences are like. (Boyle 2011, p. 1)

More specific claims about how this Transparency might explain how we come to know our own beliefs and other states often follow Gareth Evans's (1982) remarks that, 'In making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward' (Evans 1982, p. 255). Despite a lack of argument, the claim is at least prima facie plausible, and the Transparency

'intuition' enjoys a good deal of support.[107] However, the intuition is accompanied by a 'puzzle'[108] (see e.g. Byrne 2005): How is it that conclusions about the '*p*-ishness of the world' (Schwitzgebel 2011), or 'how things stand in the world at large' (Boyle 2011, p. 5) *alone* can reveal the answers to questions concerning contingent psychological facts about a particular individual?[109] Responses to this puzzle can vary considerably. Before looking at two of them, it will be helpful to have in mind a clear expression of some common ground between proponents of the approach:

> 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. Indeed, it seems that, for various sorts of mental states, there is in the normal case no other way to attend to them: all there is for me to contemplate in my sensation of blue is the (apparent) blueness of some worldly thing, and all there is for me to attend to in my belief that P is the (apparent) fact that P. (Boyle 2011, p. 3).

In the next section, I briefly discuss a Transparency account that places *making up one's mind* in a position of central importance in explaining what is special and distinctive about self-knowledge.

2. Transparent deliberation

Evans (1982) provides the following example, which we can take for the purposes of discussion to be a standard case of doxastic self-knowledge:

[107] Evans's remarks, for instance, 'strike many as one of those things that are obvious once pointed out' (Byrne 2011a, p. 204).
[108] Some authors suggest that a number of puzzles arise from Transparency (e.g. Moran 2011). This one will be the main focus here. (See also Ch. 5).
[109] cf. Moran (2003): 'how can a question referring to a matter of empirical psychological fact about a particular person be legitimately answered without appeal to the evidence about that person, but rather by appeal to a quite independent body of evidence?' (p. 413).

> If someone asks me 'Do you think there is going to be a third world war?,' I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' (Evans 1982, p. 225)

This fits the profile of the intuition as described in §1, although notably, it appears to omit detail that would be helpful in answering the puzzle. One attempt to provide that detail is to gloss the cognitive transaction that occurs—that is, to allow conclusions about the world to be (or become) conclusions about a subject's psychology—as 'automatic' or 'immediate'. What counts for 'immediacy' in this respect often boils down to an absence of 'inference' (see e.g. Moran 2001, p. 91).[110] (In Ch. 2, I called this 'Epistemic Immediacy'.) And so, on an arguably natural reading of Evans's world war example above,[111] the self-knowledge procedure that is intended to explain what is special about first-person ascriptions of mental states is a process of deliberation that results in an 'immediate' transition from a conclusion about the world to a conclusion about oneself:

> I address myself to the question of my state of mind in a deliberative spirit, deciding and declaring myself on the matter, and [do] not confront the question as a purely psychological one about the beliefs of someone who happens to be me. (Moran 2001, p. 63)[112]

This view suggests one can know that one believes that *p* by considering the question of whether *p* is true and that knowledge comes about *immediately* once we 'make up our mind' (cf. Byrne 2005, p. 84). For the purposes of discussion, I will call this the 'deliberative view' of Transparency.

---

[110] Moran (2001) suggests that immediacy is 'in a way that does not depend on any external "medium," and which involves no inference from anything else' (p. 91). For an expression of the thought as it refers to a related case of Transparency, see Shah (2003, p. 447)

[111] Byrne (2011a) suggests that the reading is uncharitable (p. 208, fn. 11)

[112] Moran (2001) makes a distinction between theoretical and deliberative questions, with the former 'answered by discovery of the fact of which one was ignorant' and the latter 'answered by a decision or commitment of some sort' (p. 58).

2.1 Objections to the deliberative view

On the basis of this outline, we can see how the view might measure up to a number of the criteria addressed in chapter two. The factors going in its favour include that it meets the Peculiarity (P) condition, just in case the Transparency intuition is true (i.e. that first-person cannot be used to sufficiently explain knowledge in the third-person case); it is an Economical view (i.e. no additional introspective faculty is required such as is the case with Inner Sense theories and their descendants); and so on.[113] There are, however, a number of criteria where the view might come up short. For the present purposes, I will focus on two: Immediacy (both psychological and epistemic variants), and Uniformity.

In chapter two, I suggested there are two readings of Immediacy worthy of attention: psychological and epistemic.[114] An initial challenge for the deliberative view is that it looks incompatible with psychological immediacy (see Cassam 2014), because the process of arriving at a conclusion sees the subject attending to 'outward phenomena' by engaging in a conscious activity that aims at resolving an issue (see Owens 2011, p. 262; Ch. 4), and this kind of conscious train of thought is not plausibly immediate in the psychological sense. Of course, deliberation need not be fully conscious in that sense, and decisive elements within this kind of reasoning are not always introspectively available.[115] But even on a modest view of deliberation there must plausibly be some recognition that the deliberative question at issue is the one the subject is 'striving to answer' (Shah 2003, p. 466) and so, on either view of deliberation, attending to the relevant phenomena is a process with distinct parts.[116] Additionally, a number of the view's proponents (e.g. Boyle 2009, 2011a) do not suggest that concluding that *p automatically* puts one in a position to believe that one believes that *p*. On the contrary, it has become a staple of the literature

[113] Since this view is not the main focus of the chapter, my aim is not to provide it with the full critique that it deserves (and the view has already received a good deal of attention in the literature).
[114] Cassam (2014) offers the same distinction.
[115] I briefly discuss these possibilities in Appendix 1
[116] It has been noted that knowing one has reached a conclusion on this view is a 'piece of self-knowledge that needs accounting for' (Cassam 2014, p. 111).

to dispute over what the missing step might be.[117] Options include inference (e.g. Byrne 2005) and 'reflection' (Boyle 2011). But since neither of these are obviously psychologically immediate, the deliberative view must overcome two obstacles to meet that condition: the apparent mediacy of deliberation, and the potential mediacy of any step that follows it.

However, psychological immediacy does not capture what some deliberative views are trying to emphasise. The point can be—and often is—an epistemic one, concerned with the claim that our self-ascriptions are not *inferred* from 'anything else' (Moran 2001, p. 91). In this case the problem is how to make sense of the immediacy claim in light of the fact that that our conclusion about the world is reached by—that is based on—the relevant outward phenomena. Consider the following example response to Evans's third world war question:

> Sandrine concludes that there will be a third world war due to a series of factors that she takes to go in favour of that conclusion: (i) a global rise in the popularity of extreme nationalist rhetoric; (ii) a scapegoating of immigrants and minorities for domestic economic decline; (iii) a breakdown in international trade and arms agreements; (iv) the repealing of human rights; and (v) her recognition that a combination of such factors have preceded global warfare in the past.

Plausibly, when questioned about why she thinks there will be a third world war, Sandrine will offer these factors in her defence: Sandrine has reached her conclusion on the *basis* of these factors; that they are her reasons for her thinking the way she does. And so the deliberative view might also have a difficulty explaining epistemic immediacy.

In the view's defence, notions of inference can vary considerably. On one such notion, Sandrine has not inferred. The propositions that correspond to

[117] Byrne (2005, 2011a) suggests inference; Boyle (2011) 'reflection'.

Sandrine's thoughts (i) to (v) are not 'premisses in an inference to the conclusion' (see Cassam 2007b, p. 163) that she believes there will be a third world war. Granted the right notion of inference, then, epistemic immediacy—where 'epistemic' refers to the presence of inference[118]—and can be present in even quite sophisticated thought processes. The notion of inference that allows for this is quite narrow, and so the value of the result may be diminished, since the more exacting one's notion of inference, the less interesting non-inferential knowledge is likely to be (see Ch. 2; also Cassam 2009). However, because this is not the only difficulty for the deliberative view, we can grant a suitable notion of inference, and move on. The deliberative view, then, is not plausibly psychologically immediate, but can be epistemically immediate (given the 'right' notion inference). I will now briefly turn to two further challenges for the deliberative view: (i) that it fails to provide an account of self-knowledge for all cases of first-person privilege; and (ii) that it fails to provide a full account of first-person privilege for belief.

The first criticism is well trodden:[119] even if such a view plausibly fulfils all other desiderata, it is difficult to see how it could provide a uniform account of self-knowledge (see Ch. 2, §4). Because the view takes the deliberative (*active*) case to be central, it fits mental states that are reason-sensitive, such as beliefs, but is ill suited to account for mental states that are *passive*, such as sensations; it seems clear that self-knowledge of some states is both 'non-observational *and* non-deliberative' (e.g. appetites and 'unconquerable' emotions, Boyle 2009, pp. 138–9, fn. 8); and one might ask why deliberative self-knowledge is the important or 'fundamental' form of self-knowledge (Boyle 2009, p. 140), that is: 'the "one that makes the difference" between first-person

[118] Cassam (2014) argues that the deliberative view is not epistemically immediate (on his notion of it) because 'believing that you believe that P comes in part from your having justification to believe other supporting propositions' (p. 112). So one can retain this notion of inference, and still be 'in trouble' (p. 111) if the claim is meant to be that on the deliberative view of Transparency, self-knowledge is epistemically immediate in a more general sense.

[119] See Boyle (2009) for a discussion of those subscribing to the 'Uniformity Assumption'.

awareness and the kinds of awareness we might have of the mental states of another person' (p. 139).

The issue becomes pressing because it is not just sensations, appetites, and unconquerable emotions that appear to be non-deliberatively accessible. This is the second challenge. 'Brute', 'unreasoned', or 'recalcitrant' (p. 138) forms of attitudes that are deliberatively accessible appear to be first-personally accessible without deliberation, and access to belief and desire in many typical cases does not require one to make up one's mind either. My mind is already made up on the location of Nelson's Column, and whether I would like to return to Vienna (cf. Byrne 2011a; Boyle 2009). (Notably, these examples do look like plausible cases of psychologically immediate self-knowledge.) So besides a questionable explanation of Immediacy, the deliberative view is unhelpful with regards to a range of mental states (i.e. it is not a Uniform account), and for some unremarkable cases of attitudes such as belief and desire (i.e. some standard cases doxastic self-knowledge). In the next section, I discuss an alternative view of Transparency that incorporates these unremarkable cases of self-knowledge and aims to provide a Uniform account of self-knowledge.

3. Transparent inference

In the last section, I discussed a version of the Transparency approach that focuses on cases of making up one's mind, and pointed to a number of objections to that view. In this section, I outline an alternative view that tries to fill out some important detail thought missing in Evans's (1982) remarks about self-knowledge of belief in terms of a world-to-mind inference, and show how it fares better against the desiderata laid out in chapter two.

Alex Byrne's Transparency view suggests the following:

> Suppose that I examine the evidence and conclude that there will be a third world war. Now what? Evans does not explicitly address this question, but the

> natural answer is that the next step involves an inference from world to mind: I infer that I believe that there will be a third world war from the single premiss that there will be one. (Byrne 2011a)

An immediate difficulty with the suggestion is acknowledged by critics and proponents alike—the problem with reasoning from 'p' to 'I believe that p' is that it is a poor pattern of inference: it is 'neither deductively valid nor inductively strong'[120] (Byrne 2011a, pp. 203–204). One critic (Boyle 2011) as urged that it is the kind of inference only a 'madman' could draw and this alone should be enough to show that an Inferentialist interpretation of doxastic transparency should be abandoned. I will leave this issue aside for the moment (I return to it in Chs. 4 and 5).

On Byrne's view, doxastic self-knowledge is achieved by means of an epistemic rule (Byrne 2005) or doxastic schema (Byrne 2011a).[121] The rule for self-ascribing beliefs is BEL: 'If p, believe that you believe that p' (Byrne 2005, p. 95). As applied to Evans's 'third world war' example, to answer the question of whether I believe there will be a third world war, I consider the evidence relevant to the question 'will there be a third world war?' If I conclude that there will be one, I believe that I believe that there will be a third world war. The 'Gallois-style' schema—which is meant to be equivalent to the rule—can be illustrated as follows:

> *P*
> I believe that *P*

In general, the approach appears to be effective. Take the following example using the rule form of the procedure:

[120] 'that *p* is the case does not even make it likely that one believes it is the case' (Byrne 2005, p. 95)
[121] They are meant to be two ways of expressing the same procedure.

> suppose that p is 'it is raining'. To establish the antecedent of BEL, I look out of the window and occurrently judge, correctly, that it is raining. Because I recognize that it is raining, I implement the consequent, and so I come to believe that I believe it is raining … This latter belief will then be true. In general, whenever one implements the consequent of BEL because [one] recognizes the truth of the antecedent, the resulting second-order belief will be true (Gertler 2011a, p. 3).[122]

If we accept, for the moment, the suggestion that the procedure is 'strongly self-verifying' in lieu knowledge-conduciveness[123] we can see how the account is meant to meet two criteria: Epistemic Security (what Byrne calls *Privileged Access*) because 'belief's about one's own mental states are more likely to amount to knowledge' than beliefs about the mental states of others (Byrne 2011a, p. 202); and Peculiarity (his *Peculiar Access*), because 'the method only works in one's own case: inferring that Andre believes that p from the premiss that p will often lead one astray' (p. 207). (For a more detailed discussion Peculiarity on this account, see Ch. 4.)

Like the deliberative view, the account is also Economical, but unlike the deliberative view, it promises a *uniform* explanation of self-knowledge. The account, of course, is clearly not epistemically immediate by the measure in chapter two (due to the presence of inference), but it can be psychologically immediate as long as the subjects are typically unaware of the inference.[124] This is a promising start, so we might see how the theory fares in some other respects. I begin by highlighting a common style of objection to Transparency accounts.

---

[122] Byrne talks of belief rather than judgement, taking the latter to muddy the water, but the example can be reworked using belief.

[123] Byrne's (2011a) argument here suggests the burden of proof, to show how the procedure is not knowledge-conducive, rests with the opponent. Although he does offer a number of suggestions (e.g. that the judgements are 'safe') in its favour. (see Ch. 4).

[124] While it might not be made explicit in the material cited here, Byrne does not intend to suggest this inference to be one that we make consciously. This matter was made explicit in response to questions at the *Varieties of Self-Knowledge* workshop in Harvard University (2016).

3.1 The contamination objection

The inferential Transparency view looks amenable to cases of making up one's mind, even if—unlike the deliberative view—it does not take these to be the fundamental, or core, cases of self-knowledge. The general concern is that it is difficult to see how Transparency accounts are meant to enable a subject to be knowledgeable of a state she is in prior to the initiation of whatever Transparency procedure is deployed. One potential response for the proponent of the deliberative view is unsatisfactory: we do have access to such states but knowledge of them is not the *fundamental* case of self-knowledge. The inferential Transparency view (Byrne 2005, 2011a), however, aims at explaining all cases of first-person privilege in the same way, and acknowledges that deliberation is not the standard or fundamental case of doxastic self-knowledge. In order to see whether that approach has something more illuminating to say about the problem, we can look at a clear articulation of the objection.

Brie Gertler (2011a) suggests that following the 'BEL' rule (and by extension the doxastic schema) fails to account for a feature of Epistemic Security[125] in need of explanation. The feature it fails to account for is that: 'If I have no belief that p (at t_1) but consider whether I have a belief that *p* (at t_2) I will not self-attribute a belief that *p* without creating a new belief' (*Ibid.*; see also Ch. 2, §5). The concern is this: since following such a rule or schema can result in the formation of a new belief, it cannot be a successful method of assaying what one believes or judges 'at any moment other than the moment I complete my attempt' to reason in accord with that rule or schema' (Gertler 2011a, p. 5).[126] This presents a genuine challenge for the approach: a procedure that allows for the formation of a new belief that *p* in response to a question

[125] Gertler is responding in part to Byrne (2005) and so uses his terminology. In the case of Epistemic Security (his Privileged Access), I have argued that Byrne's formulation of the feature is a good one and adopted it (Ch. 2).

[126] As we saw in chapter two, this condition may need to be amended to include cases in which one has no belief that *p* (at t_1), and one forms a new belief that *p* (at t_2), but does not self-attribute the belief as a belief one held at (at t_1)—that is, one recognises that this is a belief one has *only now* come to have. (This looks like a good bit of self-knowledge, rather than a self-knowledge failure.) This can be put aside for the purposes of this discussion.

about whether one believes that *p* is not a reliable method of assaying what one *already* believes.[127] So a procedure will fail to explain Epistemic Security, then, as long as it fails to explain this feature of Epistemic Security.

The beginnings of a response to this problem lie in Byrne's account, which seeks to explain knowledge in both deliberative and non-deliberative cases. While one might deliberate in response to a question in some cases—when initially considering or reconsidering an issue—in many cases, my mind is 'already made up' and:

> no deliberation about whether p immediately precedes my forming of the belief that I believe p. I conclude that I believe that Obama was born in Hawaii, not after considering the evidence, but simply by recalling the fact that Obama as born in Hawaii. The (partial) explanation of why this procedure yields knowledge is exactly the same in both cases: I reason in accord with the doxastic schema, which is strongly self-verifying. (Byrne 2011a, p. 208)[128]

Since retained beliefs form the bulk of our attitudes, it makes sense that the standard case of doxastic self-knowledge will be a case that deals with already formed beliefs rather than ones one is currently forming. Beyond this suggestive remark, Byrne does not expand a great deal. However, intuitively, one might see how recalling the fact that *p* could begin to answer the concern outlined above while still meeting the desiderata: factual recall is not ordinarily seen as involving deliberation, and so together with some means of self-ascription we have the beginnings of a perfectly good way knowing the belief in place prior to the onset of the procedure. Believing that *p* is also, plausibly, a block on the formation of directly conflicting beliefs (e.g. the belief that not *p*) (see Byrne 2011a), and, the procedure will meet the Transparency desideratum, since the

[127] See the formulation of Privileged Access in Ch. 2 for more detail about what might count as failure in this case.

[128] Byrne suggests that Moran's claim the primary case of self-knowledge is a matter of 'making up one's mind' looks, 'On the face of it … like a conclusion drawn from an overly restricted diet of examples' (p. 208)

subject need only consider the relevant facts, rather than some aspect of her psychological makeup.

However, this does not completely answer the challenge, because the explanation of knowledge is meant to be exactly the same in both deliberative and non-deliberative cases. This suggests that whatever way one ends up arriving at the judgement '*p*', one will be in an excellent position to know one's beliefs if one reasons in accord with the doxastic schema. We have no reason, so far, to think that deliberation will not occur in cases where one has no prior belief or, perhaps, where one has reason, now, to doubt a prior belief. If in cases where S has no belief, or where something in the procedure causes doubt about an existing belief, she can still form a belief by putting into operation the doxastic schema. Thus, as a response to Gertler's concern it is, at best, partial in its current form.

This leaves us in an interesting position. The introduction of memory provides an initially promising response to this style of objection, but it is not obviously the response that Byrne has in mind. In an attempt to sketch that response, it is first worth noting that while Byrne suggests that the 'explanation of why the procedure yields knowledge' is same in both cases, the procedures—at least with regards to arriving at the judgement '*p*'—are importantly different.

Following the Transparency procedure involves the selection of one out of two possible operations (i.e. means of *arriving at p*) depending on whether one is responding to a matter that has already been settled, or whether one is considering the matter afresh. We can tentatively reflect this difference in the schema. When one considers the matter for the first time:

[Deliberate] p
I believe that p

And if the matter has already been settled (and there are no obvious defeaters):

[Recal] p
I believe that p

The question of how one schema is initiated over the other without introducing some other (possibly non-transparent) bit of self-knowledge will need to be addressed at some point. However, assuming this can be done, the procedure now appears to be a reasonably promising way around the contamination objection. For once one has formed the belief that *p*, one will not re-enter the deliberation process when prompted again, and therefore will not risk the formation and self-attribution of a new belief that one may take to have been in place before the initiation of the procedure. What seems already clear from this solution is that much will depend on the ability of memory to play that kind of role. The epistemology of memory, in that case, will be critical to whether a theory of self-knowledge can explain *Preserved Access* (see Ch. 2).

Before discussing which accounts of factual memory might be a good or bad fit for this role, I will first try to strengthen the case for the view that there are important differences between the recall and deliberation versions of the schema.

4. Mnemic and deliberative schemas

In the last section I suggested that highlighting the difference between cases of recall and cases of deliberation provides an initially promising response to a common concern about Transparency approaches. Here, I will point to several important differences between the two ways of responding an inquiry into what one thinks. The two are different in at least three respects: (i) one is psychologically immediate and the other is not; (ii) one aims at resolving an issue and the other does not; and (iii) one is more likely to be a problem with regards to new belief formation—that is, of the kind that motivates the contamination objection.

4.1 Immediacy and recall

In §2, I suggested that self-knowledge on the deliberation view of Transparency was not plausibly psychologically immediate, even though it may be espistemically immediate (i.e. assuming that inference is the only important kind of mediation). The inferential Transparency view, on the other hand, can explain psychological immediacy but not epistemic immediacy. In cases where one makes up one's mind, even on the inferential view, doxastic self-knowledge will not plausibly be psychologically immediate. It can explain psychological immediacy in the standard case of doxastic self-knowledge, however, which it takes to be non-deliberative. But it will not automatically explain it. Whether it does will depend on one's view of factual memory. (One's view of factual memory will also figure into a number of other differences.)

4.2 Deliberation and recall

In order for the suggestion that the two schemas are substantively different to have any teeth, we will need to show in what respects arriving at *p* via recall is different to arriving at *p* via deliberation. A natural place to start is with the features of deliberation, since I have already said something about the kind of process this is meant to be: a conscious activity that aims at resolving an issue (see §2.1; Owens 2011, p. 262). We can expand these features as follows: deliberation is (a) an activity, (b) aimed at resolving an issue, (c) which manifests some recognition of a questions, and deliberative questions are (d) typically transparent to other considerations (e.g. to factual inquiry) (see Appendix 1, §9).[129]

One might think that these features look like decent candidates to describe factual recall, at least as it operates in the procedure outlined above. It meets at least a number: it is an (a) activity; (c) that manifests recognition of a question—at least when deployed in response to a certain kind of stimulus (e.g.

[129] In Appendix 1, I argue that '(d)' is preferable to 'conscious'. The general point, that the features of deliberation and factual memory come apart in an important respect, is also made in that section.

a request for information); and (d) it is transparent (in this case to factual inquiry). The matter that will be decisive is whether factual recall (b) aims at resolving an issue in the same sense as deliberation. If it does, there is nothing to the suggestion that the two are different ways of arriving at the conclusion *p*. (*Ibid*.) But factual recall does not aim at 'resolving an issue' in that sense. One might think it does if one subscribes to a particular view of factual memory that sees recall as reconstructing our stored reasons or evidence for a conclusion; a process that weighs and balances items of evidence until one answer wins out. But factual memory can't typically be a kind of argument: 'The witness himself does not argue "I recall the collision occurring just after the thunder-clap, so probably the collision occurred just after the thunder-clap"' (Ryle 1949, p. 250).

Our interim conclusion is that non-deliberative (mnemic) and deliberative versions of the doxastic schema differ in two important respects: (i) the mnemic version of the schema is compatible with psychological immediacy whereas the deliberative version is not; and (ii) the mnemic version of schema and the deliberative version of the schema have different features in at least one respect. An additional difference between the two is their aptness to form new beliefs.

4.3 New belief formation

Deliberation is clearly a suitable way to form a new belief. Factual recall is not (or at least not obviously).[130] Whenever the subject engages wholly or partially in deliberation, there is a chance that a new belief will be formed. As long as the subject engages purely in a process of factual recall, the subject will either come up with a pre-existing belief, or come up empty handed. The ability to come up empty handed will be key to two features listed in our desiderata: Preserved Access and Agnotic Access. In short, deliberating in response to the 'Do you

[130] It has been argued, for example, that memory is not merely the preservation of contents (Lackey 2007), although a discussion of that argument is beyond the scope of this work.

think that *p*?' question does not—at least not alone—explain how one might explain either, whereas recalling that *p* in response to that question has a shot at explaining both.

All three provisional conclusions will directly depend upon the view of memory to which one subscribes (see Ch. 4 for a further discussion of accounts of memory). However, we can draw an interim conclusion about the prospects of the account explaining the implication of Epistemic Security highlighted by Gertler and some other desiderata: (1) the ability of the account to meet these desiderata will depend upon the epistemology of memory intended by the phrase 'recalling the fact that *p*'; (2) the clearest case of meeting the desiderata behind the concern addressed by the contamination objection, is a *pure* case of recall—that is, one that resists, or does not give way to a case of making up one's mind.

A difficulty for the particular Transparency view being addressed (i.e. Byrne 2005, 2011a) is that a version of contamination objection appears to re-emerge due to the account's attempt to provide a Uniform (Ch. 2) account of first-person privilege. That is, even if one were to accept that upon the 'right' view of memory, doxastic self-knowledge can avoid the contamination objection—that is, given appropriate amends to the account—the problem appears again, in a slightly more recalcitrant form when the account attempts to explain bouletic self-knowledge.

In the next section, I outline the inferential Transparency view's attempt to explain Uniformity.

5. Transparency and Uniformity

I have mentioned a number of reasons for thinking there is something to the Uniformity Assumption (see Ch. 2, and above). Certainly, Evans's (1982) remarks suggest he thought a Uniform Transparency account of self-knowledge is a possibility, and inferential view we have been discussing (e.g. Byrne 2011a) counts it among its explicit desiderata. (See Ch. 2, §4 for further discussion of this point.) We might, at least, accept on these bases that if Transparency procedures are 'cognitively possible' and have a 'tendency to generate correct answers' there is no obvious reason we would not deploy them frequently: 'We might … consider: "I want X" from X is good, "I'm afraid of X" from X is dangerous, "I hate X" from X is horrible, etc.' (Schwitzgebel 2011, p. 15).

Alex Byrne's inferential Transparency view forms perhaps the most comprehensive attempt to argue for a Uniform Transparency theory. In this section I outline attempts to describe Transparency procedures for thoughts and imaginings (2011c), desire (2005), and intention (2011a) before suggesting that the difference between the mnemic and deliberative versions of the schemas may reintroduce a version of the concern addressed above.

For illustrative purposes, Byrne's (2011c) Transparency rule for thoughts and imaginings is:

> THINK: If the inner voice speaks about *x*, believe that you are thinking about *x*

Byrne acknowledges that there is no actual inner voice or image, and so 'knowledge' of those objects must be impossible, although he suggests that we do attempt to follow the rule, and 'attempting' in this case, may be as close as we get to Epistemic Security in that case of thinking. Access to thoughts and imaginings will be *Peculiar* because only the subject in question will 'hear' the voice or 'see' the image in question.

A number of burning questions arise about THINK,[131] but the procedures for desire and intention, and their relation to belief, are worthy of greater focus for the purposes of this discussion. For self-knowledge of desires, Byrne (see e.g. 2005, p. 100) suggests something like the following:

DES: If ϕ-ing is a desirable option, believe you want to ϕ

And for intention, he proposes the (Gallois-style) 'bouletic' schema (2011a, p. 216):

I will ϕ
I intend to ϕ

Besides any intuitive misgivings about the suitability of Transparency procedures for these states, a number of things are notable about the attempts. Firstly, they are not strongly self-verifying in the way that reasoning in accordance with the doxastic schema is (Byrne 2011a), and so Epistemic Security may still need to be explained for these procedures.[132] Secondly, Peculiarity is questionable for the *bouletic* schema. Unlike the doxastic schema, it is not clear that following the third-person version of the schema is meant to be markedly less successful:[133]

Erica will ϕ
Erica intends to ϕ

These issues aside, it is the connection between the epistemologies of belief, desire, and intention that are most relevant to the present discussion. The following section will outline this connection and its importance to the account.

[131] For instance, the argument relies heavily on the discredited Perky Experiment (see Thomas 2013) that suggests a confusion between 'images and percepts', and the controversial Humean thesis that mental imaginings are merely downgraded perceptual impressions.
[132] Byrne (2011a) recognizes this, but does not address the point fully.
[133] I discuss Byrne's general approach to explaining Peculiarity in Ch. 4.

5.1 Procedures for belief, desire, and intention

The procedures for desire and intention will need further explanation: at the very least, in their current shape they will over-generate self-attributions for both states. I may, for example, always recognise the general *desirability* of *x* over *y*, while either being drawn to *y* or failing to be drawn to *x* in the way that could be suitably described as desire. Sometimes we do what we need to do, or what we 'feel like' in the circumstances, without judging it to be preferable, or better than other options. So, as things stand with the rule, one will self-attribute desires that one does not have. To counteract the problem, Byrne introduces some defeating conditions:

> Suppose one knows that ϕing is a desirable option, and considers the question of whether one wants to ϕ. One will not follow DES and conclude one wants to ϕ, if one believes (a) that one intends to ψ, (b) that ψing is incompatible with ϕing, and (c) that ψing is neither desirable nor better overall than ϕing. (Byrne 2011b)

These defeating conditions mean that one will not, upon considering whether one wants to go swimming near a warm sunny beach, believe that one wants to do so if, that is, one also believes (a) one intends to go to work, (b) that working is incompatible with swimming in such circumstances, and (c) that working is neither desirable nor better overall than swimming near a warm sunny beach. (At least if one does, it is by some means other than the rule for desire.) If, like the present author, a subject always seems to want the former but always seems to end up doing the latter, this looks like a strange result. Nevertheless, intuitions about desire and desirability can become confused, and perhaps the example does not do justice to the account. Importantly, the general strategy of allowing competing intentions to defeat the procedure looks promising when it comes to limiting the problem of over-generation, and so we can accept the defeating conditions as presented for the sake of argument.

What is clear from the defeating conditions is that 'the complete epistemology of desire partly depends on the epistemology of intention' (Byrne 2011b). So we should briefly examine the procedure for intention.

The schema for intention—the *bouletic* schema—will also over-generate self-attributions of intention. In part, this is because we can know the consequences of some of our actions without intending them: the *doctrine of double effect*. I may know, for instance, that I will wear out my training shoes by running long distances, but this does not mean that I intend to wear out my training shoes whenever I run long distances (Bratman 1984; in Byrne 2011a, p. 217). Equally, 'a tactical bomber … intends to destroy a factory and confidently expects his raid to have the side-effect of killing ten thousand civilians'. Although he does not intend to kill them, he knows he will (Bennett 1981; also in Byrne 2011a). Reasoning in accordance with the *bouletic* schema in this case looks set to result in one self-attributing intentions that one clearly does not have. Byrne's response to this difficulty is to adapt Anscombe's remarks regarding knowing what one is doing 'without observation':

> [One] can know what one is (or will be) doing 'without observation'. And those present and future actions that can be known 'without observation' are those that one intends to perform (see Byrne 2011, p. 218)[134].

Byrne's argument proceeds roughly as follows:

1. One can know what one is doing (or will do) without evidence that one is doing (or will do) it;
2. Those things that one knows one is doing (or will do) without evidence are the things that one intends to do;
3. In considering the question of what one intends, we take into account the following: 'if one's belief that one believes one will ϕ rests on good evidence that one will ϕ' one is not reasoning in accordance with the *bouletic* schema.

[134] Anscombe's own remarks refer to present actions. The idea that they can be adapted for future actions is contentious.

If the argument works, Byrne suggests, then Epistemic Security (E) is explained, because 'if one reasons in accord with the schema … and is mindful of the defeating conditions … then one will arrive at a true belief about one's intention' (p. 219). Peculiarity (P) is also explained because in third-person attribution, the defeating condition is 'almost invariably present': if I believe that Erica will walk into a lamppost, then 'I will think that this belief rests on good evidence' and the inference from *Erica will* ϕ to the conclusion *Erica intends to* ϕ will 'often be unwarranted' (*Ibid.*).[135] Let us grant that this deals with some initial concerns for the moment and apply the fix to an example.

I believe I will drive through town on Tuesday, and I believe I will stop several times en route. Leaving open the issue of why I believe the former, I can come to believe the latter by learning about the congested traffic, traffic lights, junctions, frequent floods, belligerent cyclists, and drunken pedestrians. If I take these factors to be good evidence that I will stop several times, then I will not reason in accord with the schema, and will not conclude that I intend to stop several times en route. The outcome is roughly in line with *double effect* intuitions, in that I can believe my progress will be halted because I will be driving through town, but I do not intend my progress to be halted on every occasion I think it will be.

However, some cases look less clear. Assume that I believe that I will drive through town next Tuesday, and I believe I will stop several times. Assume also that I believe the latter because I know what happens every time I drive through town on Tuesday and I know I drive through town on Tuesdays. Now, knowing that I drive through town on Tuesdays can be evidence that I will drive through town *next* Tuesday (i.e. not only good evidence that I will stop). And yet, intuitively, I can still intend to do so. Does the evidence in this case mean that I do not follow the rule and do not self-attribute the intention to drive through town on Tuesday? Not according to Byrne (2011a). The *bouletic*

[135] Note that the intuition can be easily skewed by a different selection of examples: Erica will go to the pub, Erica will get on a train tomorrow, Erica will eat breakfast before noon, all seem perfectly compatible with her intending to do these things.

schema is defeated when I believe that I believe that *p* 'because (and *only* because) I have good evidence for it' (p. 218; my emphasis). But what does it mean to say that I believe that I believe that *p* because (but '*not only*' because) I have good evidence for it?

The answer, I think, requires a piece of self-knowledge that Byrne has not accounted for—that is, reliable access to (i) the decisive factors for what we believe, (ii) our judgements about decisive factors for what we believe, or (iii) some other indication of what we intend. While some of the options are more plausible than others, none are particularly helpful to the account. Let us consider the two clear-cut cases before looking at which of (i) to (iii) is most likely to fit with the account. The cases are:

(A) S believes that she believes that she will ϕ because she believes that she has good evidence that she will ϕ

(B) S believes that she believes that she will ϕ but does not believe she has good evidence for her belief that she will ϕ

In both of (A) and (B), we can predict the outcome with regards to S following the schema and concluding that she will ϕ: in (A) she will not reason in accord with the schema and conclude that she intends to ϕ; in (B) she will reason in accord with the schema and conclude that she intends to ϕ. While we have not been told quite what the judgements in cases like (B) look like, we can hazard a guess: no evidence comes up whenever S considers the matter (i.e. she *comes up empty handed* when it comes to evidence for that belief) and so the schema is not defeated. However, that picture cannot be complete if there is possibility of the following scenario:

(C) S believes that she believes that she will ϕ, and believes she has good evidence for her belief that she will ϕ, but does not believe that she believes that she will ϕ solely because of her evidence that she will ϕ.

On (C), which matches my example above, the account predicts that S will reason in accord with the schema and conclude that she intends to ϕ. But in this case, it cannot be that she has drawn a blank in considering whether she believes that she believes she will ϕ on the basis of good evidence. Some evidence has shown up, and (on the face of it) the evidence is good. Since good evidence plus a gap or absence of (e.g. further) good evidence is just good evidence, what is missing from the picture must be one of three things: (i) S has access to the causes of her beliefs; (ii) S has access to her *judgements* about the causes of her beliefs; or (iii) there is a mental 'flag' (perhaps a bit of phenomenology or a judgement) that provides S with a clue about whether her beliefs about her beliefs that she will ϕ are based on any good evidence that presents itself.[136]

The prospects for (i) are not very good: there is no argument for this kind of self-knowledge in the account; it has come under sustained attack in both philosophical (Dennett 1969; Kornblith 2012; perhaps even Kant 1785);[137] and psychological literature (Nisbett and Wilson 1977; Johansson et al. 2005; Hall et al. 2012). The prospects for (ii) are slightly better: access to our judgements about the causes of our beliefs is both plausible—it goes unquestioned in much literature that denies us genuine access of the same kind, and is arguably Transparent (Byrne 2011a).[138] It is difficult to see how (iii) will avoid becoming a (non-transparent) indicator of intention that would leave the

[136] There are other possibilities. For example, S could judge that her evidence, while good, does not explain why she believes she is going to drive through town. For the purposes of discussion, I have assumed that this could be classed a judgement, from the first-person perspective, that the evidence is not 'good' enough for the belief in question.
[137] See Ch. 2.
[138] 'from a first-person point of view, an enquiry into one's evidence is (near enough) extensionally equivalent to an inquiry into one's beliefs' (p. 218)

bouletic schema without a great deal of work to do,[139] but in any case, the account makes no pretence of providing for such feature. The most promising response to the difficulty, then, is (ii), although if (i) did turn out to be plausible, it would do the trick.

Using our example above in conjunction with (ii), I now have access to my evidence and also to my judgements about decisive factors in favour of my belief that I will ϕ. So, plausibly, when considering whether my belief that I believe is based on good evidence, I sometimes gather my evidence but judge that it is not the evidence that caused me to believe that I will ϕ. Thus, I can proceed to reason in accord with the schema and conclude that I intend to ϕ. What is wrong with this version of the story?

Firstly, it will need to be shown how judgements about decisive factors in favour of 'my' believing that something is the case could be cast in terms of Transparency. Secondly, it is prone to counterexamples—in which my evidence (or my judgements about its quality), or my judgements about the decisive factors in favour of my belief, or both are unavailable[140]—and it relies on a particular view of memory in which evidence (or judgements about its quality) and judgements about decisive factors in one's beliefs are ordinarily first-personally accessible. This latter possibility brings with it the re-emergence of the concern expressed by the contamination objection.

In the final sections of the chapter, I rephrase the contamination objection in terms of a discussion of doxastic deliberation in Shah and Velleman

[139] Let's say, for instance, that S gets a special headache in such instances. Unless the headaches were somehow reliable indicator of intention and/or some other related state, it would still not explain why the subject would reason in accordance with the schema in the face of 'good evidence' for her belief that she believes she will ϕ. But if such an indicator is available, it isn't obvious that we need another means of detection.

[140] Consider a case in which I believe that my belief that I will ϕ is based on good evidence. However, both that belief, and that evidence are temporarily inaccessible. In such cases of, the subject will reason in accord with the schema and conclude that she intends to ϕ. For example: S is clumsy. He trips, stumbles and wonders whether he will fall, and thereby land on the floor. Ordinarily he would say that the belief is based on evidence (he usually ends up on the floor when he falls), but in this instance, neither belief nor evidence are accessible amid the panic and he concludes that he *intends* to land on the floor. We can also adjust the example for forgetfulness. Mere panic and distraction, or forgetfulness, ought not to result in the self-attribution of intentions, even if they occasionally mean we forget intentions we do have.

(2005) (§6). I then offer three attempts to locate the account of memory that might assuage that version of the concern with regards to the bouletic schema, and suggest that the obvious solution—which sees us ordinarily retaining evidence, or metadata related to its quality—is largely a myth. This leaves the inferential Transparency view with a problem that, on the face of it, can only be resolved by supplying the details of an appropriate epistemology of memory.

6. Transparency and doxastic deliberation

In the last section, we saw how the epistemology of intention supports the epistemology of desire, and how doxastic self-knowledge supports the epistemology of intention. Other things being equal, the account has met a number of our desiderata (see Ch. 2). Some of the finer detail needs may need further comment, however, going in its favour, there is a story about *Peculiarity* (P), *Immediacy* (I), and *Epistemic Security* (E) (I discuss these further in Ch. 4). The account is *Economical* and makes a spirited attempt at *Uniformity*. We have the beginnings of an explanation of how the account could provide *Preserved Access* for doxastic self-knowledge (albeit not the one intended by Byrne). There is still a concern with regard to the epistemology of bouletic self-knowledge, and potentially, therefore, for the account as a whole.

In this section, I expand upon my explanation of how inferential Transparency view might be able to explain *Preserved Access* for belief and discuss how it differs with Byrne's own account. In order to do so, I will return to the central case of doxastic self-knowledge. This will help to bring to the surface the difficulty that intention creates for Byrne's (2011a) account.

The question 'do you think that *p*' can be taken more than one way: (i) as an invitation for one to consider whether *p* (i.e. afresh), or (ii) as a question about whether one *already believes that p* (see Shah and Velleman 2005). In order to be able to meet *Preserved Access* a self-knowledge procedure must be able to

yield answers to questions about what one already believes. This is precisely the work that I have suggested can be done by Byrne's account as long as we take the mnemic version of the schema to be standard case of doxastic self-knowledge and have an appropriate epistemology of memory. I will briefly compare factual recall on Byrne's account to Shah and Velleman's (2005) discussion of attending to one's spontaneous responses to the questions of the kind we have been looking at:

> If the question is whether I already believe that p, one can assay the relevant state of mind by posing the question *whether p* and seeing what one is spontaneously inclined to answer. In this procedure, the question *whether p* serves as a stimulus applied to oneself. (Shah and Velleman 2005, p. 506)

On Shah and Velleman's (2005) account, the spontaneity of the response plays a vital role in ensuring that we are able to assay what we already believe as opposed to entering into the deliberative process:

> One comes to know what one already thinks by seeing what one says—that is, what one says in response to the question *whether p*. But the procedure requires one to refrain from any reasoning as to *whether p*, since that reasoning might alter the state of mind that one is trying to assay. Hence, asking oneself *whether p* must be a brute stimulus in this case rather than an invitation to reasoning. (Shah and Velleman 2005, p. 506)

In responding to the question whether I believe that *p*—taken as an inquiry into my state prior to the initiation of the procedure—I must avoid *any* reasoning as to *whether p* in order to avoid contaminating the answer:

> One cannot engage in reasoning aimed at answering the *question whether p* if one wants to find out what one *already believes*, because such reasoning would contaminate the result by possibly altering the state that one is trying to assay. (Shah and Velleman 2005, p. 507)

On Shah and Velleman's (2005) account, we know what we already believe by listening to what we 'spontaneously inclined to answer'—it is a way of *giving voice* to one's belief (fn. 29). However, this is a controversial, and at best partial, response to how it is that we know what we believe. 'Simply hearing oneself' utter something in response to a stimulus is not a good reason for thinking that one believes it 'any more than if one were to sneeze in response to the stimulus' (Moran 2011, p. 221). But while Shah and Velleman (2005) may have the wrong—or an incomplete—story about how we know what we already believe, they have highlighted an important consideration about how we can fail to know what we already believe.

Just as 'hearing one's' spontaneous response to a stimulus is only meant to be a way of knowing that one already believes that *p* because one has not engaged in reasoning as to whether *p*, recalling that *p* is only a way of knowing whether one believes that *p* as long as one does not contaminate the result by deliberating over whether *p*. The proposal here is that 'recalling the fact that *p*' can play the *preserving* role for a Byrne-style account that 'hearing oneself' is meant to play in Shah and Velleman's (2005) account.

We can see how this idea fares by considering Byrne's (2011a) examples, firstly in the case of belief. In responding to the appropriate question, 'I conclude that I believe that Obama was born in Hawaii, not after considering the evidence, but simply by recalling the fact that Obama was born in Hawaii' (p. 208). I *recall the fact*. I do not consider the evidence available to me to arrive at my answer; considering the available evidence is in any case a good way to potentially contaminate the result of the inquiry; and as long as the matter is settled for me, considering the evidence should not even be an option. Contrast

this case, however, with the kind of procedure that Byrne (2011a) has in mind for *bouletic* self-knowledge.

In the *bouletic* schema, the subject must be able to consider whether her belief that she believes she will ϕ is based on good evidence that she will ϕ. In order to do this, the subject must either assess or reflect on her belief that she will ϕ in light of her available evidence and the norms she accepts,[141] or make use of a retained judgement about the quality of that evidence. In the former case, there is a clear risk of contamination by reasoning as to whether *p*.[142] But, there is a question about whether there is a risk in the latter case too. On a plausible epistemology of belief, to have made up one's mind on an issue is to have 'closed the books on it' (Owens 1999, p. 317–18). Not only does one no longer require one's evidence for the ensuing belief, but plausibly neither does one any longer require an explicit positive assessment of that evidence. To suggest that we typically retain these assessments suggests two things: firstly, that we often form, encode, and retain beliefs about the quality of evidence for a decision we have already reached as well as the forming (and encoding) the ensuing belief itself, and secondly, that we might need to check again on our original assessment of our evidence (and/or the evidence itself) for the belief. But this would be unusual behaviour for ordinary believers. Typically, the fact that one believes that *p* is sufficient endorsement of *p,* and one doesn't need to retain any further endorsement (see Harman 1986). When one does, it is either a sign that *p* is likely to be challenged,[143] or that one hasn't quite closed the book on the matter.

The question for Byrne's account of bouletic self-knowledge is how to understand 'inquiry into one's evidence' (p. 218). On the view outlined above, it appears to suggest that the question is not settled for the subject, or could be re-opened by a fresh assessment of evidence. If this is correct, there are

[141] This is what I have called 'Evaluative Access' (Ch. 2)

[142] Assessing one's evidence or reasons for an attitude can result in a change to that attitude whenever there is a change in the environment, or a change in the individual (see Appendix 1).

[143] Owens (1999) suggests that 'deliberately retaining evidence for future consultation is a sign of doubt, an attitude appropriate to the scientist' but 'unsuited to the everyday believer (p. 317).

implications for the account's ability to meet the Preserved Access desiderata, its claim to Uniformity, and there would be restrictions on the account of memory that could be employed. Since Byrne's (2011) remarks on factual memory are suggestive, in the next section I consider a number of candidates for the kind of role that is being considered here.

7. Memory, evidence, and beliefs about evidence

In this section I briefly consider three candidate views of memory that one may be tempted to employ as a solution to the problem with the *bouletic* schema and its relation to do doxastic self-knowledge, above. The first is the view that factual memory delivers one's evidence for believing that one believes one will ϕ (perhaps by a similarly transparent means to which it delivers the content of one's first-order belief in the standard case of doxastic self-knowledge). Such views of memory are available. Experiential Foundationalism, for example, has it that propositional memory must carry with it an image or 'memory seeming' (Senor 2013)—possibly provided by episodic memory—in order to be justified. But Experiential Foundationalism faces the objection that it either leaves the majority of our beliefs unjustified—by only accounting for occurrent memory beliefs—or 'epistemically unaccounted for … by only providing conditions for the justification of occurrent memory beliefs' (*Ibid*.). Even if the evidence were available, considering its quality is a potential way to re-open the question of whether *p*, and is therefore unhelpful if one is trying to assay what one already believes (i.e. as opposed to forming a new belief). This view of memory, then, is unlikely to provide the kind of solution acceptable to a Transparency account of self-knowledge.

A prominent understanding of the relationship between memory and evidence allows for the fact that we often forget our reasons for adopting many

of the beliefs we currently hold (see e.g. Owens 1999; see also Burge 1993; Dummett 1993):

> Suppose I remember that Hitler committed suicide. I don't remember how I learnt this, nor can I lay my hands on anything that might count as (direct) evidence in favour of it. This is the situation we find ourselves in with the bulk of our factual beliefs: how do you know the boiling point of water or the dates of the First World War? (Owens 1999, p. 313)

On this view, memory preserves the content and rationality of the belief: the belief that *p*, rather than 'a sort of evidence for p (either prima facie or inductive)' (Owens 1999 p. 317f.):

> Once a question is decided, we close the books on it and throw away the key: deliberately retaining evidence is a sign of doubt … Memory is a faculty which preserves the probative and motivational force of evidence beyond the point at which that evidence has been forgotten. (Owens 1999 p. 317–18)

In order for factual memory to reliably provide the evidence for, as well as the content of belief, we would have to depart from this view of memory. As mentioned, however, the evidence itself might not be required as part of defeating conditions of the *bouletic* schema. What is required is that we take into consideration whether the 'belief that one believes one will ϕ rests on good evidence that one will ϕ', and this either requires either (i) the evidence that one will ϕ itself, or (ii) some retained assessment of that evidence. Initially, (ii) also appears to be at odds with the idea that 'believing that p is to have finished enquiring into p by forming the view that p' (Owens 1999, p. 317f.): it suggests that there is a matter to settle. Becoming conscious of the possibility that one's belief does not rest on good evidence is, after all, something the ordinary

believer is unlikely to do without some risk of re-opening the question of whether to believe it.

One might, however, point to a feature of the Preservationist account that might fulfil the function without the risk. On at least some Preservationist accounts, memory preserving both the content of the belief and 'probative and motivational force of the evidence' (Owens 1999). Perhaps this feature of factual memory provides a simple response to the difficulty created by the bouletic schema, by providing a means by which a subject may gauge whether her evidence is good. But this will not do either. The preservation of probative and motivational force is not a kind of judgement about the quality of one's evidence (i.e. it is not a judgement of the kind 'and the evidence was good' that one might append to any of one's first-order belief). So, if it is an item that is available to introspection at all, it is an item for which our access must be explained, and it is difficult to see how a Transparency procedure might explain it.

As yet, we have not identified the epistemology of memory that might explain the defeating conditions for the *bouletic* schema. Retention of evidence in some cases is possible, but standard retention of evidence is both implausible and plausibly risky with regards to the possibility of forming a new belief. Retention of judgements about the quality of evidence is also possible, but again, standard retention of such judgements looks unnecessary and implausible, at least on the Preservationist view of memory.[144] It is not obvious why ordinary believers would standardly form, encode, and retain beliefs about the quality of their evidence, nor is it clear on which view of memory the retention of such judgements would be standard practice.

At least one model of memory, however, comes equipped with a self-monitoring module that might help to explain the presence of (or access to) evidence-assessment data, and may get around some of the difficulties for the

[144] It also looks at least explanatorily unnecessary on Conservatism about memory, since we already have warrant to continue believing in the absence of defeating conditions. I will not argue that point here, however.

bouletic schema by, for example, simple and localised memory failure. This view suggests that the memory system comes equipped with a '*memory-monitoring module*' that directly detects the presence of a target in the 'memory store' even when that target is not currently accessible (see e.g. Yaniv and Meyer 1987; here in Koriat 1995). Thus, a 'feeling of knowing' might indicate the presence of information that one cannot currently retrieve. The model has been thought to explain, among other things, the tip-of-the-tongue phenomenon and the apparent accuracy of 'feelings of knowing'. And although the approach may initially look an unlikely fit for the Transparency intuition, from the subject's perspective, the process need be barely noticeable:

> Computer users are familiar with the concept of a *directory*. A directory contains only the names of the files stored on a computer disk; not the content of the files … when a computer is asked to retrieve a file from memory, the first step is to consult the *directory* … analogous to 'monitoring'. (Koriat 1995, p. 100f.)

Thus, information about the presence or quality of evidence could be stored separate to the belief itself. On the face of it, the analysis of factual memory above (e.g. Owens 1999) could apply, and still allow for independent information about the presence and quality of evidence to be stored in the monitoring system. We have little reason, so far, to suspect that retrieving 'directory-level' information about the presence or quality of evidence with regard to a specific belief would create the kind of difficulties I have pointed to in the proposed solution above. This response would suppose a two-stage process in which a 'content' search is only initiated if directory-level information is available. Something like this model may provide the kind of information required for the *bouletic* schema, while *insulating* memory content from inquiries into the reliability of evidence (and thereby potential contamination).

Unfortunately, there is also little reason to suppose that a memory-monitoring module is required to explain tip-of-the-tongue phenomena or the apparent accuracy of feelings of knowing. According to an alternative account (Koriat 1995):

> people have no knowledge of their memory over and above what they can retrieve from it. They cannot monitor directly the presence of information which they cannot access. (Koriat 1995, pp. 102f.)

Tip-of-the-tongue phenomena, and the apparent reliability of feelings of knowing can be satisfactorily explained by cues that 'reside in the products of the retrieval process itself', such as 'fragments of the target, semantic attributes, episodic information' etc. (*Ibid.*). So even if a memory-monitoring module could plausibly perform the requisite functions, and operate in keeping with the transparency intuition, defending such a model would be difficult to square the motivation for an economic theory of self-knowledge (especially since such a model of memory is suspiciously close to contemporary versions of the inner sense view that Transparency is concerned to eschew).[145]

This discussion of possible solutions has not aimed to exhaust all options for the Byrne-style account. The main purpose has been to demonstrate the tight connection between memory and self-knowledge on the account. The appeal of the Byrne-style account depends to an important degree upon the role of factual memory. But subsequent complications arising due to the inter-dependence of the epistemologies of belief, desire, and intention on such an account suggest that the account is, at best, incomplete. Further success for this style of account—specifically in terms of ensuring Preserved Access, but also with implications for other desiderata—is likely to depend on identification of a model of memory able to fulfil the required role while preserving Transparency intuition, and adherence to economy and uniformity. However, of the three

[145] Schwitzgebel (2011) remarks on the striking similarity of contemporary inner sense style accounts and memory.

possibilities considered, none appear to be likely candidates for resolving the problem raised by the epistemology of intention.

Conclusion

I have argued that the inferential Transparency view—focusing specifically on Byrne's (2005, 2011a) account—of doxastic self-knowledge is initially promising with regard to a range of desiderata outlined in chapter two. In particular, it promises the resources to respond to the objection—common to a range of accounts—that Transparent self-knowledge procedures fail to explain an implication of Epistemic Security by allowing a belief to be formed or altered by the procedure. I outlined a way in which making recall the standard case of doxastic self-knowledge might be a promising response to the objection, if one is clear about the epistemology of memory being assumed on such an account (although this is not what Byrne had in mind). By replacing deliberation as the standard case of the Transparency procedure, factual recall offers a prima facie reliable means of access to pre-existing beliefs that appears to fulfil a number of desiderata constraining a theory of self-knowledge. If we accept one view of memory that I have discussed here (e.g. Owens 1999), for instance, the inferential Transparency view would explain Psychological Immediacy and Preserved Access (thereby allowing the Epistemic Security criterion to be met); it is both Economical, and Transparent. The explanation of success with regard to these desiderata would substantively be down to the role that memory plays in the procedure.

However, I have argued that the specific requirements of the *bouletic* schema reintroduce the objection to Transparency procedures by requiring a subject assess her evidence (or judgements about its quality). In order for Preserved Access to be explained by the *bouletic* schema, a plausible epistemology of memory would need to be identified. In the last section, I

examined three candidate views of memory and concluded that none of them look promising.

Among the main aims of the chapter has been to show that the success of the view will rely upon the view of factual memory assumed on the account. I have suggested one way in which the epistemology of memory might help such an account meet a number of desiderata. This explanation looks unfit for the account's explanation of *bouletic* self-knowledge. But since none of the other candidate views of memory look likely to provide the appropriate detail, the account's overall success with regard to the desiderata will likely be the result of filling out relevant detail about the view of memory being assumed.

In the absence of that detail, we have seen how the epistemology of memory can contribute to the ability of a theory of self-knowledge to meet a number of the main and ideal desiderata. In the next chapter, I examine in greater depth the extent to which the epistemology of memory might explain or shed light upon the main desiderata for a theory of self-knowledge more generally.

4

Memory and Self-Knowledge

Introduction

In chapter one, I argued that our thinking about introspective failure and memory failure converge in a both commonsense and theoretical contexts, and in simple and complicated cases. I concluded (i) that memory plays an important role in explaining a good deal of what we sometimes describe as introspective failure; (ii) that we might see whether that convergence extends to introspective success, perhaps, thereby, shedding light on some intractable problems in the epistemology of self-knowledge; and (iii) that such an inquiry should proceed by highlighting the desiderata against which the success of a theory of self-knowledge can be measured. Chapter two laid out those desiderata. In chapter three, I set a promising approach to self-knowledge against these desiderata and argued that a specific epistemology of memory completes the explanation of how the account can meet a number of criteria. Specifically, I argued that a particular view of memory would enable the approach to meet an implication of the Epistemic Security condition; namely, Preserved Access. I concluded (iv) that the epistemology of memory plays an important role in explaining introspective success on such a view; and (v) that this strengthens the case for an inquiry into the extent to which the epistemology of memory can help to explain what is thought to be special or interesting about knowledge in the domain more generally.

The main task of this chapter is to take some further initial steps in that inquiry. Since the convergence of the two epistemologies has largely escaped the explicit—or at least detailed—attention of many contemporary philosophers in the field, the chapter will be largely exploratory: I am seeking to assess some of

the costs and benefits of available options rather than to propose or defend a specific theory.

I consider whether memory can be understood to explain, or shed light upon, any of the features described in our list of desiderata by focusing initially on those listed as minimum criteria (see Ch. 2). If there are sufficient options when it comes to the main desiderata, there is a good case for constructing a test theory of self-knowledge in which the epistemology of memory is explanatory of memory in the domain (see Ch. 5). I argue that, on several independently plausible views, memory can explain these features at least partially. If my arguments are successful, we can draw one of two conclusions: either the purportedly distinctive features of self-knowledge are more common than initially thought (and thus perhaps require no special explanation);[146] or, the epistemology of memory provides a partial explanation of why we take self-knowledge to have that particular set of features.

In §1, I briefly outline my use of memory terms by discussing a number of distinctions in the literature. This leaves me with a number of features that restrict the accounts of memory suitable for the purposes of the discussion. In §2, §3, and §4, I deal with the main business of the chapter—the question of whether and to what extent memory, as described, can be understood to explain the purportedly distinctive features of self-knowledge outlined as the minimum criteria for a theory: Peculiarity (P); Immediacy (I); and Epistemic Security (E), respectively.

1. Kinds of memory

The term 'memory' is used to point to a broad range of phenomena (see e.g. Sutton 2012; Byrne 2010; Matthen 2010), which sometimes appear in difficult

[146] Cassam (2009) reaches as similar conclusion with regards to the supposed *baselessness* of self-knowledge, suggesting that on one understanding it could not be a feature of self-knowledge at all, and on another it is commonplace, and thus requires no special explanation.

combinations: I remember her reassuring embrace; I (sometimes) remember why I came upstairs, where my keys are, and how to reset the boiler; I remember to check my emails, and that I own a copy of *Sense and Sensibilia*; I remember the feeling of warm sand between my toes; and remember thinking that I would prefer to be on the sand, reading *Sense and Sensibilia*, than checking my emails; and so on. Some kinds of memory barely register, in ordinary discourse, as memory at all:[147] the positioning and movement of a proficient typist's fingers is referred to as a 'skill', and my cognizance of the fact that 7 x 5 = 35 simply as 'knowledge'.[148] In this section, I assess a number of distinctions that have arisen to cope with the range of phenomena, and settle on a distinction between factual memory and memory experience. In §2.1, I outline some features of factual memory that are relevant to the inquiry.

In what I will refer to as the traditional hierarchy, remembering an event—such as standing in the sea—is commonly referred to as *episodic* memory; whereas remembering a fact—such as 'Sand stayed at Valldemossa'—is commonly referred to as *semantic* memory. Both have success conditions that are, ostensibly, directly related to states of affairs:[149] I only remember standing in the sea if I stood in the sea; I only remember that 'Sand stayed at Valldemossa' if she did. Both varieties are sometimes referred to as 'declarative' memory. Non-declarative memory—for example, remembering 'how' to re-chain a bicycle—is less ostensibly bound to particular states of affairs for its success: it does not reflect 'the world or the past in the same sense' (Sutton 2012). Cases of 'remembering how' also typically lack a conscious element present in memory of facts and events, and this points to another distinction in the traditional hierarchy: the term 'implicit' memory is used to describe skills and abilities that require no *conscious* 'memories'; 'explicit' memory is broadly associated with conscious 'memories' (see e.g. Bermúdez, forthcoming).

[147] See Appendix 1.
[148] 'Remember in this use is often … an allowable paraphrase of the verb "to know"' (Ryle 1949, p. 248)
[149] Sutton (2012) suggests that both episodic and semantic memory 'aim at truth' although I will set aside questions of teleology for present purposes.

On the basis of these distinctions, our traditional hierarchy carves up memory along the following lines: declarative (explicit) memory refers to 'conscious recollections of facts and events' (i.e. *semantic* and *episodic* memories), while non-declarative (implicit) memory refers to 'a heterogenous collection of abilities whereby experience alters behavior nonconsciously without providing access to any memory content' (Squire 1992, p. 233, see Fig. 1).[150]

The traditional hierarchy has a number of problems. Some cases of 'remembering how' (supposedly non-declarative, implicit memory) make use of conscious memories—semantic, episodic, or both—particularly for complicated or relatively new tasks, and may require specific (conscious) effort to retrieve the relevant information. And similar discomfitures are present elsewhere. Episodic memory is generally considered to be memory for 'personally experienced events' as opposed to mere facts (see e.g. Tulving 2001), but consider the following:

> Suppose I was so drunk at the party that I cannot recall dancing with a lampshade on my head. The next day I learn of this mortifying episode; later I remember what I learned, that I was dancing at the party in inappropriate headgear. I remember "a personally experienced event", or "what happened where and when", but it is semantic memory, not episodic. Contrariwise, suppose I have seen many skunks, and on that basis can recall what skunks look like. When I recall what skunks look like, I visualize a prototypical skunk, a perceptual amalgam of the various skunks I have encountered. Such a memory is best classified (at least initially) with the paradigmatic episodic memories—recalling seeing a skunk in my garden this morning, for instance. Yet it is not a memory of a personally experienced event. (Byrne 2010)

Despite the utility of some of its distinctions, the traditional hierarchy is insufficiently robust to accommodate a number of plausible memory scenarios.

[150] Also cited in Byrne (2010)

One response to such difficulties is to discard the semantic–episodic distinction, perhaps by viewing both as 'part of an integrated memory system, grounded in the sensory, perceptual and motor systems, and distributed across key brain regions' (McRae and Jones 2013). This is appealing in that it dispenses with the problematic distinction and, in particular, strips 'semantic' memory of its monolithic, 'amodal' status, but in doing so it makes sensuous content a necessary element of all declarative memory, and this risks leaving some straightforward cases of memory—for example, arithmetical knowledge—looking mysterious. Take our memory of a word like 'apple' on the 'integrated-system' model:

> the meaning of a word is grounded in the sensorimotor systems … Hence, when one thinks of an apple, knowledge regarding motoric grasping, chewing, sights, sounds, and tastes used to encode episodic experiences of an apple are reinstated via sensorimotor simulation. (McRae and Jones 2013)

Assuming that all memory of words and objects is supposed to work this way, the first thing to note is the unusualness of the activity being described. 'Thinking of' or *about* words, or objects, is a specific kind of activity, unrepresentative of the directed roles that memory of facts plays on a day-to-day basis. Taking this kind of activity as the standard case of memory risks leaving us with an impoverished understanding of what it is to remember basic facts. Consider the following example:

> Suzie is required to sit an exam in which is must be able to state the capital cities of a number of countries. The list includes Benin, which is a new one on Suzie. Fortunately, Suzie has a reliable source of information on such matters. The source indicates that the capital city of Benin is Porto-Novo and, relying on the source, she passes the test.

For the integrated-system model to be right, we must be able to answer the questions: What is the sensorimotor grounding for Suzie's belief that the capital city of Benin is Porto-Novo that enables her to recall it? What are the 'sights, sounds, and tastes' reinstated when Suzie thinks about Porto-Novo? Assuming that Suzie's source is purely testimonial, the sensorimotor grounding cannot easily be of the variety described in the 'apple' case. 'Thinking of Porto-Novo' for Suzie—unlike 'thinking of apples'—cannot be much more than thinking of its relation to Benin, or her testimonial source. But information about testimonial sources looks like a bad fit and, in any case, is information with which she can lose touch. In the standard case, the focus of the testimonial transaction is not the vendor, but what he offers. (Memory would be woefully inefficient if in order to remember that I bought a newspaper last week, I had to remember the newsagent's tie.) Even if sensuous source-data are always encoded, unless they cannot be lost, the integrated-system model fails to provide a sufficient explanation of an important class of judgments that would otherwise fall into the semantic, declarative category.[151]

What is helpful about the traditional hierarchy is that it allows us to distinguish between two varieties of memory: broadly, the variety best described as in examples of 'remembering' in common discourse, and the variety commonly associated with—and 'an allowable paraphrase of' (Ryle 1949, p. 248)—the verb 'to know'. What we need for the current purposes is a better way to distinguish between the two (not no distinction at all).

A helpful alternative contrasts 'factual memory', which preserves propositional content and is phenomenologically poor, with 'memory experiences', which are (comparatively) phenomenologically rich (Teroni 2015) and consist in some 'preserved acquaintance' (*Ibid.*) or 'cognitive contact' (Byrne 2010) with their subject matter. So, while my memory experiences of the walk to campus consist, in part, of some 'preserved acquaintance' with the

[151] Arguably, that for which we do have 'accessible' source-data could only represent a tiny fraction of what we overall know or believe (cf. Owens 1999).

constituent sights, sounds, and objects I encounter along the way, there is not a great deal that it is like for me to know that the body of water I cross is, hydrologically speaking, a continuation of the River Ure. Recalling the latter *might* conjure up a morass of associated images and other sensory data, but these data are not necessary for the fact to be recalled. There are more complicated cases: in the absence of a specific task—or in the presence of an usual one, such as 'thinking of'—memory might deliver a mixed bag of experiences and facts; on some occasions, memory experiences might aid in the retrieval of facts and vice versa; and we might make deliberate use of some associations to help with the encoding or retrieval of information. But none of this suggests that there are no clear cases on either side of the distinction, nor that one kind of memory is necessary for the other.[152]

1.1 Factual memory

With this distinction in place, we can begin to fill in some detail on the variety of memory that will be the main focus of discussion. If factual memory and memory experiences are epistemically independent, then we have all but ruled out some detail. 'Evidentialism' about memory, for instance, is the view that a subject is 'rational in believing that *p* iff believing that *p* fits the evidence the subject has' (McGrath 2007), but if factual memory and memory experience are epistemically independent, the rationality of believing that *p* on the basis of an experience that she recalls can be brought into question. What is more, since it is implausible that we could reliably 'dredge up' the evidence for more than a 'tiny subset of our beliefs at any one time' (Owens 1999), Evidentialism would leave us unjustified in the majority of our beliefs, even if our initial grounds for believing were perfectly acceptable.

Two approaches that do not face this problem are 'Conservatism' and 'Preservationism'. On the former, one is prima facie entitled to persist in believing that *p* (just) if one already believes that *p*. This view has a good deal of

[152] As Ryle (1949) remarks, there is 'no "must" about' the connection (p. 250).

intuitive force, especially if one also subscribes to the (Reidian) view that we have 'a prima facie entitlement to presume, of any belief, that it is well-formed and well-maintained and so is worthy of trust' (see McGrath 2007). However, it has some unusual consequences.[153] Prominent among them is the concern that the passage of time alone would transform the retaining of a belief from unreasonable to reasonable.[154]

Preservationism also absolves the subject of the requirement to retain her reasons and evidence for currently held beliefs, but in this case by preserving whatever rationality was in place when the belief was formed. Memory preserves 'the probative and motivational force of evidence beyond the point at which that evidence has been forgotten' (Owens 1999, p. 318). Thus, on the Preservationist's view, if it was not reasonable to hold that *p* in the first place, it will not become reasonable to do so purely by virtue of the fact that I continue to hold that *p*.[155] It faces the objection that memory does not appear to be preservation alone (see Lackey 2007).

For the remainder of this chapter, I will take Evidentialism to be false and the disputed territory to be contested by Conservatism and Preservationism. For the most part, there will be no need to decide between the two, because both views are compatible with a range of features required to explain factual memory. (Where there is, I will make this explicit.) These features are: (i) phenomenological paucity, and (ii) a prima facie epistemic authority, that allows us to (iii) relinquish reasons or evidence for our attitudes while being warranted in retaining those attitudes as long as that prima facie authority has not been defeated.[156] With these features in place, we can set

[153] McGrath (2007) discusses what he takes to be the 'best three': (i) 'that conservatism wrongly privileges our own beliefs over others' beliefs', (ii) that it 'allows mere belief to make one rational in believing something when one previously wasn't', and (iii) that 'it allows mere belief to provide an extra epistemic boost to a subject who, prior to forming the belief that *p*, already was rational in believing that *p*' (p. 14).

[154] This is a variation on Burge's (1997) 'conversion' objection (see McGrath 2007, p. 14). Space unfortunately prohibits lengthy discussion of the debate. (Thanks to Richard Flockemann for some helpful informal correspondence.)

[155] Various forms of the position have been supported by, for example, Burge (1993), Dummett (1994), and Owens (1999). Detractors include McGrath (2007) and Lackey (2007).

[156] In contrast, memory experiences exhibit (i) phenomenal richness (comparative to factual memory), and (ii) a preserved acquaintance, or cognitive contact, with some event or object encountered by the subject.

about comparing the features of memory with the main desiderata from chapter two.

1.2 Distinctive features of self-knowledge

Chief among the desiderata from chapter two were three common features. Explaining or eliminating some variation of these has been the pre-occupation of much literature:

(P) Peculiarity
(I) Immediacy
(E) Epistemic Security

In chapter two, I suggested they form the minimum criteria for a theory. In this section, I argue that the features of factual memory can help to explain why it is that we take self-knowledge to be peculiar (P), immediate (I), and epistemically secure (E). I will address them in that order.

2. Memory and first-person peculiarity

The Peculiarity thesis suggests that S's method of acquiring (or making conscious)[157] a belief that she believes that *p* is unavailable to anyone else. I formulated this in explanatory terms (Ch. 2, §1), as follows:

> ***Peculiarity***—a method or procedure by pointing to which it is possible, satisfactorily, to explain how S comes to knows S's mental states, and that cannot be used satisfactorily to explain how S comes to know the mental states of others.

[157] This makes room for approaches to self-knowledge that see the subject with a tacit belief that becomes explicit by means of reflection (see e.g. Boyle 2012).

In this section, I suggest that, intuitively, we have reason to think a person's access to her memory is peculiar in the sense intended by the (P) thesis (§2.1). I then assess a number of ways in which one might attempt to defend that intuition. In §2.2, I contrast two varieties of recall—recall of first-order beliefs, and recall of second-order beliefs—and compare the degree to which they might be help to explain Peculiarity. In §2.3, I discuss whether supplementing the inferential Transparency view with the appropriate epistemology of memory might better explain how that approach explains Peculiarity. These are not the only options available, but they will serve to demonstrate that a number of options for a sufficient explanation of Peculiarity are available within our thinking about memory.

2.1 The intuitive peculiarity of memory

Intuitively, we might suggest that on any plausible view of memory, it cannot help but meet the Peculiarity condition, since—science fiction and fantasy aside—Bruce has the kind of access to his memory that no-one else has, and if Bruce were to explain how he knows that *p* by pointing to the fact that Jennifer recalls that *p*, we would proclaim his explanation had come up short of sufficient (i.e. to be missing some crucial step, such as Jennifer also telling him she recalls that *p*). The view has a kind of appeal it is difficult to deny, but pinpointing precisely what makes that view appealing is less straightforward.

To begin with, we can try to make the intuition that access to one's memory is first-personally peculiar explicit. A first-pass description might look like this:

> (IPAM) If S can sometimes know she believes that *p* by recalling that she believes that *p*, S knows it in a way unavailable to anyone else.

This description makes it clear that at least two issues need attention: firstly, it merely states the first-person peculiarity of memory, and secondly it is ambiguous between two readings of recalling that *p*—one on which a first-order belief is recalled, and one on which a second-order belief is recalled.

One might choose to address the first matter by removing one obvious obstacle to the possibility that IPAM is a good intuition. (This can be dealt with briefly since the issue has been broadly addressed elsewhere, see e.g. Ch. 2.) A Parity theorist about self-knowledge resists the idea that there is any difference in kind between first-person and third-person access to the mind (e.g. Ryle 1949; Carruthers 2011). There is little reason to suppose she would reject this peculiarity when talking about self-knowledge, but accept it when talking about some other cognitive faculty. She might, for instance, object that intuitions about the peculiarity of first-person access to memory are just another example of mistaken intuition about the mind.[158]

Actual Parity theorists, on the other hand, appear to be more amenable than this: they are less concerned with access to some aspects of our mental lives than when they are tackling the subject of our introspective abilities more generally. Both Ryle (1949) and Carruthers (2011) rely heavily upon silent soliloquy and memory. Ryle (1949), for instance, suggests that when memory works promptly, I can 'catch myself' ϕ-ing, though not 'in the same sense' that I catch someone else ϕ-ing (p. 148).[159] And memory, on Ryle's view provides us with a 'mass of data' that contributes to our views of our behaviour and thoughts (p. 149) even if that does not amount to 'Privileged Access'.[160] In any case, the suggestion that Parity theorists deny any significant first-person access to the mind appears to be an exaggeration (see Byrne 2010), with the

[158] Mistaken, but sometimes useful: assuming that we have excellent access to our minds could allow for quick decision making (see Carruthers 2011).

[159] Ryle uses swearing as an example, which alone may leave the passage open to alternative interpretations, but he also suggests, 'I can report the calculations I have been doing in my head' (Ryle 1949, p. 148), which I take to limit the available interpretations. See also Alex Byrne (2012) in reference to what I have called the Parity Thesis: 'Surely I don't know that I feel an itch, or see a duck, in the same way I know that you feel an itch or see a duck!'.

[160] The point being that 'more data' doesn't always make us better judges, although being in an espistemically advantageous position is not what is at issue here.

'authentic' (Ryle 1949, p. 148) processes of memory carrying some of introspection's load for a number of such theorists (see Ryle 1949; Carruthers 2011). The actual Parity theorist, then, does not typically pose a serious problem for the intuitive view expressed above. This does not address the problem posed by the notional Parity theorist, of course, although there is a sense in which she need not be concerned either.

In a minimal sense, the thought that there is *something* different about first-person and third-person access to memory is fairly uncontroversial. Even philosophers who defend the thesis that we can literally perceive the mental states of others (see e.g. McNeill 2012) do not suggest that our *perceiving* that someone is in a state is the same as their *being* in that state.[161] The difficulty is that first-person peculiarity in this minimal sense would see access to memory on a par with seeing, hearing, and feeling, etc., such that S being the subject who sees *x* would be sufficient to suggest her access is *peculiar*. In a sense, this is right, but it cannot be all there is to it. We can see, hear, feel and recall that *p* without knowing that we see, hear, feel and recall that *p*. Whatever asymmetry is granted to S in virtue of the fact that she is using her own faculties is not quite enough to explain how it is that she can know that what they come up with refers to her. So, more needs to be said about how pointing to S recalling that *p* sufficiently explains how she knows that she believes that *p*.

The appeal of (IPAM), I take it, is at least in part due the ambiguity between belief contents. In the next section, I highlight the differences between the two contents and their relevance to our present inquiry.

2.2 Recalling first-order and second-order beliefs

For the purposes of discussion, take the following hypothesis to capture the relevant intuition about the peculiarity of access to memory:

[161] The claim is rather that by sometimes seeing 'aspects of each others' mental lives', we 'thereby come to have non-inferential knowledge of them' (McNeill 2012, p. 573)

(PAM) One can know that one ϕs that p by recalling that one ϕs that p, thereby knowing that one ϕs that p in a way unavailable to others.

One thing to note about (PAM) is that it leaves in place the ambiguity about precisely what is being recalled. We can situate the ambiguity in a case of self-ascription:

> I ask Laura if she believes that there is anything wrong with gay men having consensual sexual intercourse, and she answers that she sees nothing wrong with it … she might partly be calling up the moral facts (or putative moral facts) from memory, much as a schoolchild might call up California's capital from memory … Laura's self-ascription might be partly driven by … her memory of having explicitly endorsed similar propositions in the past. (Schwitzgebel 2009, pp. 48f.)[162]

The example contains two plausible cases of memory simpliciter: there is little doubt that we can enjoy both, and both are partial explanations of doxastic self-knowledge. One involves the kind of result we might expect from a Transparency account of self-knowledge: the question of whether Laura thinks something is wrong is settled by what she takes to be the salient facts. The other case sees Laura considering what she has thought in the past, and this can be relevant to her conclusion in at least some cases. The two cases are not the same kind, and it is not obvious that both are cases of factual recall. But they can be made to look that way:

> Just as a stored representation of the fact that "Plato taught Aristotle" might influence a variety of my judgments … a stored representation of "I believe that Plato taught Aristotle" could do the same. It would be a strange incapacity

[162] Schwitzgebel (2009) is arguing for a pluralist view of self-knowledge of belief (i.e. these are listed among numerous ways in which Laura might come to know her belief in this case). My use of the example is purely to indicate a particular diversity in memory based self-ascriptions.

> if the latter were not the sort of thing I could directly remember (Schwitzgebel 2009, pp. 51f.)

So, if the two are different, we need to be quite clear about the implications of that difference when it comes to the reliability of memory based self-ascriptions. We can contrast the two cases as follows:

> (R*p*) S recalls the fact that *p*
> (RB*p*) S recalls that she believes that *p*

Characterizing (R*p*) in light of what has been said about factual memory can be fairly straightforward: S recalls some factual content answerable only to states of affairs in the world or the past. When S recalls that 'The 'Minute' Waltz is in D-flat Major', for instance, she recalls a salient fact about that waltz, not something about her own psychology. In that respect the operation is Transparent.

By contrast, characterizing (RB*p*) is less straightforward. One possibility is that (RB*p*) describes a memory experience—for example, being suddenly convinced of truth of *p*. On this understanding (RB*p*) is akin to cases in which one recalls becoming or being sad, say, at a funeral, or being in pain, say, at the hospital. The sadness and pain are not themselves recalled, only some mark of one's past awareness of them. Likewise, belief that one believes that 'The 'Minute' Waltz is in D-flat Major' is not what the memory provides, but just some mark of becoming at one time aware of that belief. While the content is still ostensibly factual, it better fits the profile of a memory experience because it relies upon preserved contact or acquaintance with an event or experience. An appropriate way to describe (RB*p*) in light of our distinction is that 'S recalls believing that *p*', where 'believing' picks out a putative episode or event in

which the subject took herself to be in a particular state (or to have some range of features that are equivalent to or evidence for that state).[163]

Eric Schwitzgebel (2009) appears to consider something like (RB*p*) a genuine means of doxastic self-knowledge.[164] It is worthwhile briefly considering whether how such a suggestion might be fleshed out, since it would be a very swift result for the present inquiry. Firstly, we can see a number of opportunities to fall into error, among them following: (i) S could have a 'false memory' of judging that she was in a particular state (e.g. S judges that she was in state F when she was not, in fact, in state F); and (ii) S could be mistaken about the kind of state she was in (e.g. S takes herself to have believed that *p* when in fact she hoped that *p*). So if this is a way of coming to know our minds it leaves us appropriately fallible. On the other hand, there is a variety of error into which the subject looks unlikely to fall into—(iii) the identity of the individual making such a judgement—and this is a promising candidate for Peculiarity, and potentially for Epistemic Security (see §4).[165]

We might be tempted, then, to conclude that (RB*p*) judgements are important and interesting cases of self-knowledge: important because recalling an episode of believing that *p* is one way to allow a subject to make sense of an otherwise mysterious change of heart; interesting for our purposes because they appear to possess a number of features relevant to the current inquiry. If the judgements possess the kind of immunity suggested in (iii), not only are they first-personally peculiar, and potentially more secure, they also require no further self-ascription. On the downside, it is difficult to see how such judgements could be Transparent (see Ch. 5), and—more pressingly—the

[163] She may recall, for instance, feelings of endorsement, etc. See Cassam (forthcoming) for a discussion of psychological evidence for our beliefs.

[164] Schwitzgebel's pluralism (2009) and scepticism (2008) do not leave him with a difficult set of strict criteria to meet with regards to introspective success. However, his remarks here do appear to be a genuine attempt to point to a procedure for doxastic self-knowledge. He also takes memory to be a good candidate for self-scanning mechanisms (2009).

[165] I refer here to the property of immunity from error through misidentification (IEM) in what Bermúdez (2012) calls autobiographical memory. (See §4, Ch. 5).

epistemic distinctiveness of (RB*p*) alone is not a sufficient explanation of how S knows that she believes that *p*.

As we have already seen, memory experiences that sometimes accompany a belief are neither necessary nor sufficient for recalling that belief (Teroni 2015). In many cases, the experience is lost, but the belief remains (I frequently forget my evidence for what I now believe). In others, the belief is lost even when the experience remains (I remember thinking many things that I now take to be foolish). This epistemic independence means that the procedure will only issue in knowledge when the episode of believing is ongoing, but since correctly recalling that I judged *p* to be true in no way suggests that I still judge it true, deploying (RB*p*) as a general guide to current states would be ill-advised.

An additional worry arises from research that suggests autobiographical memory functions so as to preserve and protect a coherent picture of the self rather than to reflect the world or the past (Conway and Loveday 2015). Since these memory experiences are autobiographical in that they relate directly to events in one's past, if the conclusions of the research are correct it is possible that (traditional) knowledge-conduciveness is out of reach.[166] Thus, while (RB*p*) is a natural example of how memory might provide self-attributed mental states, and one that initially looks promising with regard to Peculiarity as well as a number of other desiderata for a theory of self-knowledge, it is a poor response to the Peculiarity criterion alone insofar as it is does not sufficiently explain how the subject might know that she believes that *p*, simply by recalling her *believing that p*. It does, however, point to important ways in which judgements might be thought to meet the Peculiarity condition—namely, by being automatically self-ascriptive, possessing the immunity property, and revealing (in some cases) information about attitudes in place prior to the initiation of the procedure.

[166] Bermúdez (forthcoming) expresses a similar concern about coherence models of memory and the prospects for knowledge.

Now compare the benefits and deficits of (RB*p*) judgements with our standard case of factual memory: S recalls the fact that *p* (R*p*). The first thing to note is that, unlike (RB*p*), (R*p*) is not automatically self-ascriptive: one can recall of a subject that *p* without realizing that one is the subject. So this kind of memory judgement does not possess the immunity property potentially possessed by (RB*p*). One might conclude if (R*p*) is to be a part of a procedure for doxastic self-knowledge, it will need an additional self-ascriptive step (as with the Transparency views discussed in Ch. 3). Chapter three addressed a number of options in this respect: one is a process of 'reflection' (see Boyle 2011), by which a subject comes to judge explicitly that which she already tacitly knows; and another is an inference from world-to-mind (e.g. Byrne 2011a). In chapter three, I took the inferential approach to be more promising in light of our desiderata, so I will continue with that model for the purposes of discussion here. Here is an adapted version of Alex Byrne's Gallois-style doxastic schema with recall made explicit as the operation that leads to the judgement '*p*'.[167]

Rp
I believe that p

As a way of knowing, it avoids two problems associated with (RB*p*) because (i) it does not rely upon cognitive contact with some past event, and (ii) it delivers a current, first-order belief rather than the mark of a past self-attribution of some state. (Concerns about the distorting affects of autobiographical memories are also avoided because, plausibly, we are not dealing with memory experience.) Since the procedure does not face these obstacles, we can assume for the moment that it is a *credible* way of knowing one's mind. The question is whether it is a peculiar way. Answering this question (§2.3) requires that I briefly re-open a discussion briefly addressed in chapter three.

[167] The addition of 'R' to represent recall merely highlights the means by which the judgement *p* is reached, rather than indicating, for instance, that the subject recognizes that she is 'recalling'.

In this section, I examined a potential ambiguity that may be the source of the intuition behind (PAM). The ambiguity lay in the belief content that the procedure delivers: (RB*p*) has a number of features that make it a promising explanation of (PAM). Alone, however, it does not look like a sufficiently robust explanation of Peculiarity because recalling an episode of believing that *p* is a poor overall guide to one's current beliefs. Contrariwise, (R*p*) is an excellent guide to what one currently believes, but does not look like a sufficient explanation of Peculiarity because it does not, alone, provide a means of self-ascription. In the next section, I explore whether (R*p*) can help to explain Peculiarity by contributing to the inferential Transparency view mentioned above.

2.3 Peculiarity and the doxastic schema

In chapter three, I agreed with Byrne (2011a) that following the doxastic schema explains his Peculiar Access condition, 'because, the method only works in one's own case: inferring that Andre beliefs that *p* from the premiss that *p* will often lead one astray' (Byrne 2011a, p. 207). This was an oversimplification. In this section, I argue that making the role of memory in a doxastic schema explicit helps to explain Peculiarity on that account.

According to Byrne's (2011a) view, a way of knowing one's mind fits the bill if it is available to S and no-one else. However, the difference between first-person and third-person deployment of the procedure, as Byrne (2011a)[168] has it, is a matter of comparative reliability. Inferring that Andre believes that *p* from the premiss that *p* is, of course, an option for Alex, even if it is far less reliable than reasoning in accord with the schema with regard to one's own beliefs. The worst that can be said of the third-person use of the schema is that we wouldn't normally recommend it.[169] But things that work aren't always

[168] Alex Byrne has reiterated the position, for example, at *Varieties of Self-Knowledge Workshop*: Harvard University, March 2016.

[169] I raised this issue with Alex Byrne in private correspondence following the *Varieties of Self-Knowledge* workshop (2016) mentioned above. At the time, I took the position iterated at the workshop to be a

things we'd recommend. Cleaning oil spills happens to be a surprisingly effective method of discovering new marine life,[170] but since enthusiasm for the conditions that make it possible is understandably thin on the ground we are not likely to try to bring them about purely for the purposes of research. Byrne's (2011a) Peculiar Access is cast in terms of availability, but since it is available for third-person use, his attempt to explain Peculiar Access is, by his own lights, incomplete.

However, because it is the utility of the method for the third-person that Byrne clearly has in mind, we might try to make sense of it in those terms. But that does not look wholly promising either. While it seems obvious that the third-person use of the method would be less fruitful than the first-person method in this case, calculating how deficient the third-person method is would be difficult. In any case, the third-person use looks attractive as a quick, easy, and somewhat successful method of predicting the behaviour of others, and one that we plausibly make use of with some frequency. Both humans and the objects of their experience are highly predictable. In general, humans need only to negotiate (i.e. form true beliefs about) a strictly limited environment in order to survive; and do manage to negotiate that environment quite successfully for a number of years without mastering the notion of belief (see e.g. Gopnik and Astington 1988). In order to interpret others we need to attribute to them largely true beliefs (Davidson 1973), and it is unlikely that we ever quite manage to shake the childhood practice of thinking that people know, pretty much, what we know. (We are surprised, after all, when people make foolish decisions with what we take to be highly predictable outcomes.) The third-person use of the doxastic schema, supplemented with appropriate defeating conditions, is unlikely to be all that bad.

---

departure from the position as it can be found in print. However, it is clear that this is a faithful interpretation of his position as it stands at the time of writing. Professor Byrne responded to say that he ought to have said that it was a 'isn't a good idea' to follow the schema in the third person case.

[170] A Woods Hole Marine Biological Laboratory study found that, in fact, the large majority of 'new marine life' was discovered that way over a period of fifty years leading up to the survey (Shultz 2016)

Fortunately, we need not pursue the exercise in comparative reliability. Unlike Byrne's original formulation of Peculiar Access, the revised formulation in Peculiarity (P) does not face the same problem of explaining why reasoning in accordance with such a schema can be considered peculiar. Whatever can be said for the third-person use of the doxastic schema, it cannot be pointed to alone if one wishes to sufficiently explain why S knows what she knows. While it might be good enough to 'get us through the day' (Dunning 2014),[171] the fact that we can get away with using (almost certainly very many) deviant cognitive practices (see Ch. 1) does not sufficiently explain how we know the contents of others' minds. In short, on my formulation (P), Byrne's (2011a) procedure for doxastic self-knowledge is more plausibly *peculiar* than on his own view.

But here we run into an awkward problem, because it is also questionable whether the first-person use of the doxastic schema is a sufficient explanation of how we know our own minds, and this is also required upon (P). As we have seen (see Ch. 3), the world-to-mind inference is at best shaky, and at worst mad: it is a highly efficient source of true beliefs but, other than that, it looks nearly as bad as the third-person use. Byrne (2005) sketches a solution in terms of 'safety'—the beliefs yielded by the schema could not easily have been false—but stops short of endorsing it as a demonstration that reasoning in accord with the schema is knowledge conducive (Byrne 2011a, pp. 206f.). Such attempts to complete the explanation may turn out well, but I would like to offer an alternative way to explain how reasoning in accordance with (the recall version of) the schema can explain why the procedure is a peculiar way to know.

In chapter three, I compared two ways of responding to an inquiry about what one thinks: a deliberative procedure and a non-deliberative (or mnemic) procedure (also §2.2 above). Both are activities that manifest a recognition of a question, and are transparent to factual inquiry. Unlike deliberation, however,

[171] David Dunning (2014) artfully makes a similar point about a range of deficient cognitive practices. All we really need from our pattern recognition and theorizing abilities is that they they get us to 'an age when we can procreate'. If they are good enough to do that, we have little reason to jettison them.

the mnemic procedure does not aim at resolving an issue (see Appendix 1). In deliberation, recognition of the question provides direction or purpose during the period of cogitation (see Shah 2003, p. 466) on the matter to be resolved. The mnemic procedure, however, requires no such directed period of cogitation because the matter is already resolved: one may try to rehearse one's reasons or evidence for thinking what one does, but doing so will at best risk re-opening the inquiry, thus potentially contaminating the results of one's attempt to self-know (Shah and Velleman 2005); at worst one will confabulate (see Appendix 1). In typical cases of factual recall one has no access to one's original reasons or evidence for a belief (see e.g. Owens 1999). What is delivered by the mnemic procedure is the encoded attitude content. This is where the role of memory can help to complete the explanation of Peculiarity, and thereby resist one style of objection (see Ch. 3).

It has been argued that some Transparency accounts leave a bit of self-knowledge unaccounted for (e.g. Moran 2011; Cassam, forthcoming). For example, Shah and Velleman (2005) suggest that self-knowledge is a matter of attending to our spontaneous responses to brute stimuli. But, as seen in earlier chapters, unless we are willing to accept just about anything that we happen to do,[172] that is, at the time a response might naturally occur, as giving voice to an underlying state, this is not a complete explanation of self-knowledge. A similar objection has been leveled against Byrne's (2005, 2011a) account (see Ch. 5; Casssam, forthcoming). The missing detail can be found, at least partially, in the epistemology of factual memory already outlined.

On either of the two approaches discussed, recalling that *p* is more than simply overhearing oneself (using either one's inner or outer voice) say 'p'. Our recall of *p* is retrieved with a kind of force in each case: on Preservationist accounts, memory preserves the 'probative and motivational force of evidence beyond the point at which that evidence has been forgotten' through a 'belief-fixing influence' (Owens 1999, p. 318); on Conservatism, this force is provided

[172] Moran (2011) uses the example of sneezing.

by a warrant to continue to believe what we already believe. Judging that *p* following the initiation of the self-knowledge procedure is not like *eavesdropping*. At the point that one recalls that *p* one is both committed to *p* and—all being well—entitled to retain the belief that *p*.

This is not a full response to the problem, but it does highlight an important difference between self-knowledge procedures—recalling that *p* is not the same as *deliberation*, nor is it the same as *eavesdropping*—and it goes further than some accounts in explaining why we might think non-deliberative Transparency procedures are legitimate ways of knowing. Two issues await discussion elsewhere. The first is how to make sense of the 'probative and motivational' force of Preservationism; the second is how to make sense of the idea that positive epistemic status can be gleaned simply by believing over time. With regards to the first, the relevant metaphysics of memory is beyond the scope of this thesis. The second will be discussed in the section on Epistemic Security below.

In this section, I suggested that first-person nature of access to one's memory is highly intuitive. However, we may need something to support that intuition if we are to take seriously the suggestion that memory can, at least partially, explain Peculiarity in a theory of self-knowledge. The intuition is best supported by cases such as (RB*p*)—that is, when understood as involving the recall of some episode of believing rather than as factual recall. However, Peculiarity is not fully explained by such cases. (I return to this issue in §4, and Ch. 5.) Less intuitively, cases of (R*p*) can also shed light on how other accounts might explain Peculiarity, but more modestly—that is, by supplementing the explanation of how the Transparency procedure can be regarded a way of knowing. In the next section I move on to Immediacy (I).

3. The Immediacy thesis

In chapter two, I concluded that there are two plausible readings of the Immediacy thesis. The first is a purely psychological thesis; the second an epistemic thesis. Explaining psychological immediacy was retained as a minimum criterion—the commonsense case for psychological immediacy is strong. Epistemic immediacy was retained as an ideal desideratum. The purpose of this section is to show that Psychological Immediacy is explained by the epistemology of memory with which we have been concerned. Although I will not argue directly for the epistemic version of the thesis, I will also outline a way in which we might see self-knowledge in that light.

The central claim of the Psychological Immediacy is that, sometimes, when a subject inquires as to her current state, the result appears without any introspectively detectable train of thought. While the relation between psychology and epistemology is complicated,[173] most plausibly this leaves open the question of whether epistemic inference features in any cognitive transaction between inquiry and result.

Examples of psychologically immediate thought processes are not difficult to find, and many now accept that much of our thinking is unconscious, or fast, or direct (see e.g. Kahneman 2011) in contrast with typical cases of deliberation discussed in this work. In some cases, simply 'sufficiently internalizing' (Wikforss 2004) the most complicated thought processes or theories can grant the appearance of *directness* or *immediacy*. All that is required for these terms to be true in the psychological sense is that whatever cognitive processes that are deployed can be deployed very quickly, or sufficiently beneath the level of conscious accessibility, that their stages are not apparent to the subject. Plausibly this can be true of both a simple pain response or something as complicated as the physicist's 'perception' of sub-atomic particles (see Wikforss 2004).

[173] Like Cassam (forthcoming), I will demure from be drawn into discussing that relation in any depth here on the grounds that a sufficiently treatment of the issue would require more space than can be afforded here.

Meeting this condition is something that a theory of self-knowledge can clearly fail to do (see Ch. 3; Cassam 2014), but once we have rejected the epistemic link between factual memory and memory experiences, it is clear how cases of recall (i.e. as employed in Byrne's account) can begin to meet the condition: in standard cases of factual memory, retaining and recalling belief content does not require that evidence or reasons are retained or recalled (in standard cases one relinquishes them). We might take standard factual recall, then, as a paradigm case of a psychologically immediate operation.

That cannot be all there is to it, of course, because straightforward recall cases—such as proposed in Byrne (2011a)—unlike some cases involving memory experience—need an adjunctive step for self-ascription (see §2). The options most commonly referred to so far in this work are reflection and inference. We can take a closer look at inference by pointing to a 'customary distinction' within 'the sphere of theoretical reasoning' (Boghossian 2014) between two distinct systems:

> *System 1* operates automatically and quickly, with little or no effort and no sense of voluntary control.
>
> *System 2* allocates attention to the effortful mental activities that demand it, including complex computations. The operations of System 2 are often associated with the subjective experience of agency, choice, and concentration.
>
> (Kahneman 2011, pp. 20–21)[174]

The features of *System 2* are broadly comparable with cases of deliberation. To use Evans's example, when considering whether one thinks there will be a third world war—and thereby considers facts that are salient to the question of whether there will be one—one is, at least usually, engaged in an activity that

[174] See also Boghossian (2014). Boghossian points out that a lot of reasoning seems to fall 'between these two extremes', although since the distinction is purely illustrative, here, rigorous treatment will prove a distraction.

consumes a portion of one's conscious attention,[175] and that, at least usually, is accompanied by a sense of agency (that some take to be epistemically significant). Recalling the fact that *p*, by contrast, is standardly a better fit for System 1 thinking: it is either effortless—as with personal details, dates, and common knowledge—or largely involuntary.

Since the involuntariness of memory may not strike everyone one as intuitive it is worth saying a little more. One might fancy that one can initiate a probe the recesses of one's memory, and in a sense this is true—one can initiate the process. However, that is likely where voluntary control ends. When a fact seems temporarily beyond one's grasp, probing is done largely in the spirit of hope. One can probe at related or partial facts in the hope that the misplaced information will spring to mind, but if it does, it does so spontaneously. The most plausible way to explain the phenomenon is 'chaining'. In memory tasks, 'the contents of an involuntary memory sometimes trigger additional involuntary memories' (Mace 2006). Plausibly, one can present oneself with a stimulus for a retrieval attempt by posing myself a question that is either self-directed (Do I think that *p*?) or world-directed (Is it the case that *p*?). But one's control beyond that point is questionable.

So, in standard cases, the first step in the doxastic schema is psychologically immediate, at least insofar as it is broadly aligned with System 1 thinking. (We have good reason to think that it is epistemically immediate too, since recalling that *p* is not a case of inferring that *p*.)[176] The adjunctive step in the procedure—the step that takes a subject from *de re* cogitation to *de se* cogitation—is less straightforward. If the step is inferential, it will not be epistemically immediate. But, on some conceptions of inference, it will not be psychologically immediate either, and so even the non-deliberative Transparency view will not meet a minimum criterion (i.e. the psychological variant of Immediacy).

[175] I argue elsewhere (see Appendix 1) that this does not mean that such thought need always be conscious.
[176] See e.g. Ryle (1949, p. 250)

On some views, inference is a 'person-level, conscious and voluntary, not sub-personal, sub-conscious and automatic' (Boghossian 2014, p. 3), such as in the following example (p. 2):

> (1) It rained last night
> I combine this with my knowledge that
> (2) If it rained last night then the streets are wet.
> To conclude:
> So,
> (3) The streets are wet.

This example sees inference resembling, or sitting between, *System 1* thinking and *System 2* thinking. While it is 'quick, relatively automatic and not particularly demanding on the resources of attention' like *System 1*, it is apparently a 'conscious, voluntary mental action', like *System 2* (p. 2).[177] If this is, broadly speaking, the correct view of inference, then the presence of inference not only rules out epistemic immediacy, but psychological immediacy too. In the next section, I argue that this result is down to some indiscriminate handling of the term inference, and propose a distinction between single-component and dual-component inference as a way to settle the matter with regard to our current inquiry.

3.1 Psychological immediacy and inference

Both 'immediacy' and 'inference' are terms philosophers have felt entitled to use without satisfactorily defining (Boghossian 2014),[178] or at least without garnering a great deal of consensus. With regard to our present inquiry, this leaves us with an unfortunately broad range of views. The views include: (a) inference does not require a conscious element;[179] (b) inference can be quick

[177] It is worth noting that the term 'conscious' has been introduced by Boghossian (2014) here (perhaps as a gloss for 'subjective experience of agency') and does not explicitly feature in Kahneman's (2011) description of the two Systems.
[178] Boghossian's (2014) point refers to inference.
[179] This may be the minority view, but it is not uncommon, and has found apparent adherents in Mill (1882/2009); Byrne (e.g. 2011a); Carruthers (2011), among others. It is the view I am drawn to, although I need not argue for it directly here.

and automatic, even if there is some minimal degree of conscious activity; (c) 'epistemic' inferences (e.g. McNeill 2014) have a conscious element, but inferences in a broader sense need not.[180] The whole self-knowledge procedure (i.e. two steps) can be psychologically immediate on (a) and (b), but will be epistemically mediate. On (c) the procedure can be both psychologically immediate and epistemically immediate.

In its most basic form, inference appears to have a single component. This can be characterized either as (i) a transition that proceeds from a premiss to a conclusion (e.g. Cassam 2007b; McNeill 2014), or (i*) a 'non-accidental transition between belief contents' (Boyle 2011, p. 4). Since the latter would appear to neglect the valid, if uninteresting, inference, 'if *p*, then *p*', I will take the former to be the basic form. We can call this 'single-component' inference. Single-component inference does not capture everything that is intended by 'inference' in some of its uses. One appealing addition might be that the transition occurs because the subject is 'cognizant of other truths as providing justification' for that conclusion (Frege 1979; in Boghossian 2014), or she recognizes the link between the two. Since (i) is still necessary for inference on this picture, we might call the more exacting kind of inference the dual-component view. Single-component inferences are easy to come by: 'classical computers' perform inferences in that sense (McNeill 2014). But since classical computers are not cognizant of 'other truths providing justification for the conclusion' any putative consciousness requirement must lie within the second component (ii).

We should grant that single-component inferences are not restricted to one or more varieties of computer. If we do, then the concern that an inferential approach to self-knowledge is not psychologically immediate rests on the assumption that the kind of inference being performed is of the dual-component kind. But this is neither necessary for the kind of schema being

[180] McNeill (2014) seems to have such a distinction in mind in claiming that computers perform inferences that result in 'non-inferential *warrant*' although he does not expand.

discussed, nor does it appear to be what proponents inferential Transparency views have in mind (see e.g. Byrne 2011a). Indeed, since the pattern of inference is self-verifying but quite bad, it likely to be important for the success of the procedure that the workings of the transition are beyond the subject's conscious reach: [181] if the subject were to recognize the questionable link between the two, she might also recognize that the truth of *p* doesn't even make it likely that she believes that *p* (see Byrne 2005, 2011a).

Provisionally, we can conclude that in the standard (i.e. non-deliberative) case, both one's means of arriving at the judgement *p*, and the transition from that judgement to the self-ascription, can be—and on the basis of what we have said probably are—psychologically immediate. For those drawn to dual-component accounts of inference, it might be remarked that the kind of inference proponents of this view have in mind is not 'epistemic' in their sense. However, this will mean that the method is arguably both inferential (in the single-component sense) *and* epistemically immediate. In the next section, I address this possibility, and argue that we should resist the conclusion.

3.2 Epistemic immediacy and inference

The more substantive version of the Immediacy thesis is epistemological. Accounts that refer to this feature suggest that self-knowledge is distinctive because it is epistemically immediate or non-inferential (see Moran 200;1Ch. 2;). The main claim of this section is that, in the standard non-deliberative case, the first component in the inferential self-knowledge procedure (arguably unlike the deliberative case) is epistemically immediate, and that this is likely to be enough to explain why we take self-knowledge to immediate in that sense. However, this is not enough to suggest that the procedure is epistemically

[181] Byrne (2011a) is mindful that the pattern of inference in question is 'neither deductively valid nor inductively strong' (p. 204); and, in response to questions (at *Varieties of Self-Knowledge Workshop*, Harvard 2016) has indicated that the transition is not one of which the subject is likely to be conscious. Yet, he suggested, there is no question whether it is *inference*, only whether it is knowledge-conducive.

immediate: we must acknowledge the presence of inference on either view presented in the last section.

Evidentialism about memory and some dual-factor accounts of inference share a common flaw. Both portray their respective transactions as involving a subject with substantive and continued access to decisive factors in the adoption of her attitudes. Both recalling and reasoning by inference are seen as involving fully (or substantively) introspectable trains of thought allowing the subject ongoing access to premiss, conclusion, and—on the view of epistemic inference above—some recognition of how the two are linked. This view is optimistic given the breadth of psychological research on subjects' access to their cognitive processes (see Nisbett and Wilson 1977; Appendix 1) and is epistemically implausible at least in the memory case (see e.g. Owens 1999), and probably both: even Descartes and Kant had their doubts about this kind of access to our minds (see Ch. 2).

Rejecting Evidentialism eliminates a mistaken view of factual memory. One does not ordinarily recall one's evidence and arrive (afresh) at a conclusion, nor does one ordinarily recall one's conclusion along with one's evidence, on the chance, for instance, that one is asked to back up one's claim.[182] We can do neither of those things reliably, because we are unlikely to be able to 'dredge up' the evidence more than a 'tiny subset of our beliefs at any one time', and because, usually, we have little reason to retain those reasons.

> Once a question is decided, we close the books on it and throw away the evidence: deliberately retaining evidence for future consultation is a sign of doubt, an attitude appropriate to the scientist who is interested in the likelihood of various things and has a professional obligation to suspend judgement but quite unsuited to the everyday believer. (Owens 1999)

If this view of memory is correct, there is already a good case for saying that the

[182] Although this may be appropriate thing for e.g. a scientist to do (Owens 1999).

primary component of the non-deliberative (mnemic) self-knowledge procedure is immediate in the epistemic sense: when recalling the fact that *p* in response to an appropriate inquiry, we do not infer that *p* on the basis of our reason or evidence, or from anything more 'epistemically basic'. Were it not for the adjunctive self-ascriptive step in the procedure, we could happily take this as a sufficient explanation of epistemic immediacy.

The options with regard to self-ascription from the discussion above are: (i) a jugdement is automatically self-acriptive, as in the case of self-ascriptive memory (i.e. no additional transition is required); (ii) the transition is inferential but not in the weighty sense implied by the two-component view; and (iii) the transition is non-inferential (e.g. it is 'reflective').

Automatic self-ascription (i) in factual memory would require the support of an additional thesis that guaranteed a connection between memory experience and factual memory. A form of (non-doxastic) Evidentialism, for instance, might fulfil that function, but we have ruled it out as implausible. A strong natural correlation between instances of factual memories and their corroborative memory experiences might also fulfil the function, but given what we have said about factual memory, the correlation is unlikely to be strong enough to explain epistemic immediacy.

The second suggestion (ii) trades on differences between notions of inference, and was discussed in the previous section. If epistemic immediacy is simply matter of whether the subject's transition is an inference under one description, then explaining the subject's transition under a different description of inference will be a quick response to the problem. If the critic's concern is that inference is of the two-component variety, then one can suggest that one-component (classical computer) variety of inference does not threaten epistemic immediacy. However, the difference between one-component and two-component views of inference was intended to help us settle the matter of whether a procedure can be both inferential and psychologically immediate. It

was not intended to support the idea that some inferences are epistemic and some are not.[183]

An example of non-inferential transition (iii) can be seen in Boyle's (2011) suggestion that instead of moving (non-accidentally) between belief contents, the subject moves:

> from *believing P* to *reflectively judging* (i.e., consciously thinking to himself): *I believe P*. The step … will not be an inferential transition between *contents,* but a coming to explicit acknowledgement a *condition* of which one is already aware. (Boyle 2011, p. 5)

Boyle's response to the problem grants a version of the one-component view of inference, and so prevents the second option being exploited, unless one is willing to argue that one can infer *p* from *p*.[184] However, if Boyle's reflective transition is conscious, or perhaps voluntary—that is, a natural reading of reflection—then it fails to meet the Psychological Immediacy condition.[185] And while Boyle's suggestion is an ingenious way around some difficulties faced by inferential view, there is no disguising the fact that the subject performs an 'epistemic' transition of some kind (see §3.3).

It has, in fact, been argued that Transparency accounts are not a good fit for any form of immediacy. The concern, broadly speaking, is that the subject needs to reason her way from the judgement that she ought rationally to believe that *p* to the conclusion that she believes that *p* (Cassam 2014, p. 6). To varying degrees, even if the objection is effective against some Transparency views (e.g. Moran 2001)[186], it misses its mark for both Byrne's (e.g. 2005, 2011a) and

[183] Even though this sometimes appears to be the suggestion in some literature (e.g. McNeill 2014).
[184] Inferring *p* from *p* is both possible and valid as noted above. The assumption that we cannot is another indication that our notion of inference still needs some development.
[185] There are a number of objections to the general position are not immediately relevant here.
[186] It is not entirely clear that it is, although this appears to be the kind of approach Cassam (2014) has in mind.

Boyle's (2011) accounts.[187] (Byrne's account, in particular, has a promising explanation of psychological immediacy.) Construed more broadly, however, Cassam's concern can be understood as follows: no matter which way it is dressed, some epistemic transition is required that makes claims to epistemic immediacy difficult to maintain on Transparency approaches. This should be conceded in all of the cases discussed except for self-ascriptive memories. However, there is to my knowledge no Transparency approach that has made use of this explanation.

We can conclude from this section that, of the three options, only one—(i) Automatic self-ascription—is a promising candidate for both psychological and epistemic variants of Immediacy, although there is no suggestion of this possibility in the literature under discussion.[188] Of the remaining two options, (ii) is plausibly psychologically immediate, that is, given a specific view of factual recall, but not epistemically immediate; and although (iii) strictly speaking meets epistemic immediacy by eschewing inference, it is not psychologically immediate (and is potentially neither). The discussion of memory here has left us in no worse a position than before in general, and in a better position with regard to the inferential Transparency view: we now have a more complete explanation of how the account can explain psychological immediacy.

At this point it is noteworthy that something in the structure of the Transparency approach appears to preclude supporting both forms of immediacy at the same time. I return to this issue in Ch. 5. In the next section, I turn to the Epistemic Security condition (E).

[187] Cassam (2014) chooses not interact in detail with either account, although he does respond to Byrne's account elsewhere (forthcoming). His main concerns with the position, however, do not focus on the question of immediacy.

[188] Evans's brief remarks about the subject being 'ipso facto' in a position to assert that she believes that *p* suggest he has something in mind, although this is not fleshed out.

4. Epistemic Security

In this final section of the chapter, I consider whether the epistemology of factual memory might help to explain the Epistemic Security thesis (E). After some brief initial remarks, I consider three ways in which one might conceive of the epistemic security of self-knowledge being explained, at least in part, by factual memory, given our formulation of the criterion in chapter two: (i) a marked increase in the reliability of the procedure in the first-person case; (ii) a marked difference or improvement in justification in the first-person case, and (iii) immunity from error through misidentification (IEM); and (iv) improving the deliberation-resistance of beliefs. I first address some potential initial concerns about an attempt to explain (E) via the epistemology of memory.

The idea of memory providing a first-person epistemic advantage may strike some as odd. A quick look at a list of memory biases might leave one feeling that memory is barely deserving of one's trust, let alone being a plausible candidate for bestowing epistemic advantage. We are susceptible to false memory implantation (Loftus 1975); have a tendency to believe events are more predictable once we know about them—that we 'knew it all along' (Roese and Vohs 2012); that repeated statements are more likely to be true (Hasher, Goldstein and Toppino 1977); that the past was rosier than it was (Mitchell and Thompson 1994); and that the choices we made are better than those we did not make (Mather and Johnson 2000). On top of this, we systematically compress or expand time (e.g. Burt, Kemp, and Conway 2001), and rearrange the order of events (2003). The list of biases that effect our ability to recall the events of the past is extensive (see Ch. 1). Some take this as a sign that there are no accurate memories at all.[189] And if memory is epistemically flawed in this fundamental respect, then there is little hope that it could help us to explain any first-person privilege.

[189] Martin Conway argued for the claim at the *Epistemic Innocence* Conference, University of Birmingham (2013).

There are a number of things to say against this worry. The first is that the general reliability of memory is crucial for the possibility of (memory) knowledge (Senor 2014).[190] And even if we wanted to adopt a neutral position with regards to assessing the reliability of memory, an investigation into its reliability 'would have to make use of beliefs about the past':

> Our memory is not one more informational device which we can use or not as we please: it is fundamental to all cognitive transaction, including any that would be involved in establishing the reliability of memory … An agnostic about memory could not even begin to determine which of his memories he should accept and which he should suspect. (Owens 1999, p. 313f.)

Empirical investigations of the variety that provide fascinating and useful information about our ability (or otherwise) to recall accurate information in certain contexts will not inform us on the question of the memory's general reliability (Senor 2014).

Secondly, imbalances in retention and recall, such as those listed above, are not all of the same kind, and do not all go against the subject in terms of reliability. From the sample list above, the *knew-it-all-along* effect (hindsight bias) looks plainly confabulatory, whereas the tendency to think the choices we make are better than rejected choices (choice-supportive bias) looks, potentially, to have an important epistemic role on top of its pragmatic benefits: it is difficult to see how one could stand in the right kind of relation to one's decisions if one thought from the outset that the matter of whether they were correct had not already been settled. Some memory effects appear to place the subject in a position where information encoded as relevant to that subject is more readily available. This kind of memory effect is worthy of attention when considering whether memory has anything to contribute when it comes to

[190] The term 'memory knowledge' (used here in Senor 2014), like 'memory belief' is misleading. Memory is not a 'source of learning, discovering, establishing' (Ryle 1949, p. 249).

explaining perceived asymmetries between first-person and third-person judgements.

Finally, throughout the preceding chapters we have come across a number of ways in which the epistemology of memory contributes to an explanation of an epistemic feature of self-knowledge: it helped to complete the explanation of how a non-deliberative Transparency procedure can fulfill an implication of Epistemic Security (Ch. 3); it explains how a non-deliberative Transparency procedure might be in part epistemically immediate (Ch. 3, and above); and on some views, memory appears to bestow a kind of warrant (see above). In short, there are a number of options worthy of consideration. In the next section, I consider whether epistemically beneficial memory effects might be among them.

4.1 Epistemically beneficial memory effects

One way of thinking about improving epistemic security is in terms of whether some aspect or function of memory makes first-person judgements more reliable. In this section I explore two candidate memory effects that appear to improve reliability in recall for memory content that has been encoded as relevant to the subject (although a number of effects appear to be relevant to questions in this area).[191] These are the *self-referencing effect* (SRE) and the *self-generation effect* (SGE).

The self-referencing effect' (SRE)—or family of related effects (Klein 2012)—suggests that 'relating information to oneself is a successful encoding strategy' (Gutchess et al. 2007). Judgements that we judge to be related to ourselves see 'increased levels of memory compared to making semantic judgements or relating the information to another person such as one's mother or Johnny Carson' (p. 822). Researchers dispute over which model best explains the effect, the degree to which the effect is robust, and whether it is a

unitary phenomenon. However, although findings across the literature vary, a meta-analysis (Symons and Johnson 1997) concluded that 'the SRE does occur with highly significant reliability', and that 'SR was superior to semantic and OR [i.e. other referencing] encoding in facilitating memory' (p. 386).[192]

The generation effect, or self-generation effect (SGE), suggests that memory of self-generated content is *better* than externally generated content (Slamecka and Graf 1978), so that, for instance, words that are generated by a subject are 'better remembered' (p. 592) than the same words when the subjects read them. The effect is a robust over a variety of conditions and is related to research concluding that memory of problems that are solved by a subject is better than when a subject *merely* remembers the solution (Jacoby 1978).

The upshot of the two effects is that subjects, as a rule, are (a) better at retaining and recalling information that is relevant to themselves than either information that is relevant to others, or that is person-neutral, and (b) are better at retaining and recalling information that they have generated themselves over information that they have come across (e.g. read). One might be inclined to say on the basis of these results that subjects have a clear advantage when it comes to the retention and retrieval of information deemed to be about, or produced by themselves. And this looks like an initially promising response to the epistemic security thesis in two respects: (i) it can plausibly account for why judgements about ourselves are more reliable; and (ii) it can plausibly account for why we favour cases in which we have made up our minds (see Ch. 3).

Two related concerns are (1) that there may be just as many imbalances that negatively affect a subject's ability to acquire knowledge about her mental states, and (2) given this possibility how can we demonstrate to the Parity theorist that knowledge is easier to come by in the first-person case without entering into an empirical dispute. One may respond to the first worry by a producing a taxonomy of retention and recall imbalances that would allow for

[192] That is, in the literature reviewed for the analysis.

an analysis of which positively or negatively affect memory that relates to the self. However, it would be open to competing conceptual analyses, and since some cognitive biases are explicitly memory related while others are implicitly memory related, it would be a complicated undertaking, and certainly cannot be attempted here. It is also, rightly or wrongly, unlikely to satisfy many philosophers (in particularly the notional Parity theorist). While such an exercise is promising, then, we should see if there is anything in the philosophical treatment of memory that allows for an understanding of how self-knowledge could be epistemically privileged.[193] In the next section, I will consider whether the positive epistemic status bestowed on some the epistemologies of memory with which we have been concerned, might be used to explain the feature.

4.2 Epistemic security and positive epistemic status

One way to conceive of what is epistemically distinctive about self-knowledge, concerns the strength and/or source of justification (see Ch. 2). Jordi Fernández (2013) suggests that both 'the source of my justification for my beliefs about … [mental] states' and 'its strengths or robustness' need explaining (Fernández 2013, p. 4): in *Special Access*, first-personal justification 'relies on neither reasoning nor behavioural evidence' (p. 5); and in *Strong Access*, first-personal justification is 'stronger' (p. 6). This is not the correct way to characterize Epistemic Security (see Ch. 2). However, focus in epistemology has turned to epistemic warrants and entitlements, rather than only *justification*, and the view is not incommensurate with the suggestion that a subject's beliefs about her own mental states are more likely to result in knowledge than her beliefs about the mental states of others (see Ch. 2), as long as the notion of justification is taken broadly, for instance, to mean something like 'positive epistemic status'.

[193] We have, of course, covered in some depth an attempt to explain (E) in terms of reliability—namely in Byrne's (2005. 2011a) account. However, the mechanism responsible for increased reliability is the self-verifying nature of the schema.

The main aim of this section is to show that—on the assumption that one (or both) views of memory that meet the features outlined at the beginning of this chapter is correct—the epistemology of memory can explain how first-person *justification* is both 'stronger' and arises from a different source. The view of factual memory most likely to fulfill this condition is Conservatism. While it is a plausible view, some of its consequences are considered unusual, and so its merits require emphasis. One way to do that is to show that moderate Preservationism and moderate Conservatism are closer than some literature suggests.

One might see belief as an 'epistemic commitment' (like a promise or contract) in that once the commitment has been made it provides a reason to do what has been promised 'over and above any which might have led [one] to take on that commitment in the first place' (Owens 1999). However, the analogous view of practical decision-making looks false when it comes to 'Auto-promising' (pp. 320f.). If one decides to do something that one later realizes is a silly idea, one's having decided to do it gives one no special reason to persist (p. 321). And so it is with belief. Because warrant bestowed by memory on the conservative's view is a prima facie warrant, the objection does not quite hold. If information comes to light that throws a subject's belief into doubt, that prima facie warrant is defeated. Even if the objection did work, one need not view Conservatism that way.

One might try to understand it, for instance, by appeal to the 'Reidian principle that we are prima facie entitled to presume, of any belief, that it is well-formed and well-maintained and so is worthy of trust' (McGrath 2007, p. 16). On this view, Conservatism offers a plausible response to a question that Preservationists (and Evidentialists)[194] find difficult to answer: Is it rational for me to retain a belief when it was poorly formed, but my reasons for believing it is poorly formed are no longer available?

[194] Both doxastic and non-doxastic forms of Evidentialism. For a detailed discussion see Owens (1999) and McGrath (2007)

On the Preservationist view, memory preserves belief content along with the 'probative and motivational' force of the evidence (see Owens 1999), thus allowing the evidence itself to be relinquished. This means that one is rational to believe that in the absence of defeating conditions (such as evidence against *p*) just as long as one was rational to acquire the belief that *p* in the first place. The difficulty is that once we are entitled to relinquish the reasons for our beliefs, the evidence that would usually allow one to weed out one's past errors is no longer available.

Thus, even in cases where one's 'present self' is particularly careful, one will often have no basis on which to doubt one's retained beliefs. On the face of it, the Preservationist answers 'no' to our question, since memory only preserves one's past rationality. However, since one has no (subjectively available) grounds on which to abandon the belief, this is an unintuitive response: either one is somehow rational to retain the belief in spite of its questionable provenance, or one must be willing to abandon great swathes of historical 'knowledge' (see McGrath 2007).

One response on behalf of Preservationism is to deny that retaining the belief in this case is rational but admit that it must be, *blameless*. However, because this blamelessness is *epistemic*,[195] the response still amounts to the conferral of positive epistemic status upon a belief that one ought not rationally to have adopted (see McGrath 2007, pp. 5f.). The result is surprising but plausible in the face of alternatives, and some versions of Preservationism concede the result, at least in a restricted class of cases (see Owens 1999, p. 322).[196]

If this is correct, as I have suggested, then we have a potential, and independently plausible, way in which memory can confer positive epistemic

[195] It is 'not moral or prudential' (McGrath 2007, p. 6).

[196] 'if it is reasonable for me not to reconsider my belief in Hitler's suicide, I can't be irrational in continuing to believe it, though the belief itself may be quite irrational' (p. 322). Owens's suggests the result is not peculiar to memory, but the general point that 'a subject may be entitled to think that a belief can be justified when, in fact, it cannot' (p. 325), although the Reidian principle (above) provides a more complete explanation for the 'epistemic luck' and makes memory the standard case (*Ibid*.).

status, and with it a possible response to a *Special Access* view: the source of epistemic status is a different source, for example, to the original evidence. Such positive status need not be conferred in all cases of belief, and the defeating conditions can ensure that it is not (see McGrath 2007, p. 17–19). But if we are willing to accept Conservatism, then a kind of positive epistemic status can be conferred, in appropriate cases, that is, provided in cases where beliefs are retained. In other words, it is a reply to the Special Access that is plausible in the absence of a negating argument. (How one might fit this into a specific theory, or theories, of self-knowledge is a matter for elsewhere.)

Conservatism holds that believing—*sans* defeating conditions—confers rationality. So we might ask how the positive epistemic status conferred by believing interacts with any residual positive epistemic status bestowed by retained evidence. One concern is that the former combines with the latter (McGrath 2007, p. 20) effectively awarding retained beliefs 'an extra point' (Owens 1999, p. 321). This concern is the 'extra boost' objection to Conservatism. One response to the objections is to suggest that defeating conditions for the 'bonus' epistemic status include, roughly,[197] that the conditions that bring about the conferral of a positive epistemic boost include a 'lack of special information about her past evidence', and since in 'ordinary mature adults, long-stored memory beliefs comprise the bulk of beliefs', and would appear to meet this condition, we might take Conservatism to be a good epistemology of memory (McGrath 2007, p. 21f.).

There are, however, reasons to reject this response to the epistemic boost objection if one takes seriously the dynamics of memory and the importance of the first-person perspective in rationality (McGrath 2007). For at times one forgets and later recalls the evidence for one's beliefs, and there are no strict limits on how many times one can oscillate between the two. Far from making the Conservatism more plausible, having memory confer and withdraw an epistemic status as bespoke to the demands of individual absent-mindedness,

[197] McGrath offers this as an incomplete suggestion (2007, p. 21).

would make memory an odd thing indeed. There is another way. In other cases, such as testimony (on some views), one may obtain prima facie epistemic entitlement for believing that *p* just because one is provided with that information in the absence of defeating conditions. If, in addition, one is provided with a subject's evidence in support of *p*, plausibly one has an additional warrant: firstly (i) testimony; secondly (ii) reasons that directly favour *p* (see McGrath 2007, p. 19). We can continue to add sources of warrant—perhaps one now happens upon (iii) perceptual evidence in addition to (i) and (ii)—such that the rationality of the belief is multiply-determined.[198] Multiple-determination is not abnormal. As long as the reasons in favour or believing that *p* (regardless of their source) are not incompatible, the *justification*-strengthening property prima facie warrant in the memory case is no more concerning than if one were to both hear second-hand and see first-hand that something is the case. Neither way offers a complete and robust defence of Conservatism, but they do show that a number of options are available to diminish concerns about its apparent consequences. If we are looking for a modest epistemic contribution, then Conservatism arguably provides plausible answers to both *Special Access* and *Strong Access*.

4.3 Immunity from error

A third route has been touched upon with regards to memory experiences (above). Some memories appear to be 'self-specifying' and are thus arguably 'immune to error through misidentification relative to the first person pronoun' (IEM). Judgements that are IEM are meant to exhibit the following 'property'—they do not: 'involve a identifying a particular person as oneself because the sources of information on which they are based are such that they can only provide information about oneself' (Bermúdez, forthcoming). Possession of the property is one way that first-personal can be epistemically

[198] McGrath (2007) suggests that over-determination might solve the problem in another way—namely, that additional sources of warrant do not provide an epistemic boost. Either way, it can be agreed that 'the mathematics of rational entitlement is not … simple' (p. 20).

privileged or distinct, and is therefore one way in which we might consider attempting to explain Epistemic Security. The IEM property is usually associated with 'introspection and somatic proprioception' (Bermúdez 2013, p. 212), since the information from these sources typically 'could not but be about the thinker' and hence such judgements are identification-free' (*Ibid.*) However, it has been argued that past-tense judgements based on autobiographical memory possess (IEM) (Bermúdez 2012), or at least some of them (2013, forthcoming). The obvious cases of memory judgements that possess the property IEM are those that are 'explicitly' self-specifying cases (Bermúdez, forthcoming), such as 'I remember skiing'. These are of interest, but of limited use to the current inquiry: (i) we don't often retain the evidence for a belief (in this case acquaintance with an event), and (ii) such judgements will not meet the Transparency desideratum. Things can be improved with regards to (ii) if non-explicitly self-specifying can possess the IEM property. It has been argued (Bermúdez 2013, forthcoming) that that they can. If things can be improved with regard to the first concern (i), then the IEM property offers a substantive explanation of Epistemic Security, as well as a potential solution to some of the difficulties posed by other Transparency methods with regards to Immediacy. (I return to this issue in the next chapter.)

4.4 Deliberation-resistant attitudes

As we saw in chapter three, one of the main concerns about the inferential transparency view is that the procedure is only a good indicator of one's mental states at the moment one completes one's attempt to reason in accord with the doxastic schema (see Ch. 3; Gertler 2011a). While I argued that the epistemology of memory is crucial to the success of that approach, I did not show how the correct epistemology might make a distinctly positive contribution to our epistemic situation. I will now describe how this is possible.

The plausible solution to Gertler's (2011a) concern has two parts: firstly, that the standard way to deploy the doxastic schema is in its non-

deliberative (mnemic) form; and secondly, that a retained beliefs act as a block on further deliberation. These two factors offer an explanation of how the procedure can be a reliable way to assay one's beliefs at a time other than the completion of the procedure. This explains how the self-knowledge procedure can be a reliable. It does not, however, explain how ensuring that one has the right epistemology of memory in place might positively contribute to one's epistemic situation (i.e. over and above completing an explanation of the how procedure could work).

One way of reframing the issue is this. Assume that the doxastic schema is, descriptively, our main method of knowing our minds. Since its self-verifying nature leaves the subject almost infallible with regards to her beliefs, how could two subjects relying upon this method differ? [199] This is where one Preservationist approach to memory can help.

Firstly, for any retained belief (B*p*) the degree to which a subject is rational in holding that belief can vary. On Preservationism, memory preserves both the content of that belief and its rational force, such that if a subject S believes that *p* on the basis of evidence F, the *force* of F, but not F itself is carried with the belief. If the evidence is poor, then the force of F will be weak, and vice versa. David Owens describes the feature in terms of 'cognitive inertia':

> The cognitive inertia of belief is a corollary of the rationality-preserving nature of memory. Where belief is rational, the inertial force of the belief is determined by the strength of the reasons which supported its adoption. (Owens 1999, p. 323)

Since belief is a block to further deliberation, in order for deliberation to pose a challenge to the reliability of a procedure, there must be grounds for doubt. And, plausibly, the relation between cognitive inertia and grounds for doubt is proportionate—that is, the greater the cognitive inertia, the greater the grounds

[199] Thanks to Tom Stoneham and Keith Allen for pressing me on this issue.

for doubt required to re-open the inquiry into truth. And so, one way to improve the reliability of the method with regards to assaying one's state prior to the initiation of the procedure is to ensure that the grounds for one's beliefs are strong. Thus, the more careful one is when initially making up one's mind, the more reliable the doxastic schema will be as a procedure for detecting one's mental states prior to the onset of the self-knowledge procedure.

One might object that cognitive inertia does not always originate from careful reasoning.[200] A subject might be prone to quick, confident decisions, or slow and thorough, but in both cases one's processes could be badly affected by unconscious bias. In such cases, even attempting to double check one's steps may not help improve one's first-order belief formation (see Kornblith 2011). However, such processes are still likely to improve the belief-detecting capacity of the procedure, by making it more likely that the belief in place prior to the initiation of the self-knowledge procedure has sufficient inertia to resist deliberative contamination.

Conclusion

In this chapter, I have addressed the question of whether any light can be shed upon the purportedly distinctive features of self-knowledge by three common models of factual memory. The peculiarity of first-person access to memory is highly intuitive upon any of these three models, and particularly plausible upon views that allow a subject to rationally retain a belief without evidence. Immediacy is also plausible on such views. Perhaps most surprisingly, on all three models of memory, there are options to be explored with regards to improving one's position, epistemically speaking. If one is willing to accept Evidentialism, then it is possible that one could be substantially immune to a certain kind of error; if one is a Conservative (and on some forms of

[200] 'where a belief is irrational [cognitive inertia] is determined by some other factor. Either way, a belief, once acquired, constitutes a psychological obstacle to its own revision' (Owens 1999, p. 323)

Preservationism), memory is the source of a distinctive positive epistemic status; and on Preservationism, one can improve the reliability of the doxastic schema with regards to one's ability to assay beliefs one has in place prior to the initiation of the self-knowledge procedure.

The result of the inquiry into whether the epistemology of factual memory can be thought to explain what we take to be special about self-knowledge has revealed a surprisingly broad range of options even in the most difficult cases. Although I have illustrated how some of these features may fit into theories of self-knowledge, I have not argued specifically for a theory based upon a model of memory. In the next chapter I make use of some of these features discussed in this chapter to demonstrate that a fairly standard case of doxastic self-knowledge can suitably be described as kind of recollection. I then outline a theory of doxastic self-knowledge based on this case, which can be set against the desiderata from chapter two.

5

Doxastic Recollection

Introduction

In chapter one, I argued that our thinking about introspective failure and memory failure converges both in non-theoretical and theoretical contexts. This, I suggested, was good reason to consider whether our thinking about introspective success and memory success also converge. I suggested that a positive answer to that question would likely require a positive response to a question of whether memory can help to explain what we take to be special about self-knowledge, and accordingly in chapter two I examined the plausible desiderata for any theory of self-knowledge. In chapter three I argued that the success of a prominent approach to explaining self-knowledge—the Transparency approach—is, in part, dependent upon a specific epistemology of memory. Since the epistemology of memory was found to be part of the explanation of why the Transparency approach—or a specific version of it—fulfils some of main desiderata for a theory of self-knowledge, I explored, in chapter four, the extent to which the epistemology memory might be taken to explain the main desiderata for a theory of self-knowledge more generally. The result of this inquiry was a surprisingly broad range of options. Even in the most intuitively difficult cases, such as Epistemic Security, the options are considerable. What remains of the investigation is to see whether some of these findings can be brought together in a theory that might be set against the desiderata outlined in chapter two.

The discussion so far has been heavily weighted towards the Transparency approach to self-knowledge. While there is much in this approach that is valuable, its use as an example can mute the potential contribution of

memory in the following respect: there are at least partial explanations for a number of our stated desiderata already available within the structures of individual Transparency accounts. So, for example, Epistemic Security and Peculiarity are meant to be explained by the world-to-mind inference on (a version of) what I have called the inferential Transparency view; and Epistemic Immediacy is intended to be explained by 'reflection' on (a version of) what I have called the deliberative Transparency view. While there may be a good case for the contributory role for memory in some of these accounts (see e.g. Chs. 3–4), merely filling out some missing detail in other accounts diminishes the potential to highlight how explanatorily apt the epistemology of memory can be for our thinking in this domain.

In this final chapter, I consider the degree to which it is plausible to see self-knowledge, in the case of belief, as a *kind of recollection*. Key to this will be presenting a standard case of doxastic self-knowledge that is plausibly describable in terms of recollection. Once I have presented this case, I argue that a theory of doxastic self-knowledge with the main explanatory work being done by an appropriate epistemology of memory can fulfil not only the main criteria, as discussed in chapter four, but a good range of secondary criteria (i.e. ideal and additional desiderata).

I first highlight a problem of self-ascription in literature on the Transparency approach that has been a motivating force behind a number of Transparency accounts (§1). I frame this in terms of an assumption that ought to be questioned. While it ought to be questioned, doing so here will also serve the additional purpose of bringing into sharp relief the explanatory aptness of memory in this domain. In §2, I argue for weaker characterisation of the self-ascriptive requirement of Transparency accounts of self-knowledge. In §3, I offer a sketch of a theory that can be placed against the explanatory desiderata discussed in chapter two. In §4, I measure the success of the theory against a range of desiderata. Finally, in §5, I consider some possible objections to the proposal and outline some the limits of the approach.

1. The Transparency–Transition problem

It has become orthodox in literature on the Transparency approach to suppose that something important is missing from Evans's (1982) brief remarks about self-knowledge (see e.g. Byrne 2005, 2011a; Boyle 2011; Cassam, forthcoming). In chapter one, I suggested that, at base, the ingredients of any account of introspection must include something like the capacity for correct, usually time-sensitive, attribution of some fact of the matter about oneself, to oneself. Evans's remarks suggest that all there is to trying to assay whether one believes that *p*, is to consider whether *p* is the case (p. 225). And so, what is sometimes thought to be missing in Evans's remarks is a means of getting from one conclusion—namely a conclusion about the 'world at large' (Boyle 2009)—to a conclusion about oneself. In other words, an independent act of self-attribution.

The worry, I take it, stems from a legitimate puzzle about Transparency, here expressed by Richard Moran (2003):

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

But this puzzle and the recent orthodoxy come apart in several important respects. In order to see how, it will be helpful to rehearse some of Evans's (1982) remarks on the matter:

> If someone askes me 'Do you think there will be a third world war?,' I must attend in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question 'Will there be a third world war?' (Evans 1982, p. 225)

Let us say that, following Evans's guidance, one is asked what one thinks, and attends to the appropriate outward phenomenon. On the basis of a series of unsettling events around the globe, one concludes that there will indeed be a third world war. Since the conclusion that there will be a third world war is not a conclusion about oneself, it might be tempting to think that it is no longer obvious, after all, that this is procedure 'automatically' places one in a position to make an assertion about one's belief. One may be further tempted to speculate about what the missing step might be that takes a subject from the former conclusion to the latter (cf. Byrne 2011a, p. 203). At this point one has, to a greater degree, restricted the kinds of answers that are available.

The problem, I think, must be based on hyperbole. For the picture suggests that the subject can attend to the relevant task competently, and still find herself in a position in which her conclusion has no obvious—that is, conscious first-person—connection to what she thinks. The transition that occurs after the conclusion about the world, then, can be described as an additional inferential step, a non-inferential step, or left a mystery. Since we are unlikely to accept a mystery, we are left with two approaches to rescuing the stranded subject. It is worth outlining some of the outstanding issues with both of these approaches.[201]

1.1 Inference and reflection

One response to this Transparency puzzle is to suggest that the missing step is a non-accidental shift between the belief contents, '*p*' and 'I believe that *p*' (Boyle 2011, p. 4). That is, it is an 'inference from world to mind' (Byrne 2011a, p. 203) on an minimal view of inference (see Ch. 4).[202] If one takes such an inference to present a 'serious problem' (Byrne 2011a), to be 'mad' (Boyle

[201] There are varieties of both, so I will stick to the examples that have formed the basis of the discussion in previous chapters.

[202] 'I infer that I believe that there will be a third world war from the single premiss that there will be one' (Byrne 2011, p. 208)

2011), or to be 'patently invalid' (Cassam, forthcoming), one might either seek to refine the inferential process, or look for a step that is non-inferential. One way to refine the process is to have the inference mediated by psychological evidence (e.g. Cassam, forthcoming), and one approach to the non-inferential solution is to suggest that the subject engages in 'reflection'—that is, she does not 'transition between contents' but comes 'to explicit acknowledgement of a condition of which' she is 'already tacitly aware' (Boyle 2011, p. 5).[203] But it looks like both of these solutions leave us with ways in which we either *already* know what we believe rather than 'detect' what we believe, or they fail to complete the explanation of what is missing.

Among a number of concerns about the world-to-mind inference is whether this kind of inferential approach could yield knowledge. Broadly, the concern is whether the procedure is conducive to knowledge (cf. Byrne 2011a; see also Ch. 3). Knowledge-conduciveness on the view is, nonetheless, sketched as arising from two features of the procedure: (i) 'inference from a premiss entails belief in that premiss' (Byrne 2011a, p. 206) and (ii) the procedure typically 'yields beliefs that are *safe* in the sense that they could not easily have been false' (p. 206).[204] However, one might still object to (i) in the following way:

> Logic teachers run thousands of sample inferences from premises that neither they nor anyone else in their right mind actually believes. In *reductio* arguments, one supposes that *p,* infers *q* from *p*, and then infers the falsity of *p* from the falsity or absurdity of *q*. There is obviously no question here of inference from a premiss entailing belief in that premiss (Cassam, forthcoming, p. 7).

[203] In Ch. 4 I say that Boyle's (2011) description of the minimal view of inference is not quite right.
[204] Byrne (2011a) acknowledges that the two features do not amount to a '*demonstration* that reasoning in accord with the doxastic is knowledge conducive' (p. 207).

This kind of objection sees Byrne's (2005, 2011a) inferential approach facing a difficulty similar to that already addressed with regard to Shah and Velleman's (2005) Neo-Expressivism (e.g. Ch. 3). The objection in that case was that many things that occur following a stimulus, including a variety of spontaneous vocal emanations, are not good indicators of what one thinks, and so merely attending to these cannot be all there is to knowing one's mind (see Moran 2011; also Chs 1, 3). The corresponding difficulty for the Byrne's inferential account is meant to be that thinking '*p*' is not a good reason to believe that one believes *p* either; that is, unless one 'already knows' that '*p*' 'expresses' one's belief (Cassam, forthcoming); there may be many times when simply thinking '*p*' is not a good indicator of what one believes.

The objection has limited force, because reasoning in *reductio* cases sees '*p*' as hypothetical, and reasoning hypothetically from '*p*' is not a standard case of concluding that *p*: the epistemic context provides one with a reason not to believe that that *p*.[205] Thus, the proponent of the (non-mediated) inferential view is able to respond that reasoning hypothetically about '*p*' is not a case of reasoning in accord with the self-knowledge procedure (in this case, the doxastic schema).[206] However, we might take the spirit of the objection to be this: if one is meant to go directly from reasoning about '*p*', to concluding that one believes that *p,* then, the explanation of knowledge looks incomplete. A different way to flesh out this concern is to ask—that is, at least in the deliberation case but plausibly in the non-deliberative case too—how it can be that one is able to detect *when* and/or *whether* one has reached the relevant conclusion, to thereby enable an inference to the second-order belief to take place.

[205] Cf. Gertler's (2011a) summary of the Transparency procedure in Ch. 2.

[206] Markos Valaris (2011) argues that Byrne's account does not allow for hypothetical reasoning, although it is not clear why in reasoning hypothetically one would attempt to follow the same pattern of inference that one would follow in self-knowledge case. A recognition of the kind of task at hand appears to be a pre-requisite of either both deliberative and non-deliberative versions of the schema, and there is no obvious risk of forming the belief that *p* when one has deliberately set '*p*' out to be the kind of thing one is not supposed to believe in the present case. (Thanks to Professor Stoneham for a helpful discussion about these cases.)

If one is persuaded by the a line of thought, one might try completing the explanation by making the inference 'mediated' (Casssam, forthcoming) by conscious 'mental occurrences' such as, 'judgements, feelings of conviction and experiences of agreeing'.[207] These occurrences are meant to act as evidence that one believes that *p*.

There are a number of things to note about this way of addressing the issue. Firstly, unlike Byrne's (2011a) inferential model, the evidence-mediated inference is not self-verifying: feelings of conviction (etc.) about *p* do not entail that one believes that *p*, and so the procedure will not be guaranteed to produce true second-order beliefs. Secondly, because the doxastic schema explains Epistemic Security in terms of the strongly self-verifying nature of the procedure, the evidence-mediated view of inference will leave us without an explanation of that feature.[208] Thirdly, it is not clear how such an adapted account would account for the Transparency of first-person thinking, since the source of the inference is no longer a judgement about the world, as such, but ultimately a judgment about one's 'feelings'.[209] Fourthly, our access to the relevant mental occurrences—and especially judgements—is assumed rather than explained.[210]

On this fourth point, the options available for explanation include: (i) that knowledge of one's own judgements is 'unmediated, direct knowledge', and (ii) that 'knowledge of one's own judgements is also inferential', perhaps based on a 'sense of cognitive ease or settledness' (p. 11). But neither of these options immediately answers the question of how such an account would qualify as a way of knowing, since adding 'direct' to 'knowledge' alone contributes nothing to an explanation of how S knows that *p*; and it is not clear how a sense

[207] This is what Cassam (forthcoming) calls the 'mediated inference model' (MIM).
[208] This is fine as long as one adheres to the relevant arm of the Parity Thesis, although this has its own challenges (see Ch. 2).
[209] Cassam (e.g. 2014) has argued extensively against the Transparency approach to self-knowledge, but a number of the arguments are not affective against developments of the approach such as Byrne (2011a) and Boyle (2011), and do not directly tackle the phenomenon of Transparency that I have retained here as a desideratum.
[210] Cassam (forthcoming) recognises that the plausibility of the account will rely upon the account of knowledge of judgements.

of cognitive ease can be regarded as the grounds for an inference to the conclusion that one believes[211] (especially, if concluding that *p* cannot do that work). The mediated-inference model, then, also faces the familiar objection—namely, that listening to one's thoughts and feelings—that is, without already knowing that they express one's views—is a questionable, rather than sufficient, explanation of knowledge.

At this point we might begin to notice a pattern. Once one accepts that Evans's (1982) routine leaves the subject stranded with a first-personally irrelevant belief or judgement '*p*', and accepts that the self-ascriptive act is an inferential one, then one faces a choice: either one must accept that arriving at the conclusion involves a kind of self-knowledge for which one has not accounted,[212] or one must prevent a regress by reference to direct knowledge or a psychological state. Neither response looks wholly convincing.

One might instead turn to the reflective approach. But this approach faces a similar difficulty. On this view, the relevant state is 'knowingly believing', and the self-knowledge procedure is not a detection mechanism for one's belief, but a way to make one's tacit knowledge explicit (see Boyle 2011). Because the reflective procedure is not intended to reveal anything to the subject she does not already know, it can be objected there is a 'sense in which subjects *are* omniscient about their own beliefs' (Casssam, forthcoming, p. 14). Since explaining how one finds out what one believes (etc.) was meant to be the business of a theory, Boyle's elegant (2011) proposal seems to miss the point. And since pointing to omniscience (alone) is an insufficient explanation of knowledge, this approach too leaves knowledge less-than-fully explained. As

[211] Cassam's (forthcoming) point is that judgement is conceptually linked to belief. However, it is difficult to see how a cognitive sense of ease could play the same role, that is, if it is to be the basis of judgement. There is no obvious conceptual link between 'ease' and 'judgement'. One may have a sense of cognitive ease when it comes to simple explanations that one knows to be false, for example. One may have been taught the Rutherforf-Bohr model of the atom, in one's formative years, and find it 'easy' think about atoms in that way, while all the while knowing that it violates the Uncertainty Principle, and therefore cannot be the correct model. Having this latter information does not prevent one from feeling the ease and familiarity of representing atoms as tiny solar systems. Contrariwise, believing that Utilitarianism is the correct moral stance may well leave one feeling 'uneasy' in many day-to-day situations. Knowing that one believes is more than feeling easy or settled.

[212] Both Cassam (2014) and Stoneham (private correspondence) have raised similar points.

Ryle (1949) notes: 'becoming conscious' of something is not obviously the kind of thing one can use as a 'final appeal' when answering questions about how one knows (p. 143).

The two options, then, have a somewhat limited explanatory appeal (since we have adopted an explanatory view of knowledge, this is a problem). Either one accepts that the subject is already in possession of some relevant knowledge, and awaits an explanation of *that* knowledge, or one accepts that one has not yet quite explained the subject's knowledge. Neither option is terribly satisfying, and I do not aim, here, to offer a new defence of either approach.[213] In the next section, instead, I make an effort to clarify the problem by pointing to some unusual features of what we might now call the Transparency–Transition assumption.

1.2 The Transparency–Transition assumption

In the last section I highlighted common obstacle in Transparency accounts of self-knowledge. I mentioned two attempts to respond to a Transparency puzzle that run into the similar difficulties. Part of the explanation for this difficulty is that both approaches appear to subscribe to the same assumption. On both accounts, the Transparency procedure involves two steps—first coming to a conclusion about the world and then activating a mechanism that allows one to move from that thought, to a subsequent thought about oneself. The unusual conceit is that without this latter step the subject has no way of being aware that the thought she has arrived says something about *her*. While this way of thinking about Transparent self-knowledge has clearly been seductive, it is in a number of cases misleading, and should not go unchallenged. Let us try to capture the assumption:

[213] Works are forthcoming from the main proponents of both positions that will no doubt contain helpful clarification on these issues.

(TT) A subject engaged in Transparent cogitation cannot correctly attribute the output of that cogitation (t_1), without a subsequent ascriptive transaction (t_2)

In self-knowledge cases, of course, the subsequent attribution is self-directed. So the (TT) assumption sees the subject engaged in a two-step procedure that always proceeds in the same manner.

The assumption is unusual in a number of respects. It casts the subject as engaged with and responding to a query *about herself*, and coming up with an answer that she cannot recognise as an answer that refers to herself without the additional step. On the assumption that the notions of reflection and inference are more or less standard,[214] it should be relatively unremarkable when the two elements of the procedure come apart, such that for instance the procedure is harmlessly curtailed. Unless the combination of the two elements in this case is irresistible—a matter that would need further explanation—there is no reason to suppose that a curtailment of the process would come across as especially odd.[215] Here we can make use of an example:

Petra asks Preeti whether she wants a slice of cake. Preeti considers whether cake is a desirable thing, and makes a call (t_1). At this point Preeti's thought process is curtailed, that is, prior to the completion of stage two (t_2).[216]

What is the best way to describe Preeti's state of mind now? How will she respond to Petra's question? On Byrne's account, Preeti will think some thought along these lines: 'Cake [or that slice of cake] is desirable', but cannot

[214] Although, please see note my earlier concerns on these matters (Chs 2,4).
[215] That is, unless the combination in the first-person case is irresistible, in which case this is another feature of self-knowledge in need of explanation.
[216] It has been noted that the desire case is unusual in that one can end up in Preeti's predicament without any particular curtailment of the thought process (i.e. one sees the desirability of the cake, but the matter of whether on wants it is left open). I sympathise with the point, which I take to demonstrate a possible limit of the Transparency approach. Instead, we can have Petra ask something like 'Do you think the desk is brown?', and adjust the responses accordingly.

answer in a way that makes it directly clear to Petra that the conclusion she reached is in response to the specific question posed.[217] On Boyle's account, Preeti will already know whether or not she wants the cake, but will have nothing to say in response to the question anyway, because the belief that she desires cake has not been made conscious. Intuitively there is something wrong with both of these scenarios. The intuition the scenarios rub against, I take it, is that these are simple and ordinary cases of neglecting to infer or reflect; the kind of cases that would be explicable if one were to say: 'I guess I didn't think about it (enough, properly, or fully, etc.)'. If self-knowledge is really just putting into place a kind of reasoning supplemented by inference or reflection, as is the case these accounts, then there ought to be nothing more to these examples than there is to any other case of failing to think things through.[218] The task now is to consider why it might intuitively be wrong in these cases to think otherwise.

To attempt to clarify the intuition, we can first take some partly analogous scenarios that we might expect to occur on this model of thinking:

(1) There is an exchange in which Jesse gives James a festive gift, which James accepts in good faith. However, James's thought process is curtailed, and he doesn't know for whom the present is intended.

(2) Sam and Arjun are talking about what Bob thinks. Arjun poses this to Sam in the following way: 'What does Bob think?'. Sam considers the relevant factors and concludes '*p*'. However, Sam's thought process is curtailed, and Sam cannot attribute '*p*' to Bob.

Intuitively, at least, it does not seem that (TT) is true in case (1) because James has already accepted the gift and does not need an additional reason to recognise it as his. And it does not seem that (TT) is true in case (2) because Sam has

[217] Of course, conversationally, Petra may take 'cake is good' to mean 'yes please'.
[218] For example, a simple arithmetical equation.

already accepted the task as being a one about what Bob thinks. Any question of to whom the thought applies has already been settled. On the basis of these examples, one might venture to conclude that (TT) looks implausible as a general rule. So we might take it to be a feature of self-knowledge cases only.

Why should we accept that? The natural response may be to suggest that numbered examples above miss something important about a feature of first-person thought, namely, that the thinking related to the questions *whether p*, and *whether I think that p* naturally elide. However, this would present a potentially unwelcome asymmetry between first-person and third-person thought: assuming that the examples above are roughly on target but are intuitively odd, accepting that they would be fine in first-person cases would now leave the first-person with an apparent disadvantage.

For those concerned that bare intuitions can be misleading, there are further reasons to suspect that there is something wrong with the pictures as presented. Firstly, the picture does not fit well with the activities that it is commonly assumed the subject is engaged in. Secondly, the pictures appear to presuppose a specific view of Transparency that is not obviously the right one. I will briefly outline both of objections.

1.3 Deliberation and the TT assumption

On both Byrne's (2011a) and Boyle's (2011) views, deliberation is a bone fide way of arriving at a conclusion that comprises the first step in our Transparency procedures. Philosophers are not always clear about what they mean when using bits of jargon, but on some fairly standard accounts of deliberation, it is the kind of thing that requires either conscious or tacit recognition of the task that has been set.

David Owens, for instance, suggests that deliberation is a conscious activity that aims at resolving an issue (2011, p. 262), and so it has the following features: (a) it is conscious in the respect that it 'occupies the deliberator's attention'; (b) it is an activity in so far as the deliberator is 'trying to do

something … [such as] make a decision'; (c) it is directed at the resolution of an issue or question; and (d) typically it does this by focusing on the salient features of the world rather than psychological concepts (i.e. it is Transparent). (See Owens 2011, p. 262; Appendix 1, §8.) If this is broadly correct, then the subject—in directing her attention to the resolution of an issue or question—must be consciously aware of that issue or question. If we take that issue or question to be 'Do you think that *p*', where the subject understands the referent of 'you', then subject's predicament at the end of step one of the process is quite curious. She has been conscious all the while that she is attending to a question that refers to what she herself thinks, but is left with a conclusion, ostensibly about an altogether different matter, and is suddenly unable to view the answer as one that refers to her (i.e. unless she completes the second step). This would be a fairly specific, and quite unusual incapacity (self-knowledge failure or a type of acute *amnesia*, if one is so inclined). Had one not already bought into a particular assumption about transparency procedures, it would be natural to reject the possibility that this could reasonably describe the standard case of Transparent self-knowledge. Since the subject's thought process is directed towards resolving an issue, she should ceteris paribus[219] clearly *already* be in a position to resolve the matter to its proper conclusion—namely, a conclusion that respects the phrasing of the question, or pays due regard to the task at hand.

Perhaps, however, it is wrong to accept that deliberation is a conscious process in the way implied by my comments above. After all, it is beginning to look as though models that suggest reasoning is best described as a fully conscious train of thought have missed something important: much of our thinking seems to fall beneath that level, and much reasoning that can be

[219] There are a number of ways in which a subject can go wrong in such cases. She might for instance, reach no conclusion on the basis of the evidence, reach a conclusion on the basis of perverse reasoning, get distracted and not complete the task, etc. The point here is whether having completed one stage of a specific kind of task consciously directed at resolving a question about *herself*, an ordinary reasoned might fail to recognise the subject to which the conclusion refers.

described that way—at least in many day-to-day cases—is inefficacious[220] (see e.g. Kahneman 2011). One might at least accept that if deliberation is a fully conscious train of thought, then it is much more rare than might once have been thought. If one does, then one might reject the idea that deliberation in ordinary self-knowledge procedures is likely to be of the fully conscious variety. (See Appendix 1 for further discussion.) In this case we need an alternative notion of deliberation. An alternative would not have the question always at the forefront of a subject's mind or being 'explicitly' posed (Shah 2003, p. 466) by herself or another. However, even on this view of deliberation, the subject's thinking cannot simply be a 'stretch of directionless cogitation'. In order to be a case of deliberation—that is, the kind of activity that aims at settling an issue—the subject's thinking must 'manifest some recognition that this is the question that he is striving to answer' (*Ibid.*).

So, even without a notion of deliberation on which it is a fully conscious activity, it is difficult to see how a subject could end up in a situation in which—that is, following a period of cogitation that manifests recognition of a question that refers to herself—not being able to recognise that the outcome of that cogitation as an outcome that refers to herself without some additional stage or step in her reasoning.

I do not mean to suggest here that such a sequence of cognitive transactions is not possible, or that it is not one method by which one could plausibly come about—other things being equal—knowledge of one's mind. It is unlikely, however, unless one has already accepted the (TT) assumption, one will be inclined to think that the subject—in standard cases such as Evans's example above—could get herself into a mess, from which she can only be extracted by putting into place some additional bit of reasoning.[221] Why then, might one be tempted to think that this is the case? Perhaps the reason is that one has a particular view of Transparency. In the next section, I briefly outline a

[220] That is, the *decisive* elements in human reasoning are not always consciously available (Appendix 1; §9).
[221] It is also difficult to see how it would be possible, since she would (ostensibly) have little reason to engage in the second bit of reasoning if the original epistemic project has been lost.

small number of views of transparency to see if a motivation for the (TT) assumption can be garnered from them.

1.4 Transparency and the TT assumption

The general observation behind the Transparency of first-person thought, at least in the doxastic self-knowledge case,[222] is that one can and typically does approach questions of what one believes without 'essential reference' to oneself or one's belief (see Edgley 1969, p. 90; in Moran 2001; Appendix 1, §12). This in itself is no motivation for the (TT) assumption. So perhaps it can be located in more specific notions of Transparency. Below are outlines of three fairly standard notions of Transparency:

> (1) The questions 'Do I think that *p*?' and 'Is *p* the case?' are first-personally indistinguishable (Edgley 1969, p. 90; in Moran 2001).
>
> (2) 'I can know various aspects 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 … there is in the normal case no other way to attend to them' (Boyle 2011, p. 3).
>
> (3) '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* … whenever you are in a position to assert that *p,* you are ipso facto in a position to assert 'I believe that *p*' (Evans 1982, pp. 225f.)

Nowhere in (1) to (3) is it stated or implied that one ought to adopt something along the lines of the (TT) assumption; a natural reading of (2) would suppose no further action would be required for knowledge of one's mental states, and (3) explicitly states that by the very act of putting into operation the procedure

[222] Deliberative Transparency can come in a number of contrasting forms. One can, for instance, deliberate on 'what I should believe' or 'what I do believe' (cf. Shah and Velleman 2005).

for answering the question whether *p*, one is in a position to answer the question whether *I believe that p*. The conclusion that one might tentatively draw from this is that the requirement of (TT) must be extraneous to general characterisations of Transparency (at least those articulations here).

What the characterisations above suggest is a that thoughts about *whether I believe that p* and *whether p* elide in a way that sees the former give way, or to be settled by the latter, or that the decisive elements in one's cogitations about the former are nothing more or less than the matters that settle the latter. But none of that rules out the fact that the subject can remain mindful of the original task, and none of it implies that there is a tendency to conclude one's cogitations with a judgement that detaches the subject from her conclusion in a way that requires reunification at a later point.

The assumption (TT), then, is neither required or implied by fairly standard characterisations of Transparency; it is counter-intuitive to suppose that an ordinary subject would be prone to a kind of failure that (TT) implies given the kinds of activity she is undertaking and has ostensibly accepted; and on standard conceptions of one accepted route to responding to that task the subject engaging in the task must at least be sufficiently cognisant of the question she is striving to answer for the question to give direction to her thoughts. Why, then, ought we think that there is reason to accept the (TT) assumption? The answer, I suspect, is that it is evident that one can have a belief with the content '*p*', without realising that one believes that *p*. What this amounts to is that there must be some act or element of self-attribution for a subject's beliefs to be recognised as her own. However, this neither requires nor implies that in standard cases of self-knowledge by Transparency procedure, the subject will typically find herself with a cognition that is not acknowledged as her own. In the next section, I will outline a way of fulfilling this condition that does not require that one accept the (TT) assumption.

2. Transparency and self-attribution

In chapter one, I suggested that self-ascription is one of the few basic ingredients of introspection. If that is right, plausible self-ascription must be a part of any theory that aims to do the explanatory work of introspection. In §1 of this chapter, I argued against an assumed pattern of self-ascription that has been the cause of a number of disputes in the literature. Now I turn to what I take to be a more plausible way to characterise the requirement for self-ascription. In this section I argue that the correct way to characterise the general requirement of a theory to explain self-ascription is to dispense with the temporally ordered components implied by (TT). And, because I have accepted Transparency as a desideratum for a theory (see Ch. 2), I will need to provide a plausible case in which the output of a Transparency procedure leaves one in the same eventual position implied by (TT). I do this by describing a form of judgement that is a plausible product of a Transparency procedure and yet does not require an independent act of self-attribution.

Not all thoughts or judgements require an independent act of attribution to be recognised as one's own, but then not all thoughts are Transparent in the relevant sense. Cogito-like judgements, which are noted for their 'contextually self-verifying' properties, are notable also for their structurally self-ascriptive properties. Take an example from Burge (1996, p. 92): 'I am thinking that there are physical entities' (IT). Judgements like IT are *infallible* in the following sense: 'One cannot err if one does not think it, and if one does think it one cannot err'. The combination of understanding the thought and engaging in it, make it true (p. 92). Because the judgement is also *structurally self-ascribing*, when one understands the thought, and engages with it, there is no question about to whom the thought refers. Such thoughts contrast significantly with what is assumed to be the case with thoughts resulting from the first stage of Transparency procedures. Those thoughts—conceived of as matters independent of the subject—require an additional step clearly not required in

cogito cases. However, since they are implausible candidates for Transparency judgements, and so they do not count against the (TT) pattern.

Cogito-like thoughts are not the only variety not to require a second, ascriptive step. At least some thoughts arising from autobiographical memory, for instance, are argued to possess the same or a similar property: immunity from error through misidentification (IEM). It is doubtful that 'logical' immunity from error—that is, 'strong IEM'—is a property of autobiographical memories.[223] However, it is doubtful that any judgements possess that property (Bermúdez, forthcoming), and a weaker version of the property is still possesses the features relevant here.

To avoid confusion with strong IEM, I will generally refer to 'Auto-identification'.[224] Good examples of auto-identification come in the form of 'explicitly recollective' autobiographical memories (*Ibid*.).[225] However, these too can be poor candidates for judgements that might arise from a Transparency procedure, since they contain the first-person pronoun. (Judgements explicitly containing the first-person pronoun are unlikely to be accepted as 'transparency' judgements.)[226] Not all autobiographical memory judgements are 'explicitly recollective', but at least some that are not 'explicitly recollective' still auto-identify.[227] They auto-identify when they are past-tense judgements that have both an 'experiential basis'—what I have discussed as acquaintance with an event, or cognitive contact (Ch. 4)—and their 'present tense analog has the immunity property' (Bermúdez, forthcoming).[228] This is relevant because

[223] The worry is this: 'quasi-memories' of someone else's experience are a possibility. If they are a possibility 'then so are errors of misidentification': so those judgements cannot be 'logically immune' (Bermudez, forthcoming).

[224] Following Bermudez (forthcoming) and Evans (1982) dichotomy: 'identification-dependent' and 'identification-free'.

[225] For example, 'I recall falling over'.

[226] The matter is not so clear-cut given variation in notions of Transparency. It is notable, for instance, that Byrne's (2011a) schema for intention starts from the premiss 'I will **Φ**'.

[227] Bermúdez (2012) argued that autobiographical judgements all possess IEM, but considers this a mistake in Bermúdez (forthcoming).

[228] 'They possess it [the relevant property] when, and only when, the recalled experiences are such that they would have warranted a present-tense judgment that would itself have had the immunity property … In this way, therefore, the immunity status of past-tense memory judgements is inherited from the epistemic features of the original experience' (Bermúdez, forthcoming).

some autobiographical memory judgements that do not contain the first-person pronoun—and are thereby are plausible candidates for the kind of judgements that can be produced by the first stage of a Transparency procedure—will not require an independent act of self-ascription.

This finding may have its own implications for the connection between memory and self-knowledge.[229] However, because the main focus here is not autobiographical memory, I point to it in order to emphasise the fact that the pattern of cognitions implied by (TT) should not be adopted as a general characterisation of the self-ascription requirement, even in Transparency cases. Nevertheless, because we must build some means of self-ascription into our theory, here is an alternative to (TT) that captures the more general Transparency requirement for introspection:

> (TAC) A self-knowledge procedure that meets the Transparency condition must explain how ascription occurs.

The main difference between (TT) and the Transparency ascription condition (TAC) is that the temporal ordering of cognitions suggested by (TT) has been removed. Even if the (TT) pattern occurs in some Transparency judgements, it is not the correct way to characterise the entire class of judgements in question. A certain class of judgements that can be produced by engaging a Transparency procedure do not require the second stage implied by (TT). We can conclude for the moment, then, that (TT) is a questionable assumption when measured against commonsense cases, when considering the kind of activity that a subject is engaging in, is neither required nor implied by characterisation of doxastic (and intentional) Transparency, nor does it correctly portray the full range of judgements that can issue from Transparent cognition.

[229] Bermúdez discusses what he calls and 'interdependence between memory and self-consciousness'.

With (TAC) in place, we are now able to move on to outline the remaining part of a simple mnemic theory of self-knowledge, and to set it against the desiderata laid discussed in chapter two.

3. Doxastic self-knowledge as recollection

In chapter three we saw how, given an appropriate epistemology of memory, standard case of self-knowledge sees the subject *recalling the fact that p* (see also Byrne 2011a, p. 208). In the theory under discussion at that point, *recalling that p* was always the first of two steps. I have suggested above that this is wrong. In fact, there appear to be a number of cases in which the self-ascription is explained by other means.

In §1, I suggest that—barring an unusual kind of amnesia or introspective failure—if the subject is mindful of the question to which she is attending, she will generally already be aware that the answer relates to her. This suggestion found support in common notions of how the appropriate shaping of a subject's thinking in response to a question must at least involve some recognition that she is responding to that question (§1.2). In §2, I suggest that Transparent reasoning can issue in judgements that are auto-ascriptive without containing the first-person pronoun, thus bypassing the need for further attribution while still respecting the Transparency intuition. We can describe these cases as follows:

(A) Task-identified cognitions
(B) Auto-identified cognitions

Both (A) and (B) are initially plausible cases of judgements that avoid the need for an independent act of self-ascription while meeting the self-ascription condition (TAC).

We might be tempted, therefore, to suggest that mnemic self-knowledge is a kind of abbreviated Transparency procedure; one that sees the subject responding to an inquiry (from herself or another) while mindful of that inquiry (A) or responding to the question autobiographically (B). In both cases, the responses can be seen as responding to an (Evans-style) inquiry. Let us consider the example: 'Do you think that the ball crossed the line?' In the first case, the subject responds by considering whatever relevant facts she has at her disposal (e.g. what the experts say) and is mindful of the question. In the second case, the subject recalls something from her past that relates to the question, for example, an image of the ball and the line and their relative positions.

At least two problems become immediately apparent if one wishes to suggest that (A) and (B) are sufficient to explain standard cases of doxastic self-knowledge. The first is that suggesting a subject is 'mindful'—while no doubt the case—does not add a great deal to the explanation. The second—highlighted in chapters three and four—is that factual memory and memory experiences are epistemically independent and, in the majority of cases, our reasons for believing are relinquished (see also e.g. Owens 1999) and therefore unavailable. Cases of (A) then are under-explained, and cases of (B) are relevant only in a tiny subset of cases. Where do we go from here?

The answer, I think, is that both cases say something important about how a subject might be seen to engage with this kind of inquiry successfully. While it is right to think that responding to the inquiry successfully ought not to be seen as avoiding a special kind of otherwise impending failure, it is not quite enough to suggest that the task alone guarantees a particular kind of success; or that our general competence in the face of such queries eliminates the need for a story about attribution. One way of completing the story—though not the only way—borrows from the structure of (B) cases while avoiding some of their deficiencies.

Thoughts of the (B) variety are auto-identifying due, in part, to their cognitive contact with some event in the subject's history. This cognitive

contact can help to explain why subjects are able to competently respond to queries that are ostensibly a muddle of two subject matters: one about the world, and one about themselves. Following the Evans-style case: Subject S, when faced with and accepting a question like 'Do you think that *p*?', initiates whatever procedure she has at her disposal for answering the question, *whether p*. Doing so standardly requires that subject recognise the task with which she is faced. And that task, of course, is the former question rather than the latter. (One reason why Transparency is interesting is that one can and does answer the first question by considering elements decisive in answering the second question; not that one does not answer the first question, but answers the second question instead). The posing of the question—either by herself or another—is an event experienced by the subject, and one with which she has cognitive contact. Ordinarily, we have seen, cognitive contact is fickle. However, the nature of the task with which the subject is engaged requires that she be cognisant of that task: it is that recognition alone which prevents her response being 'stretch of directionless cogitation' (see above). In other words, the cognitive contact with the question–event is artificially maintained by the nature of the inquiry.

We can now introduce the relevant elements of (B) into (A) to provide a more complete explanation of task-identified cognitions (TIC):

> (TIC) Whenever a subject has cognitive contact with the experience prompting the task in which she is engaged, and that cognitive contact issues in the property of auto-ascription, the ensuing judgement requires no additional independent act of ascription.

It is important to note that the presence of (TICs) does not suggest that there is no possibility of error at all. We can and do forget some activities in which we are engaged. In chapter one I mentioned the dangers of forgetting what one intends to do. If one can forget why one is minded to go upstairs, then one can

plausibly forget the kind of task one is engaged in when responding to Evans-style queries. Most plausibly, one will be more likely to forget it if the task is prolonged, difficult, or if one is somehow distracted. Such failures do not entail a loss of cognitive contact with the relevant experience, but if there is such a loss, the result is not catastrophic: even in the absence of another method of keeping her task in mind[230] such that she does not need to follow the pattern suggested by (TT) (there are options to be explored in that regard) the subject will merely need some additional method of attributing the result of the activity. So a subject can still know her mind in the absence of (TICs), although when she knows her mind via (TICs), the process has a number of advantages that (TT) processes do not have. A range of these advantages are outlined in the next section, however, (TICs) explain the theoretically supported intuition that once one consciously engages in—or manifests some recognition that one is—responding to a particular question or issue, one is unlikely to have to 're-attribute' the issuing judgement.

Since (TICs) apply to both deliberative and non-deliberative cases of doxastic self-knowledge, it will be helpful to briefly rehearse the suggestion (from Ch. 3) that the standard case of doxastic self-knowledge is non-deliberative.

3.1 Non-deliberative doxastic self-knowledge

In chapter three we saw that the standard case of doxastic self-knowledge is non-deliberative. I mentioned a number of factors in favour of that conclusion. A fairly simple and straightforward objection to the alternative is that one clearly does not have to make up one's mind each time one wishes to know what one believes.[231] In many cases my mind is already made up and there is no

[230] The ordinary notion of attention looks like a promising alternative explanation. If a subject is paying attention to the task at hand, an explanation of the kind of potential failure implied by (TT) would be the burden of the opponent. Though, I will not attempt to flesh out this idea here.

[231] We should admit that, alone, this is not a good objection. The claim in favour of deliberation can be appropriated stated with deliberation as *the most important* case rather than *the most frequent* (see e.g. Boyle 2011). (Whether we have good reason to accept this version of the claim is a different matter).

deliberation prior to self-ascription of a belief that *p*. In such cases, I simply recall the fact that *p* rather than consider the evidence (see Byrne 2011a, p. 208; Ch. 3, §3.1).

A more pressing difficulty with the alternative is that a Transparency procedure that allows for deliberation will 'contaminate the result by possibly altering the state that one is trying to assay' (Shah and Velleman 2005, p. 507). The risk of contamination by deliberation means that affected Transparency procedures are only reliable guides to what one thinks at the end of the procedure, as opposed to what one thinks prior to its initiation (see Gertler 2011a) and this leaves a feature of Epistemic Security unexplained: 'If I have no belief that p (at t_1) but consider whether I have a belief that *p* (at t_2) I will not self-attribute a belief that *p* without creating a new belief' (*Ibid.*). Thus, there is good reason to think that the standard case of doxastic self-knowledge is non-deliberative, and that if it is factual recollection (in the broad sense), then it must be recollection that is uncontaminated by reasoning as to whether *p*. The non-deliberative, that is, recollection case of doxastic self-knowledge better explains a number of desiderata outlined in chapter two. Our best candidate for a standard case of doxastic self-knowledge for our theory is a recollective case of (TIC): when S responds to an inquiry of the variety, 'Do you think *p*?' she recalls whether *p* is the case, as it were, *auto-ascriptively*. The auto-ascription is explained by means of the artificially extended cognitive contact with the stimulus event provided by the engaging with the task initiated by that event. In such cases there is no need, or indeed room, for the kind of second step found in Byrne-style and Boyle-style Transparency procedures. (Evans's remark that the subject is *ipso facto* in a position to assert 'I believe that *p*' looks plausible on this picture.)

The explanation is incomplete in two respects: (i) it lacks a full description of how the auto-ascriptive judgement concerning the inquiry combines with the result of that inquiry; and (ii) whether or not the IEM property survives the process. To some degree, these aspects of the account will

have to remain partially incomplete. However, with regards to (i) it is important to note that the Transparency views expressed above (§1.4) do not preclude the possibility that a subject can explicitly acknowledge what doxastic Transparency means: that the decisive elements in her treatment of the question 'Do I think that *p*' are precisely those that will be decisive for the question *whether p*. There appears to be no contradiction, or even tension, in the single thought: 'I will answer[232] the question of "Do I think that *p*" by reference to what is decisive for answering the question "is *p* is the case?"'. If it is possible for a subject to think such a single thought, then she can replace the variables '*p*' and 'what is decisive' with the specifics of each question or issue. How the individual components of such a thought are bound together might be open to dispute, but the results of that dispute are unlikely to lead us back to the pattern implied by (TT).

For the purposes of discussion, I will call the general view mnemic self-knowledge (MSK). The remainder of this chapter will highlight the benefits of the view by examining the extent to which the view meets the desiderata laid out in chapter two (§4), and by considering a range of possible objections (§5).

4. Meeting self-knowledge desiderata

In the last section, I outlined what I take to be, and what I have presented as, the standard case of doxastic self-knowledge, a core case of (MSK). In this section, I argue that the view can fare well against the desiderata discussed in chapter two.

4.1 Peculiarity, Immediacy, and Epistemic Security

In this sub-section I suggest that (MSK) can meet all three of the minimum criteria laid out in chapter two. I called these the PIE theses: Peculiarity (P); Immediacy (I); and Epistemic Security (E).

[232] Adjusting for tense: 'I have answered … ' or 'I am answering … '.

4.1.1 Peculiarity

Peculiarity was formulated as: 'a method or procedure by pointing to which it is possible, satisfactorily, to explain how S comes to know her mental states, and that cannot be used satisfactorily to explain how one S comes to know the mental states of others' (Ch. 2).

The doxastic version of (MSK) is an explanation of how S comes to know her belief, just as long as the relevant forms of memory are explanations of knowing. As far as I can tell there is no coherent challenge to the notion that factual memory—as described in chapter four—is a way of knowing, and it has good pedigree as an 'accredited' way of knowing.[233] Because (MSK) makes use of autobiographical memory—more properly memory experience (Ch. 4)—the following objection might be raised: the function of autobiographical memory is to preserve and protect a coherent picture of the self (Conway and Loveday 2015), not to accurately reflect the world or the past (Ch. 4, §2).[234] Because memory experiences are autobiographical in the sense that they relate to a subject's past, it may be that the use of memory experience undermines the reliability of the procedure. While there is support for what we might call the coherence view of autobiographical memory, and for the unreliability of autobiographical memory in general, there is little reason to suppose that it is pervasively unreliable in a way that would substantively threaten short-term question–answer tasks. Because memory experiences can be declarative, we have a way of assessing whether judgements about the past are accurate and false, and the general reliability of memory counts against the possibility of pervasive error (see Senor 2014).

The procedure cannot be used satisfactorily to explain how S comes to know the mental states of others, because firstly S does not have access to

[233] Memory tends to appear in standard lists of ways of knowing, for example, Ayer's (1956) list of accredited ways of knowing: 'Claims to know empirical statements may be upheld by reference to perception, or to memory, or to testimony, or to historical records, or to scientific laws' (p. 31). Questioning this conclusion is beyond the scope of this thesis.

[234] Bermúdez discusses with a similar point (forthcoming).

others' memory in the relevant sense, and auto-ascription is a first-person phenomenon. It is clear then, that (MSK) for belief meets the Peculiarity condition.

4.1.2 Immediacy

In chapter two, I concluded that there were two variants of Immediacy that are worthy of consideration. Explaining *psychological* immediacy—according to which 'subject (S) can be knowledgeable about her current mental state (C) without being able to provide her reasons or evidence for self-ascribing mental state (C)' (see Ch. 2, §2)—is a minimal criterion for a theory. Explaining *epistemic* immediacy was retained as an ideal desideratum. I described the epistemic variant as: 'subject (S) can be knowledgeable about her current mental state (C) without inferring that she is in (C) from reasons or evidence that she is in (C)' (*Ibid.*).

The doxastic version of (MSK) explains psychological immediacy because for the majority of retained beliefs, a subject will not have access to her reasons and evidence for her belief (see e.g. Owens 1999). We can see now, also, how it might explain the epistemic version of the thesis: recalling that *p* is not rehearsing an argument for *p*—recall is a 'got it' verb (Ryle 1949, p. 254).[235] So there is no inference involved in the recollection part of the procedure (p. 250). There is also plausibly no inference (or indeed reflection)[236] required for the attributive 'part' of the procedure, since the judgements are *pre*-ascribed via cognitive contact with the inquiry-event.

4.1.3 Epistemic Security

The Epistemic Security thesis (E) requires that: 'beliefs about one's mental states are more likely to amount to knowledge than one's corresponding beliefs about others' mental states' (Byrne 2011a, p. 202). There were a number of

[235]Ryle (1949), for example, makes this clear a number of times (see also e.g. p. 250).
[236] That is, at least no reflection of the variety implied by Boyle (2011).

options with regards to this criterion, and most were modest (Ch. 4). Nevertheless, modest advantage is all that is required by the condition. Broadly conceived, on the doxastic version of (MSK), (E) can arise from (i) the positive epistemic status bestowed by the retention of belief; and (ii) can be achieved by inducing cognitive inertia (that is, for instance, taking greater care when adopting beliefs). To this, we can now add (iii) a weak form of immunity from error through misidentification (IEM). IEM plausibly provides a greater degree of security than the others, however, even if there is reason to doubt IEM is available via (MSK), then the modest forms of security are arguably enough for the asymmetry implied by (E).

4.2 Uniformity, Economy, and Transparency

Having, arguably, fulfilled the main requirements of a theory of self-knowledge, we can now move on to the ideal desiderata. In chapter two, I suggested that, ceteris paribus, a satisfactory account of self-knowledge should be fundamentally uniform, explaining all cases of "first-person authority" … in the same basic way' (see Boyle 2009, p. 141), although I also suggested that there might be good reason to reject the 'uniformity assumption', along principled lines in the taxonomy of states under review. Chapter one listed six varieties of self-knowledge failure and corresponding success, and although there are reasons for optimism on in a number of cases, a full review of each cannot be undertaken here. Of the six varieties, the main focus of the discussion has been on the fifth: intentional states. If (MSK) is plausible for belief, then there are good reasons to think it will be plausible for intention and desire since the standard cases of both are mnemic—that is, one does not make up one's mind each time one wishes to know whether one intends to Φ and deliberating on the matter will—to follow Shah and Velleman's (2005) concern—'risk contaminating the response'. And there is reason for more general optimism: 'If retrospection can give us the data we need for our knowledge of some states of mind, there is no reason why it should not do so for all' (Ryle 1949, p. 148).

However, for the moment, we can afford to be cautious. The interim conclusion should be that for any state about which deliberation is a plausible option, (MSK) will be a plausible self-knowledge procedure. Beyond that, each variety of self-knowledge will require further consideration. It is worthy of note, however, that while this does not leave us with a theory of self-knowledge that explains self-knowledge for all relevant states, it does leave us with a theory that covers the same ground as Boyle's (2011) view. There is modest success, so far, with regards to uniformity.

The *Economy* desideratum suggests that, ceteris paribus, 'a theory that explains the distinctive features of self-knowledge without recourse to capacities not employed in other domains of Knowledge' (see also Byrne e.g. 2011a). Since (MSK) makes use only of general epistemic and rational capacities required for knowledge in other domains, (MSK) meets the Economy desideratum.

It also meets the Transparency desideratum, because one can come to know one's mental state without considering anything 'inner' or 'psychological'. Factual recall does not require the additional belief that one is recalling—'explicit recollection'—and while many obvious cases of IEM in memory do contain the first-person pronoun, these are not the only memory cases of IEM.

We can conclude, that with the exception of currently limited success in Uniformity, (MSK) meets the ideal desiderata. Since there are good reasons to question the uniformity assumption, this does not constitute a particular worry for the account, even if success is limited to a particular variety of self-knowledge.

4.3 Additional desiderata

Up to now, (MSK) appears is a promising way to explain what is thought to be special or distinctive about self-knowledge—with the exception of only modest success in Uniformity (a feature that requires further inquiry)—both the

minimal criteria, and ideal desiderata can be comfortably met by the view. In chapter two, I outlined a number of other plausible demands upon a theory. For the remainder of this section I will briefly outline how (MSK) meets each of these.

Agnotic Access is the requirement that a theory explain not only self-knowledge of belief, but also self-knowledge of its absence. Some commentators have seen the latter as a particular challenge for Transparency accounts, ostensibly because one is meant to be considering the world, and it is difficult to see how the world can come up empty handed (see e.g. Fernández 2013). The concern might rest upon an overly exacting notion of Transparency. (Compare, for instance, those outlined above.) Nevertheless, doxastic (MSK) meets the desideratum because, to put it bluntly, one does not recall a belief that that one has not adopted. This, of course, is not an explanation of the subject's awareness that she does not believe. However, an explanation of that awareness will, in part, be available due to a key difference between recollective and deliberative thinking. *Recalling that p*, for the most part will be immediate. Deliberating over *whether p* will not. In one case one is accessing the content of a retained belief. In the other one is settling the matter, or making up one's mind by considering factors that go in favour of one conclusion over another. Some forms of recollection can be confused with deliberation. And this confusion can give rise to a common form of self-knowledge failure (discussed at length in Appendix 1). This form of self-knowledge failure generally occurs when one is asked to recall decisive factors in the formation of one's attitudes—which, as we have seen is not ordinarily something that we are apt to do. Straightforward recollection of first-order belief content, on the other hand, is markedly different psychologically speaking: it is *psychologically* immediate, whereas the former is not. There is, of course, room for error, but no reason to think error is more likely than success. The analogous case for abilities (Ch. 2) highlights the difference: there is a reason we think it is funny to suggest that someone would try to speak Spanish to find out whether they could speak Spanish.

Preserved Access is explained because recalling that *p* is a case of detecting the belief in place prior to the initiation of the self-knowledge procedure. The inferential Transparency view left open the possibility a subject would self-ascribe a belief formed during the procedure itself. On that view such a self-ascription would count as success. On (MSK), however, the standard case is recall, which is not conducive to the production of new beliefs, at least not in the troublesome way possible on the inferential Transparency view.

I have suggested that *Evaluative Access* is enabled by *Preserved Access* (Ch. 2). Without *Preserved Access*, it is difficult to see how a subject could successfully 'assess or reflect upon her current (i.e. pre-existing) attitudes in light of her available evidence and the norms she accepts' (*Ibid.*). A full explanation of how (MSK) explains (if it does) Evaluative Access is a matter for elsewhere. However, by explaining how Preserved Access is possible, I have left the door open for that feature. In that respect it is an improvement on both inferential and reflective Transparency approaches to self-knowledge.

Finally, (MSK) will not leave the subject Self-Blind. The subject is not left with third-person only access to her mental states, and can acquire knowledge of her mental states without treating them as independent objects of observation.

5. Queries and objections

In the last section of this final chapter I would like to pre-empt a number of initial concerns about an account of self-knowledge that explains what is special about knowledge in the domain via features of memory.

Is all self-knowledge question-led? Self-knowledge on this account is seen as a response to an inquiry, but one might object that not all cases of self-knowledge follow this pattern. A number of things must be said response to this point. The first is to point out that, generally speaking, one could not be

mindful of the full set of one's attitudes, or even a significant proportion of them. For most varieties of self-knowledge discussed here, the relevant data must be 'detected' by some means. The formation of attitudes is an interesting but independent matter. The majority of our attitudes at any one time will be retained attitudes (or what some misleadingly call memory attitudes).[237] What are detected in standard cases of self-knowledge are those attitudes (cf. Byrne 2011a).

Cassam (forthcoming) calls these '"in question" (IQ)' cases and suggests that some important forms of self-knowledge are not IQ. Certainly there are cases where such a question has not been made explicit and one still finds oneself suddenly in agreement with something that one reads or hears (*Ibid.*). These are interesting phenomena, and deserve more attention than can be afforded here. Many cases are likely to involve implicit questions (one wonders whether one agrees), or questions that are the product of silent soliloquy ('Do I believe *that*?'). Neither will pose a problem (the question is posed by oneself in either case), but if there are genuine cases of attitude detection that do not contain either of these elements, then this might be thought a limitation of the theory. A plausible explanation of these phenomena may be found in involuntary memories and 'chaining', wherein 'the contents of an involuntary memory sometimes trigger additional involuntary memories' (Mace 2006; also Ch. 4). This will not be a straightforward case of (MSK). However, the project here is to describe the standard case of attitude detection, and Cassam himself notes that it is not clear whether they are cases of detection, formation, or something in between (*Ibid.*). For the moment, at least, they do not present a substantive problem for the account.[238]

[237] See Ryle (1949): 'Theorists speak sometimes of memory-knowledge, memory-belief … This is a mistake … Reminiscence and not-forgetting are neither 'sources' of knowledge, nor, if this is anything different, ways of getting to know. The former entails having learned and not forgotten; the latter is having learned and not forgotten. Neither of them is a source of learning, discovering or establishing.' (p. 249)

[238] If one hopes for a memory effect that matches the phenomenon, then a plausible explanation can be found in involuntary memories and 'chaining': 'the contents of an involuntary memory sometimes trigger additional involuntary memories' (Mace 2006).

Can the account explain self-knowledge of what one remembers? This is another genuine phenomenon that the approach has not attempted to explain, although this ought not suggest that it cannot provide an explanation. The sense of the term 'remember' in use will be crucial in the explanation. One way of making sense of the issue is to see it as a 'how' question—such as, 'How do you know that Cimetière du Père-Lachaise is in Paris?'—that has a number of distinct kinds of response: (i) 'I remember going there while in Paris' (memory experience); and (ii) a standard case of recalling that *p* where *p* is 'Cimetière du Père-Lachaise is in Paris'. Cases of (i) have not been treated in depth here, although a solution is suggested above: if the judgement has the correct properties, then it is a fairly standard case of self-ascriptive memory. Cases of (ii) are somewhat more complicated. Factual memory as presented here, is—in Ryle's (1949) terms—'allowable paraphrase of the verb "to know"' (p. 248). In the absence of associated memory experiences, knowing that one remembers in this sense will likely require an inference (one believes that *p*, but is not currently perceiving that *p* and has no memory of acquiring the belief that *p*). However, this does not substantively change the nature of procedure.

Are inference or reflection not still required in (MSK)? I have presented (MSK) has a non-inferential and non-reflective procedure, although there are a number of places where one might be tempted to insert a requirement for either. One reason for presenting (MSK) as non-inferential and non-reflective was to bring into sharp relief the potential explanatory value of memory in this domain, rather than argue for that value by using it to supplement an existing model of self-knowledge such as the inferential Transparency view (see Ch. 3). On this latter view, for instance, the strongly self-verifying procedure leaves the subject nearly infallible (Carruthers 2011) in any case, and so any epistemic contribution by memory would pale by comparison (and possibly over-determine the result). If it turned out that (MSK) was a partially inferential or reflective process, then the conclusions about Immediacy would need revisiting. However, unlike the inferential and reflective views, knowledge in the domain

would still not be substantively explained by those cogitations. Pace proponents of those views, and to paraphrase Ryle once more,[239] on my view, a good self-knower is not someone who is good at inferring and reflecting, it is someone who is good at a recollecting.

Conclusion

In this chapter I have argued that a standard case of doxastic self-knowledge can be described purely in terms of memory. I first challenged an assumption about self-ascription in the literature on the Transparency approach (§1); and argued for weaker characterisation of a self-ascription requirement in Transparency accounts of self-knowledge (§2). With this weaker condition in place, I outlined a case of doxastic self-knowledge that can be appropriately explained as a kind of recollection. In §3, I outlined a theory with memory in the main explanatory role of doxastic self-knowledge; and in §4, I set the theory against the desiderata for a theory of self-knowledge in chapter two. Finally, in §5, I pre-empted a number of questions and objections. The overall conclusion for this chapter is that a mnemic theory of self-knowledge fares well against all three varieties of desiderata for a theory of self-knowledge. The focus on memory has shed light on at least one intractable problem in the domain that arises from a well-known puzzle of Transparency approaches. The main findings are in line with the aims of the inquiry stated in chapter one: (i) it looks possible to construct a theory of self-knowledge in which the distinctive features of knowledge in the domain are sufficiently explained in memory terms; (ii) that the inquiry has shed light on some intractable problems in the self-knowledge literature.

[239] Ryle (1949) says: 'There is no such inference; and even if there were, the good witness is one who is good at recollecting, not one who is good at inferring' (p. 250).

Conclusion and Further Work

In this thesis, I set out to consider an often-neglected option in the attempts to explain what is special about knowledge of our own minds. I argued that our lexicon of introspection terms has become confused, and unhelpful in explaining knowledge in its domain. Given an interesting and broad-ranging convergence in our thinking about introspective failure and memory failure, I suggested that we might see the degree to which it is possible to think of knowledge of our own minds as a kind of remembering. I proposed that success would likely mean a positive answer to the question of whether memory can explain what is thought to be special about knowledge in the domain.

In chapter one I concluded (i) that memory can play an important role in explaining a good deal of what we describe as introspective failure, (ii) that it is worthwhile investigating whether that convergence extends to introspective success. I argued (iii) that such an inquiry should proceed by outlining the desiderata against which the success of any theory of self-knowledge can be measured. In chapter two, I considered a range of features we take to be special or distinctive about self-knowledge and produced a list of desiderata—minimal criteria, ideal desiderata, and additional criteria—against which the success of a theory of self-knowledge might be measured.

In chapter three, I set a prominent approach to self-knowledge—the Transparency approach—against those criteria and argued that a particular view of memory would enable the approach to meet an implication of the Epistemic Security condition; namely, Preserved Access. I concluded (iv) that the epistemology of memory plays an important role in explaining introspective success on such a view, and (v) that this strengthened the case for an inquiry into the extent to which the epistemology of memory might be explanatory of knowledge in that domain.

In chapter four, I explored the question of whether memory might explain, or contribute to the explanation, of the three minimal criteria from the list of desiderata (Peculiarity, Immediacy, and Epistemic Security). I concluded (vi) that a surprising number of options are available for each of the three desiderata, and (vii) that this merited the construction of a theory that might be tested against the full list of desiderata.

In chapter five, I highlighted a problem arising from the Transparency approach concerning self-ascription that has affected the ability of a number of accounts to meet come important criteria (e.g. Byrne 2011a; Boyle 2011). I argued that this problem is due to an assumption based on too strong a conception of the requirement that a Transparency account must explain self-ascription. I conclude that a weaker conception (a) better captures the range of data, and (b) better fits standard conceptions of the important features of Transparency in the literature. With this in place I highlighted a straightforward case of doxastic self-knowledge that can be appropriately described in memory terms. I constructed the outline of a theory around this case and set it against the full range of desiderata from chapter two. I concluded that the theory fares well against desiderata of all three varieties.

By focusing on the explanatory contribution of memory, it has also been possible to shed light on at least one of the intractable problems in the self-knowledge literature: an assumption related to the puzzle of Transparency that has resulted in too strict a conception of the self-ascription requirement for Transparency accounts. At the end of chapter five, I address a number of potential questions and objections, some of which will form the basis of further work. The overall conclusion is two-fold: (viii) a theory of self-knowledge with the epistemology of memory playing the main explanatory role can be surprisingly successful when set against desiderata that are common in the self-knowledge literature; (ix) an inquiry into the prospects of memory to explain features thought to be distinctive of self knowledge has the potential to shed light on some intractable problems in the literature.

One last issue remains to be resolved in this work. I have suggested, broadly, that once the correct epistemology of memory is in place, explaining what is special about self-knowledge becomes a lot more straightforward than many attempts would have us believe. Once the requisite detail has been filled out, and when faced with the prospect that it is possible to explain what is special or distinctive about self-knowledge in terms of memory, one may be faced with questions about the terms in which it is appropriate to explain certain putative facts about memory. There are two things to say, initially, in response. Firstly, our memory lexicon is comparatively rich and focused. Secondly, if all that can be achieved by an inquiry into the explanatory potential of memory in the domain of self-knowledge is to reduce two questions into one, then this is progress.

Appendix 1

Choice Blindness and Introspective Competence

1. Varieties of introspective failure

The focus of this paper is a variety of introspective failure: I make a choice, but when providing reasons, I offer reasons that could not be my reasons for that choice. There are other varieties of introspective failure, but this variety has long enjoyed attention in the literature on self-knowledge, and examples of its kind are sometimes taken to be evidence of introspective unreliability more generally. It is important to keep the difficulties presented by this variety of failure apart from more general concerns about our introspective competence.

In a review of two decades of prior research, Nisbett and Wilson (1977) provide an enormously well-cited[240] record of this failure which is taken to have implications for how we think about everyday responses to inquiries into our reasons for actions, preferences, and our trains of thought (*Ibid*.). They concluded that there 'may be little or no direct introspective access to higher order cognitive *processes*' (p. 231; my emphasis), but make no negative claims about our access to our own minds more generally, and are quite optimistic about our knowledge of mental *content*: we have 'a great deal accurate knowledge and much additional "knowledge" that is at least superior to that of any observer' (p. 255).[241] Indeed, they suggest, the 'fact' that we have such knowledge may help explain why we believe we have, in addition to knowledge of content, knowledge of our own cognitive processes (p. 255).

[240] See e.g. Cassam (2014); Carruthers (2011); Gertler (2011); Kornblith (2011); Schwitzgebel (2008); Johansson et al. (2006).

[241] See Schwitzgebel (2006) for a helpful discussion of myths surrounding this paper.

In a famous example from Nisbett and Wilson (1977), participants heavily over-selected the right-most article in an array of identical nylon stockings. When asked for their reasons, no subject spontaneously mentioned an article's position in the array, and 'virtually all' denied the possibility that position might affect the choice (pp. 243–4). Nisbett and Wilson's own explanation of these findings left correct self-attribution of reasons aleatory. Rather than direct internal access to our reasons, we self-attribute reasons based on implicit theories of 'how minds work' (Lopes 2014, pp. 27f.): if one believes a reason to be a good one for a certain attitude, that is the reason we attribute to ourselves when questioned about why we have that attitude (see Nisbett and Wilson 1977, pp. 248–9; Lopes 2014, pp. 27–8).

The research died out in the eighties, perhaps partly due to its association with this model, and a more general concern that attempts to study introspection empirically will always be subject 'to wildly differing conceptual analyses' (Johansson et al. 2006, p. 675). Whatever the reason, the demise of the research left us without a settled way of understanding the phenomenon.

Undeterred by the halt in progress, and assuming improvements in experimental design over the hiatus (Johansson et al. 2006, p. 675), Petter Johansson, Lars Hall, and colleagues have recently (2005–) achieved 'striking' results over a range of attitudes and environments 'without making any assumptions about the mechanisms of introspective misattribution' (Lopes 2014, p. 28). They have found that a surprising number of participants 'fail to notice mismatches between intention and outcome' when faced with a covertly manipulated choice' (Johansson et al. 2006, p. 673).

They call the effect *Choice Blindness* following the literature on *Change Blindness*—an effect in which participants 'fail to detect changes in a scene when the change is accompanied by some other visual disturbance' (Johansson et al. 2008, p. 142). *Change Blindness* research is taken to have serious implications for

how we think about our visual experience.[242] *Choice Blindness* research is taken to show something interesting about the 'relationship between intention, choice, and introspection' by 'surreptitiously … [manipulating] the relationship between the choice and [the] outcome that … participants experience' (Johansson et al. 2008, p. 143). It is taken to show a high degree of willingness, in non-clinical participants, to offer confabulatory explanations for manipulated choices. If the results are reliable, then the variety of introspective failure at issue is surprisingly prevalent, and this is thought to have implications for our conception of ourselves as rational creatures, and some deeply ingrained intuitions about our access to our own minds. In considering the variety of introspective failure in question, I will look primarily at the Choice Blindness research.

Some of our supposed intuitions about introspective reliability enjoy less plausibility than others in the face of everyday observation, let alone rigorous empirical research. 'Strongly Cartesian' conceptions of the mental suggest that 'nothing in our mental life is hidden from us', but no one thinks that nowadays, and it is not clear they ever did (Greenough 2010).[243] Most, at least now, are willing to accept that we can go wrong—perhaps badly and frequently wrong—when it comes to some aspects of our mental lives: we are not always in the best position to judge our emotional states (see e.g. Schwitzgebel 2008a), character traits, or deep motives (see e.g. Burge 1998).[244] Nevertheless, there is an important difference between accepting the kind of challenges associated with coming to know whether we are courageous or loyal, for instance, and widespread failure with regard to awareness of our own reasons and a corresponding willingness to confabulate on such matters. We can accept

[242] Some take the Change Blindness research to show that we have a 'drastically false conception of own visual experiences' (see Blackmore 2002, in Johansson, Hall, and Sikström 2008, p. 143) and others, more modestly, that we 'represent the world in much less detail than what was previously thought' (Johansson, Hall, and Sikström 2008, p. 143)

[243] Several of his own passages, for instance, appear to suggest that Descartes did not hold a *strongly* Cartesian view (see e.g. *Discourse on Method*, AT VI 23).

[244] These too are less recent discoveries than some might expect (see e.g. Aristotle on the importance of friendship for self-knowledge in *Nichomachean Ethics* 1170b5–7).

occasional error, or even frequent error for a given range of states, processes, and characteristics, but we tend to balk at the suggestion that introspection is 'massively and pervasively' misleading, as some have argued (see e.g. Schwitzgebel 2008a).

Part of the reason is this. We take ourselves to be rational creatures. We think that we act, believe, desire, and intend, off the bat of our reasons and sometimes our reason*ing*. Not only do we think we have access to our own reasons, we think we can weigh them up *as reasons*, and to self-regulate when we find something amiss. We think we 'engage explicitly in reason-induced changes of mind'—a feat that requires knowledge of those reasons (Burge 1998, p. 248)—and we expect the same of others. We are thus rationally criticisable, or open to 'critical challenge' (see Lopes 2014) from ourselves and from others.

Widespread confabulation seems at odds with all of this, and so our understanding of effects such as Choice Blindness—and the variety of introspective failure it exemplifies—will have important implications for our conception of ourselves as rational, and perhaps for the claim that we *must* have some knowledge of our reasons. What appears to be at stake, then, is our status as introspectively competent and rational decision makers (Davies 2015).

In this paper, I briefly outline examples of Choice Blindness research that is taken to show a high degree of willingness, in non-clinical participants, to offer confabulatory explanations for manipulated choices. I evaluate several attempts to make sense of the Choice Blindness effect and its implications for our knowledge of our own minds, and reject them as sufficient explanations of the effect. I then present an alternative explanation and consider its strengths and weaknesses, addressing two likely objections. One of the strengths is that it allows us to acknowledge our susceptibility to a certain kind of introspective failure without accepting catastrophic implications for our status as introspectively competent and rational decision-makers.

2. 'Classic' Choice Blindness and Choice Blindness across modalities

In *Classic* Choice Blindness, participants were 'shown pairs of pictures of female faces' and instructed to 'point to the face they found most attractive' (Johansson et al. 2008, p. 143). After the choice was indicated, the selected picture was given to the participants. They were asked to 'explain why they preferred the picture they now held in their hand' (p. 143). In some instances, and unbeknown to the participants, a 'double-card ploy was used to covertly exchange one face for the other'. Thus, in such cases, the 'outcome' of the choice was the opposite of what the participant 'intended' (p. 144). A surprisingly low number of participants detected the manipulation even when no time constraints were placed on the response. Even with unlimited time to explain their preference 'no more than 30% of all manipulated trials were detected' (*Ibid*.). Interestingly, a large majority of the reported explanations of preference (72%) were '*clearly confabulatory* since they reported features of the non-chosen face not possessed by the initially chosen face' (Lopes 2014; my emphasis).

Here are some common concerns: perhaps the faces were too similar with regards to level of attractiveness; or the stakes too low to show anything of wider interest (since we are not going to enter into a long-term relationship with the selected individual); there may be some important difference in the kind of reports offered in manipulated and non-manipulated trials; the laboratory conditions may affect the behaviour of the participants; reporting of detection may be dis-incentivized; and so on. But further studies refined the methodology to address many of these concerns: responses were compared over a number of dimensions including emotionality, specificity, certainty expressed, deceit, and complexity (Johansson et al. 2006); the effect has been shown in a more 'natural' environment (see e.g. Hall et al. 2010); varieties of detection were identified ('conscious', 'unconscious', and 'retrospective') and accounted for in the results (2010, p. 56f.); and research was expanded to test for Choice

Blindness in other modalities (e.g. taste and smell).

The effect was still present in sufficiently high numbers to support the initial findings. The research appears to show that in aesthetic visual, gustatory, and olfactory choices, participants often 'fail to notice mismatches between what they prefer and what they actually get … while nevertheless being prepared to offer … [qualitatively similar] reasons for why they chose the way they did' (Hall et al. 2012, p.1).

What are we to make of these findings? The researchers claim interesting implications for a fundamental assumption in theories of decision-making: we 'detect mismatches between intention and outcome, adjust our behavior in the face of error, and adapt to changing circumstances' (Johansson et al. 2005). And the results should be of interest to market researchers: 'in what sense can attitudes be *real* if people moments later fail to notice they have been reversed?' (Hall et al. 2012; my emphasis).[245] The implications for introspective reliability, especially in general, are less easily discerned.

Some preferences can be transient, subject to shifting attention or mood, and easily overturned without it being thought that this says something important about our status as rational creatures. Caprice is only rationally criticisable in certain cases. Anyone who has struggled with a dessert menu can attest to the virtues of more than one selection: the tarte aux fraises sounds enticing, but it's been a while since I've had cheesecake; I spy the chocolate gateaux, but I'm more worried about calorie intake these days. None of these considerations is in genuine conflict, so there is little mystery how it is that I can come to list the reasons counting in favour of one choice when requested, even a choice I didn't make. (Although there remains a mystery in these cases: why having made one choice, I will apparently fail to notice I receive something else.)

At least some aesthetic preferences will pose more interesting questions, but it is plausible—at least in very low-stakes cases—that analogous, and

[245] More cynical applications to marketing, and other domains, are not beyond the imagination.

equally capricious, considerations are in play in many cases. There is nothing incompatible about *liking* both blonde and red hair; or in feeling attracted now to one, and now to the other; only in *preferring* both at the same time. Since liking and preferring are different, but related, attitudes it is understandable that we might borrow some considerations from former when answering questions about the latter, especially where a preference is marginal, a choice is forced, or we are unexpectedly asked to articulate the deciding factors in our selection process. There is no reason to suggest that such preferences are always shallow and easily overturned—explaining the Choice Blindness data only requires that they often can be.

Some of our attitudes are taken to be less generally casual. Moral and political attitudes, for instance, are *serious* matters. We regularly dispute over them, taking them to be more obviously open to critical challenge. Our reasons for having such attitudes should not be capricious; the attitudes should not be subject to shifting attention or mood, and not so easily overturned. (I do not suddenly think that deceiving in order to obtain ready cash is a good thing because I feel like a change.) But there is evidence that Choice Blindness occurs for this kind of attitude too.

3. Choice Blindness and moral attitudes

Hall et al. (2012) used a 'self-transforming paper survey of moral opinions' and asked participants to 'rate on a 9-point bidirectional scale to what extent they agreed or disagreed with each statement' (Hall et al. 2012, p. 1). Participants were asked to read some of their answers aloud and to explain their ratings, but unbeknown to the participants, two of these responses were 'the reverse of the statements they had actually rated'. So, for example (Hall et al. 2012, p. 1; my emphasis):

> Large-scale governmental surveillance of email and Internet traffic ought to be *forbidden* as a means to combat international crime and terrorism

Becomes:

> Large-scale governmental surveillance of email and Internet traffic ought to be *permitted* as a means to combat international crime and terrorism

In the debriefing—which was constructed to promote the over-reporting of detection—participants were given multiple opportunities, with increasingly stronger cues to report any suspicions (p. 3). The survey covered both 'foundational issues and real world issues' (p. 5) often debated in the media (p. 1); participants were assured that there were no time limits for answering; that researchers had 'no moral or political agenda'; and that the researchers would not 'judge or argue' with the participants' opinions (p. 3). The ratings on the nine-point scale suggested that these were issues the participants 'cared about' (p. 3). Nevertheless, 69% of participants (roughly in line with the effect elsewhere) accepted at last one of the two 'statement–rating relations' (*Ibid*.) and 53% argued 'unequivocally for the opposite of their original attitude' (p. 4).

With addition of this version of the protocol, we have demonstrations that the effect can be found, to a comparable degree, in moral attitudes and for preferences across a range of modalities and environments. Even if one is inclined to think that the effect can be partially explained in the visual, gustatory, olfactory cases by the relative instability of the attitudes involved, this kind of explanation looks much less amenable to the moral attitudes case.

Debate continues over elements of the methodology. But, while it continues, corroborating evidence, suggesting that the effect is robust across a range of attitudes and environments, continues to come in (see Lind, A. et al. 2014; Parnamets, P. et al. 2015), and so there is a case for exploring suitable explanations of the phenomenon. In the next section I evaluate three

prospective explanations of the data.

4. Explanations of Choice Blindness

Before I consider two explanations of the data that are broadly amenable to the researchers' conclusions, it is worth briefly considering a third (debunking) view. On this view, participants are caused to enter into a non-ideal (quasi-delusional) cognitive state. We know that conjuring tricks work, can produce false beliefs that are contrary to the evidence and, arguably, fairly resistant to revision. The Choice Blindness set-up takes a participant from a normal state at the time of her initial choice, to a non-ideal state—in which she makes provably false utterances—by 'deluding' her into thinking she chose differently. Just as it does not follow from the extensive clinical data on psychiatric disorders that 'normal human subjects never have transparent, non-interpretive, access to their own judgements and decisions' (see Carruthers 2011, p. 42[246]), it does not follow from the surprisingly low detection rates and a high degree of willingness to confabulate in Choice Blindness experiments that 'normal human subjects' are *Choice Blind*. The more interesting features of the effect are, if anything, akin to our most startling conjuring tricks[247]—surprising only in so far as it is surprising that non-clinical participants can be convinced they are in the middle of a zombie apocalypse or willingly to conduct an armed robbery.[248] On this view there are no consequences for the population, our introspective competence, or for our reasoning abilities, as a whole.

Whatever the merits of this view it has at least two important deficiencies as presented thus far. Firstly, there are important and obvious dissimilarities between Choice Blindness research and conjuring tricks—even

[246] Carruthers is reporting on, rather than endorsing, the view.
[247] Thanks are due to Tom Stoneham for helpful correspondence on how such a view might be articulated.
[248] These examples are from UK illusionist Derren Brown's *Derren Brown: Apocalypse* (first broadcast in two parts in 2012 on Channel 4) and *The Heist* (first broadcast 2006 on Channel 4).

though the former make innovative use of latter. Stage and television magicians can pre-select and prime their subjects, sometimes conditioning them over several weeks (as in the cases above); there is no scientific methodology, or at least no full disclosure of methodology from which, for instance, replicability (and a range of other theoretical standards) can be assessed; and, of course, we are intended to see *only* the very best results. None of this is true for the empirical research under discussion. Secondly, the idea that upon this view Choice Blindness has no implications for the population as a whole is suspect. Even if this explanation were accepted, the data can be interpreted as demonstrating the extent to which simple prestidigitation can cause non-clinical patients to enter into a non-ideal (quasi-delusional) cognitive state in which they utter provably false responses to questions about their choices. This equally alarming conclusion about the fragility of our cognitive faculties has the additional frustration that it leaves us in no better position with regards to introspective failure of the variety in question—self-knowledge failure due to vulnerability to delusive cognition is self-knowledge failure none the less. I will put this view aside for the moment (though we will see later that it has some explanatory value).

In 'Feckless Reason' (2014), Dominic Lopes examines the implications of the research for the role of reasons in aesthetic responses. He suggests there are two hypotheses available to explain the effect (p. 29f.):

> (1) We do not choose for reasons; we choose and then provide reasons. The manipulation merely brings this out by setting up an unusual situation where the reasons miss their target.
>
> (2) Reasons offered for the choices do not 'target [participants'] initial choice and preference'. The belief that they chose *x* 'determines their preference' and the so reasons offered accord with their eventual preference.

On hypothesis (1), 'confabulation is the norm'; 'choices are not based on the

reasons we give' (*Ibid.*). On hypothesis (2), participants didn't confabulate but our attitudes are 'fickle' (p. 29f.); they can easily be overturned by the suggestion that we chose, preferred, or believed otherwise. Both, says Lopes, explain the data, and both do damage to our 'conception of rational decision making' since on either hypothesis 'reasoning about decisions is post hoc' (p. 30). Both of these explanations of the data are broadly amenable to the researchers' own conclusions about the degree of willingness, in non-clinical participants, to offer confabulatory explanations (or post hoc rationalization) for manipulated choices, and the suspicion that the research has implications for the population at large.

Lopes (2014) favours the second hypothesis, citing research on the distorting effects of reason-stating and reason-forming behaviour, but argues that we can square this hypothesis with our access to our own aesthetic attitudes. Choice Blindness research, he suggests, has only shown something about aesthetic appreciation—and, I assume, any appreciation and attitudes that are taxonomically equivalent—if it has shown that critical reasoning is employed in the formation of aesthetic appreciation. But it is not:

> We form aesthetic attitudes consistent with long-term behaviour partly because we do not try to explain ourselves. When we do try to explain ourselves, it seems that we tend to state reasons that imply an attitude and adopt the attitude implied by those reasons. (Lopes 2014, p. 32)

Choice Blindness research insists that we explain ourselves and, in doing so, engage in reason-stating or reason-forming behaviour, and thus it will have a distorting effect on any of the aesthetic attitudes under examination. But, the failures that occur when one critically reasons about aesthetic attitudes will only lead us to question the reliability of introspective access to those attitudes, in general, if it can be shown that they are the kind of attitudes in which such reasoning is implicit. However, 'reasoning is not implicit in such appreciation.

The difficulty is not that our reports of our reasons are often erroneous' but that 'explicitly stated or formulated reasons are post hoc and have a systematically distorting effect on our attitudes' (p. 33).

Choice Blindness, then, tells us nothing about our awareness of aesthetic attitudes—and taxonomically equivalent attitudes in which reason-stating and reason-formulating is not implicit—it only tells us what happens when try to explain those attitudes. We are:

> in some sense aware of the features of stimuli that speak in favour of one choice over another. However, this awareness is not the same as the kind of state that is either articulated verbally in making a report or mentally in preparing to make a report. (Lopes 2014, p. 34)

Thus Lopes (2014) provides a story about how we have reliable access to a certain class of attitudes that is consistent with the findings of Choice Blindness research, and an explanation of why we get things wrong when we try to formulate or report upon reasons for having those attitudes. Some of our attitudes are *pre-critical* in that our awareness of features which 'speak in favour' of one choice is pre-critical. Reason-formulating and reason-stating are systematically distorting (p. 33) and so in trying to formulate or state reasons for our pre-critical attitudes we distort them and end up reporting, instead, upon a different attitude to the one in place prior to questioning.

I do not think this is the right story to explain the Choice Blindness data across the range of environments and attitudes discussed above. Things aren't quite as bad as the two hypotheses make out and, in some important respects, neither hypothesis is a good fit for the data. In the next section I discuss where Lopes goes right, and where I think he goes wrong.

5. Stability, success, and attitude distortion

In the last section I discussed three potential explanations of Choice Blindness. The first explanation, I suggested, suffers from two important deficiencies and so was (at least temporarily) put aside. The second two, considered by Lopes (2014), are said to explain the available data, but in doing so do damage to our claims to rational decision-making. Lopes (2014) suggests the second hypothesis, combined with some notion of pre-critical awareness can explain the Choice Blindness data while leaving our access to our preferences relatively intact. In this section, I consider whether Lopes's explanation can be adapted to provide us with a workable model of Choice Blindness and offer four concerns that count against it.

Firstly, while this model would appear a good fit for *Classic* Choice Blindness and perhaps for Choice blindness in gustatory and olfactory cases too, it is a less plausible explanation of Choice Blindness in moral attitudes (see §3). For the explanation to work for moral attitudes, moral attitudes must be the kinds of things in which reason-forming or reason-stating is not implicit.[249] We might not normally—at least seriously—expect someone to defend their preference when it comes to desserts, or flavours or odours more generally. While there will be some constraints on what we are willing to accept as reasons stated with regard to any such preferences,[250] flipping between such preferences, even relatively frequently, is unlikely to draw much criticism. But the constraints on, or expectations regarding, our moral attitudes, for instance, seem to be more exacting—we expect to offer and receive reasons when challenges are made, and flipping between opposing moral views on a regular

[249] Much will depend here on what we mean by 'implicit'. If we take it to mean 'always to be found in' or 'essentially connected with', then there is a case for the suggestion that reason-forming and reason-stating is not implicit in moral attitudes (indeed there may be no variety of attitudes in which it is). However, taken to mean 'suggested by, though not directly expressed', then moral attitudes look to be precisely of this variety.

[250] Pure gibberish, or the stating of some consensually absent feature, for instance, would not do.

basis is likely to be seen as troubling.[251] The use of 'hotly debated' moral issues, makes exposure to reason-stating and reason-forming seem less likely.

Secondly, the claim that on the first hypothesis the participant confabulates (Lopes 2014, p. 29), but on the second she does not (p. 30), needs grounding in some suitable definition of confabulation. But no such definition is offered. On some definitions of confabulation, participants may be seen to confabulate on both of these hypotheses (i.e. when taken in by the manipulation).[252] A workable model of Choice Blindness would require a working notion of confabulation with which to differentiate between confabulation and post-hoc rationalization. (I return to this issue in §6.)

Thirdly, further support would be required for the claim that reason-forming and reason-stating are systematically distorting of attitudes. To this end we are presented with studies that imply a sharp distinction between 'focusing on' and 'analyzing' an attitude (Lopes 2014, pp. 30–1). Analyzing—or reasoning more generally as the position appears to become—is *distorting of*, while 'focusing on' is *preserving of* attitudes. Whether or not such a sharp distinction can be supported, the conclusions of these studies cited are never so strong as to suggest any *systematic* distortion. (In summarizing one study, Lopes's describes the conclusion as follows: 'in certain circumstances giving reasons for an attitude reduces its consistency with behaviour' (p. 31).) And we have good reason to think that some varieties of reasoning could not be systematically distorting and still successfully perform the role for which they are deployed. (Hypothetical reasoning looks like it may be one such variety.)

Finally, the data does not appear to demand either of the hypotheses on offer. Neither is a particularly good fit. What the data might reasonably be taken to show is that, in the majority of cases, non-clinical participants (and so, perhaps, the population as a whole) willingly provide provably false statements

[251] This, of course, is not to suggest it is impossible to form an opinion on moral matters without doing a great deal of thinking.

[252] It is not clear how much weight rests on the distinction between confabulation and post-hoc rationalization for Lopes. Confabulation often has clinical implications, and thus may be thought to carry some stigma, but the distinction is far from clear.

about the reasons for their choices when queried, and when failing to notice those choices were manipulated. But a sizable minority, in all of the cases of Choice Blindness considered so far, is not shown to be willing to do this. A significant proportion of participants detect the manipulation in some way, and a proportion of those utter statements that are *true* of their original choice, and *not* true of the 'revealed' and manipulated choice. In the early Choice Blindness experiments (see Johansson and Hall et al. 2005), for instance, some 27% of participants were deemed to have detected the manipulation, while 11% *could not* have been offering confabulatory explanation for the manipulated choice since the reasons they offered (e.g. specific features of the selection) matched their original choice and not the 'revealed' one (see also Lopes 2014, p. 29).[253] These data are not irrelevant to the construction of a model that hopes to satisfactorily explain the effect, but neither hypothesis on offer provides the resources to explain what these participants are doing. (More naturally, what they are doing *right*.)

Together, these concerns suggest that a Lopes-style approach would require substantive revision if it were to be employed as a model of Choice Blindness. And while some features of Lopes's account of the effect have some explanatory benefits, I think these features would be better employed in a simpler model; a model that can still acknowledge that Choice Blindness research shows something interesting about our targeted variety of introspective failure, but which is less disruptive to our conceptions of reasoning and reasoners; one which provides more insight into the what occurs in detected manipulations; and is less concessive to more dramatic interpretations of the data.

[253] Lopes (2014) mentions these figures (see also §2 above), but either does not think them decisive in an analysis of the effect, or does not notice that the figures can be quite large.

6. Modeling Choice Blindness

On the basis of the discussion so far, we can outline several factors that should be taken into account for the construction of a model of Choice Blindness: (i) a tendency for introspective failure of one variety should not be taken to indicate a tendency for introspective failure more generally (or other varieties) without further argument (see §1); (ii) on the assumption that Choice Blindness is a unitary phenomenon, we should expect a uniform explanation—that is, a model should fit all attitudes and environments for which Choice Blindness has been putatively shown; (iii) if Choice Blindness is thought to demonstrate confabulation, we should have some view of confabulation against which to assess competing hypotheses; (iv) In the absence of further support for the claim that reasoning is *systematically* distorting and on the assumption that some attitude-distorting form of reasoning is in play, a plausible candidate for the variety of reasoning in question should be identified; and (v) a model of Choice Blindness must provide (or at least allow for) a plausible account of why some participants do, and some participants do not, succumb to the variety of introspective failure in question.

Before proposing such a model, it will be helpful to briefly discuss a matter at the heart of what the research is thought to show—namely, whether or not non-clinical participants willingly confabulate across a range of environments. Along the way, this discussion will allow for progress with the third (iii) desideratum of a model.

7. Confabulation

In the early 1970s Berlyne (1972) suggested that *confabulation*, somewhat like *delusion*, is a term that is 'widely employed, poorly defined and variously interpreted' (Berlyne 1972, p. 31). Some four decades later, the literature

sometimes steers clear of the 'thorny issue' of defining confabulation in favour of listing its common features (see e.g. Sullivan-Bissett 2015).[254]

Why, then, without a canonical definition of confabulation, is it remarkable to suggest that 'normal participants produce confabulatory reports' (Hall et al. 2005, p. 119)? One reason is that debates about confabulation often involve discussion of clinical syndromes where confabulation is likely to be found (e.g. split-brain, hysterical blindness) (Johansson et al. 2006, p. 675). Perhaps confusing the definiendum with common source of data, definitions of confabulation have often tended to follow the clinical theme and are 'usually defined as false narratives or statements about the world and/or the self that unintentionally arise due to some underlying pathological condition' (McVittie et al. 2014).

Accepting a definition of this variety has undesirable consequences. Combined with an acceptance of the Choice Blindness findings, it leaves us suggesting that a large majority of the population (perhaps around 70%) have some pathological condition. (This may be part of the motivation for an alternative hypothesis on which the manipulated participant is not confabulating.) But there is no need to see confabulation so tightly connected with pathology. The literature on confabulation has offered an alternative with roots going back at least a century (in Bonhoeffer 1901) that suggests a link between a failure, or a gap, in memory and a tendency to confabulate. Borrowing from this alternative, we might suggest that confabulations need only be understood as 'statements or actions that involve unintentional but obvious distortions of memory' (Moscovitch and Melo 1997, p. 1018; following Berlyne 1972). The view, in some form, is still in currency (see McVittie 2014) and allows for an understanding of confabulation without an explicit commitment to pathology.

[254] According to Sullivan-Bissett (2015) confabulatory explanations are: '(1) … false or ill-grounded; (2) are offered as the answer to a question; (3) have a motivational component; (4) fill a gap, and (5) are reported without any intention to deceive', although these are explicitly not intended to be necessary and sufficient conditions (p. 4).

Viewing confabulation in this way looks helpful in two respects: firstly, the definition can still make sense of clinical cases, because the pathological condition can be the cause of the failure, but it also makes room for non-clinical cases. Secondly, it looks a better fit for the Choice Blindness data, particularly with regard to the data unexplained by two hypotheses on offer—the fact that a significant proportion are not willing to utter false statements about manipulated choices. If I successfully recall some salient features, for instance, of the face I originally preferred, I do not offer my reasons for preferring the face now presented to me. I either talk about some features of my original choice incompatible with the manipulated choice, or I otherwise 'detect' the manipulation.

However, on this *de-stigmatized* view of confabulation, the successfully manipulated participant can be understood to be confabulating both on hypothesis one, and hypothesis two. For, in the latter case, the distortion of memory can be understood as the belief that one chose the presented, manipulated item rather than the original selection. If one does not form that belief, then one will not utter false statements about one's choice.

In short, the claim that on one hypothesis the participant confabulates, and on the other she does not, will only be true on some understandings of confabulation. On one common understanding of confabulation, it does not appear to be true.[255] On this understanding, confabulation involves a distortion of memory. And this helps to make sense of the neglected explanandum: the fact that a significant proportion of participants are not taken to utter provably false statements about their choices.

I will assume, for the purposes of this discussion, that something like this common, de-stigmatized, understanding of confabulation is the correct view (although this is not a position I will defend further here, and the understanding of Choice Blindness that I propose does not hang directly on that view). In the next section I consider two models of Choice Blindness.

[255] It appears, at least, that some further argument would be required, and this has not been offered.

8. Simple and dual process models of Choice Blindness

In §7 I suggested that a failure or distortion of memory helps to make sense of the Choice Blindness data. In proposing a model of Choice Blindness, then, we might proceed by first considering what I will call the Simple Model. This might be formulated as follows:

> **[SM*]** Choice Blindness can be explained in terms of a transition from a process of recall to a process that assesses the salient features of the subject's current environment.

While not uncontroversial, this proposal seems well founded given the discussion so far. Apart from the addition of the thought—both persistent in the literature on confabulation, and given independent plausibility by (see §7)—that some failure or gap in memory occurs in Choice Blindness, the proposal does not seem wholly incompatible with Lopes's view. This is combined with an apparent feature of the empirical findings: that successfully manipulated reports appear to draw on salient features of the subject's current environment rather than the environment as presented in the original choice.

Unfortunately, this formulation cannot explain the phenomenon alone since all it requires is that we move between two cognitive processes (one of recall, and one as yet unnamed). This is because it allows for instances in which a subject is aware of moving from one process to another, and this awareness looks unlikely to issue regularly in the kind of self-knowledge failure under discussion (it may well e.g. come up before or during the post-session debriefings as a variety of detection). What is required is that the movement

from one process to another goes undetected.[256]

> **[SM]** Choice Blindness can be explained in terms of an undetected transition from a process of *factual* recall to a process that assesses salient features of the subject's current environment.

This formulation of the Simple Model is more controversial than its predecessor in so far as it assumes that an undetected movement between cognitive processes is possible and plausible. It relies on what we might call—following the theme of ocular pathology—the *Process Blindness* assumption. A fairly radical formulation of Process Blindness might look like this:

> **[R-PB]** We have little or no access to all (or the vast majority) of our cognitive processes.

Something like this suggestion is familiar from Nisbett and Wilson's conclusions (1977, see §1) and similar theses have enjoyed support from the likes of Daniel Dennett (1969), Peter Carruthers (2011), and Hilary Kornblith (2012):

> The control of reflexes in man is subconscious, as are the stages of perceptual analysis, and in fact all information *processing*. We are not aware of the processes at all (as one might with suitable incisions and mirrors, be aware of one's digestive processes) … as Lashley says, 'No activity of the mind is ever concious' (Dennett 1969, p. 128)

And more moderately:

> It is … well-known that a very large part of the cognitive processes by which

[256] Alternatively, the 'movement' in question may be simply neglected by the subject as unimportant. (Thanks to Tom Stoneham in private correspondence.) This does seem like a genuine possibility. However, if such a movement were to consciously available to the manipulated subject, and especially in cases of purportedly more stable attitudes, one suspects that its relevance to any initial failure would become apparent in the debriefing, likely appearing as retrospective detection.

beliefs are produced is unavailable to introspection. (Kornblith 2012, p. 21)

So we might think that R-PB enjoys some independent plausibility. However, the plausibility that the assumption enjoys often rests upon interpretations of the kind of empirical research in question (as with Nisbett and Wilson 1977), and thus caution is required when employing such an assumption when attempting to explain seemingly related phenomena. Fortunately, while R-PB would—with little further argument—suffice to support the Simple Model, there is no need for such an extreme version of the thesis. A suitable version of the Choice Blindness assumption need only go so far as to say that access to cognitive processes is limited to the degree that a subject can, under a range of circumstances, fail to detect movement between two processes. And this assumption is modest, even when compared to some of the reservations expressed by those often held to have the most optimistic views of self-knowledge.[257] A moderate formulation of Process Blindness need only claim:

> **[PB]** Some cognitive processes are such that, from the first-person perspective, S's transition between those processes can go undetected by S at the time of the transition.

Because this falls within even optimistic assessments of our capacity for self-knowledge, a moderate Process Blindness assumption does not seem unwarranted and is not a high price to pay if the Simple Model affords some explanatory progress. Assuming the Simple Model has some initial plausibility, then, what remains is to assess the extent to which it *does* afford some explanatory progress. The answer, I suggest, is that it does not make enough.

Going in its favour, it does not require (as yet) that any general conclusions about introspective competence are reached from data that appears

[257] For instance, even Kant's reservations that, 'we can never, even by the most strenuous self-examination, get entirely behind our covert incentives' (Kant 1785/1997, §2, 4: 407, pp. 19–20) seem stronger than required for the purposes of the Simple Model, since it could be that under optimal conditions such processes (or any movement between them) is detected.

to make claims about a specific variety of self-knowledge failure (i); it is compatible with a view of confabulation that is persistent in the literature (iii); and (v) it offers a plausible account of why some subjects do, and some do not, succumb to the variety of self-knowledge failure in question—that is, it predicts that where a subject does not transition from one process to another, or that transition is detected, one should not utter provably false statements about the features of one's choice.

However, the proposal looks less secure in two respects. Firstly, while it offers what could well be a unified theory (ii), it appears to predict a change of attitude will occur *whenever* one slips without detection from one process to another. But there is little reason to think that assessing salient features of one's current environment will ordinarily 'overturn' one's prior attitudes leading to *de facto* confabulation as to one's reasons. Let us consider an example:

> I φ *that p* following some selection task (t_1), but under questioning (due to some failure or distortion of memory) I enter process *CP* and assess the features of my current environment salient to the question offering these features (t_2) as the reasons for my selection.

Among the numerous possible outcomes are the following: (a) I cease to φ *that p* and begin to φ *that q*; (b) I continue to φ *that p* throughout; (c) I cease to φ *that p* but quickly φ *that p* again; (d) I cease to φ *that p* but form no new attitude. Outcome (a) would show up as Choice Blindness (and is broadly in line with Lopes's explanation of the phenomenon, with the addition of an explicit role for memory). However, while (b), (c), and (d) SM, they are unlikely to result in Choice Blindness.[258] For example, (b) and (c) may come about when the salient features of one's environment are the same at t_1 and t_2 and process *CP* either provides no reason to cease φ*-ing that p* (b), or it potentially results in the

[258] In so far as they will not be found to be undetected manipulations potentially resulting in confabulation on the methodologies discussed here.

same attitude (c). In short, we can move between a process of recall and process that assesses our current environment without a change in attitude and without confabulation. The Simple Model then, is insufficient in that it over-predicts instances of Choice Blindness.

Another concern facing SM is that it falls short of identifying a plausible candidate for the variety of reasoning in question (iv) and this leaves us with little reason to suppose that the transition between recall and *CP* is a case of process blindness. Even though PB looks a reasonable assumption, it does not follow from it that one can move, without detection, between just *any* cognitive processes: for example, my 'memory experience' of putting my keys on the shelf and my thinking about where I usually leave my keys are both ways to respond to questions about the locations of my keys. They appear, however, at least as far as 'commonsense' psychology goes, to be first-personally distinguishable. It may be that factual memory is similarly distinguishable from a number of other processes. In the absence of a plausible candidate for the variety of reasoning in question, then, we may require further reason to suppose that SM will issue in plausible instances of PB.

The modifications I propose will help in both of these concerns. The proposal is as follows: the process into which the subject transitions (following the failure or distortion of factual memory), assesses features or factors that are salient to responding to an inquiry in a specific way—that is, with a view to resolving the inquiry *afresh* rather than with a view to responding via the identification of the subjects presently held attitudes. The proposal amounts to the claim that the *CP* is, at least in part, a *deliberative* process; aiming at resolving an issue presented to the subject via a consideration of those factors which, by the subject's own lights, settle the matter. This is as opposed to responding by considering any attitudes already in place. The assumption that we *ought* to be engaged in the latter rather than the former may help to explain why the results are thought to be striking. A modified version of the claim might look like this:

> **[DPM]** Choice Blindness can be explained in terms of an undetected transition from a process of *factual* recall to a process that, is at least in part, *deliberative*.

The modification to include deliberation allows for an understanding of the phenomenon in relation to an existing philosophical discourse regarding the *Transparency* of deliberation, and a related case of *Transparency* that has been the concern of much literature on self-knowledge in recent years (see e.g. Moran 2001; Byrne 2005, 2011; Boyle 2009, 2012; Gertler 2011b). However, 'deliberation' used in this sense is philosophical jargon and will require further explanation.

9. The dual process model, deliberation, and transparency

On the Dual Process model I am proposing, we can make sense of Choice Blindness by reference to these two components. The proposal is that Choice Blindness can be understood in terms of an undetected[259] transition from a process of *recall* into a process of (or partially constituted by) *deliberation*. Before attempting to support the proposal, I will explain what is meant by deliberation in this context.

Philosophical accounts of *deliberation* present it has having a range of features. David Owens' (2011) account, for instance, suggests that deliberation is a conscious activity that aims at resolving an issue (see Owens 2011, p. 262). It is (a) conscious in that it 'occupies the deliberator's attention; it is (b) an

[259] Why undetected as opposed to detected but judged to be irrelevant? One may respond in this way: many of the participants were 'utterly surprised' when they discovered their choices had been manipulated. And 84% suggested they 'would have noticed if they had been presented with mismatched outcomes' (Hall et al. 2006, p. 699) in the debriefing. Participants who detected a move away from recalling their reasons for a choice into some other activity should likely to be surprised, retrospectively, since they can attribute their being 'taken in' to the to the mistaken judgement that such a shift would be irrelevant. However, I find the idea that we *routinely* distinguish between our ongoing mental processes quite fanciful, and it looks likely that our abilities in this respect are fairly restricted. Thank you to Tom Stoneham for pressing me on this point.

activity in that the deliberator is 'trying to do something: prove a theorem, to make a decision, and so forth' (p. 262); and it is (c) directed at the resolution of some issue or question. Typically, it does this by (d) focusing on features of the world rather than on psychological concepts (p. 262). So, when we deliberate, we do not typically focus on anything inner or psychological, such as previously or currently held attitudes. Two of these features, (a) and (d), will benefit from further comment.

9.1 Deliberation as conscious activity

Why might we think that deliberation is a distinctively conscious activity? One suggestion is that there is a finite resource, 'conscious attention', and that deliberation is the kind of activity that takes up a proportion of that resource (Owens 2011). But this will not directly lead to the conclusion that deliberation is a conscious activity, since there is little reason to suppose that non-conscious activity cannot effect conscious processing, and good anecdotal evidence to suggests it can: emotional upheaval, positive or negative (e.g. grief, loss, joy); fairly straightforward stress and fatigue; and complicated life events (e.g. trauma), appear to take their toll on our cognitive performance, yet we do not suppose we are always conscious of their presence whenever they do.

It is clear that decisive elements in human reasoning are not always conscious or introspectively available (see e.g. Kahneman's 2011 'Two Systems' approach[260]), and it is not always better when it is (see e.g. Kornblith 2012; Strick et al. 2011; Ghiselin's 1952).[261] And distinguishing between conscious

[260] Daniel Kahneman's 'two systems' model contrasts automatic, quick, effortless thinking with no sense of voluntary control (System 1) and effortful and demanding mental activities 'associated with the subjective experience of agency, choice, and concentration' (System 2) (see Kahneman 2011, pp. 20f.) is becoming a standard way of referring to markedly different thought processes (see Boghossian 2014). But the approach is not without its problems. Paul Boghosssian (2014), in discussing how to characterize 'inference', suggests that some fairly common varieties of thought seem to lie somewhere between System 1 and System 2, and suggests that perhaps we are interested in reasoning that is 'System 1.5 and up' (pp. 2f.).

[261] Empirical research into unconscious thought effect (UTE) suggests that 'unconscious thought leads to better complex decisions than conscious thought' (Strick et al. 2011, p. 738). Brewster Ghiselin's illuminating (1952) anthology of self-reports from offers accounts of philosophers, scientists, mathematicians, and novelists (etc.) suggests that some striking discoveries were made without conscious intervention. For example, Henri Poincaré recalls: 'At the moment when I put my foot on the step the idea came to me without anything in my former thoughts seeming to have paved the way for it, that the

reasoning and other varieties is no simple matter (Boghossian 2014),[262] with even ostensibly conscious decision-making processes prone to infection from 'relatively unconscious' influences such as implicit bias (Brownstein 2015).

None of these factors demonstrates conclusively that *deliberation* is not a conscious activity, but it is enough to favour a characterization that does not come with this explicit requirement. Here is one such characterization:

> Deliberation of any kind is framed by a question, whether it is what to do, what to believe, what to pretend, or whatever. This does not mean that an agent has to have the question at the forefront of his mind, explicitly posing the question to himself, as it were; but unless his thinking manifests some recognition that this is the question that he is striving to answer, his stream of thought would lack the direction or purpose required for it to be an instance of deliberation about what to do or believe rather than, for example, a stretch of directionless cogitation. (Shah 2003, p. 466)

For the purposes of the discussion, then, we will take deliberation to be (a) an activity, (b) aimed at resolving an issue, (c) which manifests some recognition of a question requiring resolution, and that (d) deliberative question are typically transparent to other considerations (e.g. to factual inquiry).

9.2 Transparent deliberation

Deliberative questions are, or can be, *transparent* to other considerations—that is, engaging with the deliberative question, typically results in us doing so by engaging with some other consideration by which the deliberative question is settled. Take, for example, the deliberative question of whether to believe

---

transformations I had used to define the Fuschian functions were identical with those of the non-Euclidian geometry. I did not verify the idea … I went on with a conversation already commenced, but I felt a perfect certainty. On my return … for conscience' sake I verified the result at my leisure' (Ghiselin 1952, p. 26). The volume (and this example) is mentioned in Nisbett and Wilson (1977)

[262] Paul Boghosssian (2014), in discussing how to characterize 'inference', suggests that some fairly common varieties of thought seem to lie somewhere between System 1 and System 2, and suggests that perhaps we are interested in reasoning that is 'System 1.5 and up' (pp. 2f.).

something is the case (doxastic deliberation). In doxastic deliberation, the deliberative question *whether to believe that p*—inevitably 'gives way' to factual inquiry—that is, the factual question *whether p*, because 'the answer to the latter question will determine the answer to the former' (Shah and Velleman 2005, p. 499). Because answering the question *whether p* settles the question *whether to believe that p*,[263] whenever we engage with the latter question, there is a *slip* or *collapse* into the former into considerations that speak to the former question. Not all deliberative questions are transparent to factual inquiry—contrast *whether to believe that p* with *whether to suppose that p* or *whether to imagine that p* (see p. 499). However, it is enough for the present purposes that they typically are.

On the assumption that deliberative reasoning has the characteristics above, it is clear why engaging in such reasoning would be potentially disruptive to pre-existing attitudes when deployed. Deliberative reasoning is not in the business of preserving attitudes in place prior to its deployment. It is in the business of resolving an issue (afresh) by focusing on what the subject takes to be the facts that bear upon the matter at hand. Since deliberation is blind to current attitudes, deliberative success comes at the risk of attitude distortion, most obviously where there is has been some change in the individual or environment since, in either case, decisive factors in the reasoning process may have been altered.

Before explaining how these modifications resolve the concerns above, I will consider two possible objections to the view that a form of deliberative reasoning might be in play.[264]

[263] Just how it is meant to settle the question is open to debate. Shah and Velleman (2005) suggest the best explanation is that 'the very concept of belief includes a standard of correctness, to the effect that a belief is correct if and only if it is true' (pp. 499f.). For alternative explanations see, for example, Steglich-Petersen (2008) who suggests that transparency can be explained 'by the aim one necessarily adopts in posing that question' (p. 546), and Sullivan-Bissett (2014) who explains transparency by appealing to 'causal facts about belief formation which obtain in virtue of natural selection selecting for mechanisms which produce beliefs with true contents' (see Ch. 5).

[264] Thanks are due to Paul Noordhof for helpful correspondence on these issues.

10. Objections to the dual process model

An opponent of the proposed view may object on the following two grounds: firstly, that deliberation is not a possible candidate for the role as described; secondly, on the grounds that deliberative reasoning as described is not distinct from the factual memory process—that is, the features as described are shared by both processes.

The first objection can be expressed as follows: the formation of belief is the terminus of deliberative activity. One cannot at the same time believe that *p* and deliberate over whether *p* (see e.g. Owens 2011). Successfully manipulated subjects in Choice Blindness experiments falsely believe that the option presented to them is in fact their (original) selection, but they believe it none the less. Because they believe that the option presented to them is their (original) selection, they cannot deliberate over the features that speak in favour of one selection over another. The concern is founded in an understanding of deliberation that is already present in the literature, for example:

> belief and intention both act as a block on further deliberation. Suppose I am convinced of the honesty of my accountant. To have such a belief is not just to think that the evidence currently favours his honesty: that would be consistent with having an open mind on the question, with carefully collecting and assimilating further information and being thoroughly on one's guard. Believing my accountant to be honest, I simply don't consider whether a certain anomaly in the company' s books should undermine my faith in him: I ignore it or explain it away on the assumption that he is honest. (Owens 2011, p. 262.)

If the objection succeeds, it would clearly be damaging to the current proposal, and so it should not be dismissed without consideration. However, it rests upon two assumptions that should be treated with caution. The first, of course, is already made explicit: the formation of a *belief that p* prevents deliberation over

whether p. The second is that subject forms a specific variety of attitude upon viewing a choice presented as her own, and that attitude is *belief*. (Lopes 2014 appears to make the same assumption.) With regard to the first assumption a response might point to clear support in the literature for idea that we can 'critically reason' about our beliefs (Burge 1996); that we can reflect upon them (Kornblith 2011);[265] and—most directly (Crane 2014)—that in some cases one deliberates upon what 'one already believes':

> This can occur when one is trying to work out what one believes, or remember some fact, or draw out some consequence of what one believes … this isn't a case of forming a belief, but rather a case of revealing to oneself what one believes anyway. (Crane 2014, p. 277)

So responding to the objection by denying (or bringing into question) the first assumption is one response. However, one can also remain committed to the view that belief blocks deliberation, and still disarm the objection by casting doubt upon the validity of the second assumption. We can do this by considering whether the experimental data suggests a belief that the presented choice is the subject's own is formed at the appropriate time to prevent deliberation (i.e. when presented).

What the experiments clearly show is that subjects are able (and willing) to cite features of the choice (presented to them as their own) upon request, whether or not those features belong to their original selection. But this ability alone requires neither that such a belief is formed initially, nor the belief that the choice was their own. One can, for example, list factors in favour of one holiday destination having eventually chosen another (cf. §2). Of course, we may take the data to show something more than this. The fact that some subjects argue 'unequivocally' for the presented moral proposition (see Hall et al. 2012, p. 4), for example, may be seen as indicating a strong pro-attitude towards that

[265] Hilary Kornblith (2012) does not think that this is a very valuable practice, but doesn't appear to question whether it is possible.

proposition. But even this requires only that a strong pro-attitude formed at some relevant point during the subject–experimenter discourse—and this could be either at the *outcome* of deliberation or at its *outset*. So there is still no requirement that a belief that the presented choice is the subject's own is formed upon presentation on this reading of the phenomenon, and there is still *room* for deliberation. (Indeed, in the moral case, the idea that pre-existing attitudes replaced by opposing attitudes purely by the suggestion that one chose otherwise, and without some such intervening process seems baffling. The fact that *some* deliberation occurs partially explains how a change of heart over such prominent issues could occur.) What seems plausible, however, is that in order for the subject to engage in confabulation with regard to a presented choice, she needs to be in some respect open to that choice being *her* choice, and that this openness may evince a (loosely) pro-attitude to the choice being her choice *at the point of presentation*. However, even if it is true that one cannot *believe that p* and deliberate over *whether p*, one can, for example, *suspect that p* and deliberate over *whether p*. More generally, it is far from obvious that some form of 'prima facie view' would be sufficient to block deliberative activity. Anne's cognition *Lo, distant Indian elephant* set against the background of her knowledge that she is in South Africa, of zoos, and of colonial history, does not prevent her from investigating whether it is, in fact, an Indian elephant that she sees, or a rather young African elephant with small ears.

The first objection, then, can be disarmed by denying that belief blocks deliberation, by questioning the timing of the belief formation, or suggesting that the maximum explanatory requirement is for some 'prima facie view' that seems unlikely to prohibit deliberative activity. Since the latter two responses are compatible and stand independently of whether the thesis that belief prohibits deliberation turns out to be correct, I favour a combination of these two (though not much will rest on it for the purposes of the paper).

The second objection questions the distinction between the two processes on the basis of the features attributed to deliberation (see §8). If the

processes can be shown to have the same salient features, then the model has failed to deliver a cognitive process capable of explaining the apparent disruption of a subject's attitudes following inquiries into the reasons for her choice. Section eight (§8) concluded in the suggestion that deliberation ought to be seen as (a) an activity, (b) aimed at resolving an issue, (c) which manifests some recognition of a question and that (d) deliberative question are typically transparent to other considerations (e.g. to factual inquiry). But, as the objection goes, these features look like fine candidates for the role that factual recall plays in the model. And clearly factual recall must meet most, if not all, of these criteria: it is (a) an activity; (c) that manifests some recognition of a question—at least when it is engaged in response to the right kind of stimulus (e.g. an explicit request for information) and (d) the subject need consider no psychological intermediary in order to respond to such stimuli—a request for information about the capital of Austria need to focus on nothing more than the salient facts about Austria. So for the objection to cause problems for the proposed model, the residual issue—whether (b) factual recall aims at resolving an issue—will be decisive. If factual recall exhibits the very same features as deliberation, the proposed model tells us little about why the variety of self-knowledge failure at issue appears to be so prevalent in the Choice Blindness data.

Factual recall, at least typically, does not aim at 'resolving an issue' in the sense intended here. But it will be helpful to examine one reason for thinking that it does. One might think it does if one subscribes to a particular view of factual memory. On this view, memory knowledge is possible because we preserve our reasons (or evidence) for thinking that something is the case. When challenged about the reasons for thinking that 'Rosa Parks defied the bus driver', we are able, therefore, to 'reconstruct' our reasons for thinking that she did, and our 'memory belief' is thereby justified. If one holds this view of memory, one might be tempted to see the process of factual recall—at least when employed in response to certain kinds of stimulus—as a process of

'reconstructing' a kind of mental argument or a balancing of available evidence, which terminates when one reaches a conclusion. If one did have such a view in mind, one would have to think that factual memory exhibits all of the features attributed to deliberation. However, the account of memory on which the view depends is largely discredited (as we will see in §10).

What the second objection successfully does is to highlight both the need for more clarity on what is meant by factual memory and an apparent similarity between the relevant features of deliberation and factual memory. In the next section I defend a view of factual memory that resists the second objection. I also discuss how the respective features of deliberation and factual memory go some way to explaining how the transition between the two processes can go undetected in certain circumstances, thus contributing to the likelihood of self-knowledge failure.

11. Preservationism and deliberation

The term 'memory' is used to label a heterogeneous variety of phenomena (Sutton 2012; see also Byrne 2010) especially when referred to obliquely, in ordinary language, via the notion of 'remembering':

> I remember how to play chess and how to drive a car; I remember the date of Descartes' death; I remember playing in the snow as a child; I remember the taste and the pleasure of this morning's coffee; I remember to feed the cat every night. (Sutton 2012)

All of the uses above carry an implication of success, but for some that success is directly related to truths about states of affairs. These are sometimes divided into *episodic* and *semantic* memory: I only *remember* walking along the beach if I walked along the beach (*episodic*), and I only remember that 'the sand was wet'

if the sand was wet (*semantic*). But both the monolithic status (see e.g. Byrne 2010), and the 'truth-directedness' of these forms of memory have been the subject of some scrutiny: some putative instances of episodic memory—instances of 'observer perspective'—involve a subject seeing herself 'in the remembered scene' rather than from her 'original point of view' (Sutton 2010, p. 27), and so there is a mismatch between the original and remembered experience. It has been suggested more generally that episodic memory functions to preserve or protect a coherent picture of the self (Conway 2005) rather that than to correspond with any actual events; and even that there are *no* strictly accurate autobiographical memories (e.g. Conway and Loveday 2015).[266]

For the present purposes I will use another distinction to highlight some important differences between what I have called 'factual memory'—the variety of memory whose success conditions are directly related to states of affairs—and 'memory experiences'. Memory experiences are phenomenologically rich and include objects or events with which one has some past acquaintance (Teroni 2015): my memory experience of *Ein Deutsches Requiem*, for instance, consists in part of a 'preserved' acquaintance or 'cognitive contact' (Byrne 2010) with some elements of that composition. In contrast, factual memory, which preserves propositional content (see e.g. Owens 1999) is 'phenomenologically poor' (Teroni 2015): there is not a great deal that it is like to recall the facts like 'Brahms composed *Ein Deutsches Requiem*'. In contrast to many cases of 'remembering' above, and to memory experiences, factual memory barely registers as 'memory' at all (we often refer to instances simply as 'knowledge' or 'facts').[267]

Factual memory does not involve memory experiences (see e.g. Teroni, 2015), even though it is tempting to take the fact that memory experiences sometimes accompany factual memories as indicating some epistemic role, such

[266] Space does not permit an in-depth discussion of the positions and arguments. However, they are helpful in pointing to some of the long-standing and current debates in literature on memory.
[267] Asher Koriat highlighted the matter in conversation in Grenoble 2014.

that memory experiences provide our evidence, or justification, for our factual memories (the view I dismiss in §9). If this were correct, then the majority of our current beliefs would not be justified since, as David Owens notes: [268]

> We have probably forgotten why we adopted many of our current beliefs and even if we could dredge the evidence for them up from memory, we couldn't do this for more than a tiny subset of our beliefs at any one time. (Owens 1999)

On an alternative view, factual memory preserves the rationality of a belief (Owens 1999, pp. 319f.), regardless of whether we recall our original justification (Stoneham 2006). It thus has prima facie epistemic authority: 'we are entitled to persist in believing something remembered providing nothing comes to our notice which should make us desist' (Owens 1999, p. 319). And since *believing that p* is to have 'finished inquiring into *p* by forming the view that *p*' (p. 317) relinquishing the reasons for our attitudes is permissible, and arguably the norm:

> Once a question is decided, we close the books on it and throw away the evidence: deliberately retaining evidence for future consultation is a sign of doubt, an attitude appropriate to the scientist who is interested in the likelihood of various things and has a professional obligation to suspend judgement but quite unsuited to the everyday believer. (Owens 1999, p. 317)

On this view, factual memory has the following features: (i) has prima facie epistemic authority, which (ii) allows us to relinquish the reasons for attitudes, and (iii) it is phenomenologically poor. With regards to the disputed feature of memory—whether (b) factual recall aims at resolving an issue (i.e. from the

[268] We sometimes cite information from memory experiences when offering reasons for our beliefs. Fabrice Teroni (2015) argues that memory experiences can play an important (non-epistemic) explanatory role in some cases.

second objection from §9)—a (memory) belief is the result of an inquiry that has been settled, rather than one that requires resolution.

If this account of memory is correct,[269] the explanatory contribution made by an explicit role for factual memory is clear. Choice Blindness subjects form an attitude at the point of initial selection for which the reasons may or may not be retained (and in typical will not be). So requesting reasons for a selection already places the subject in a difficult situation. In cases where the choices are manipulated so that the opposite of their selection is presented, this difficulty is compounded since the subject now has reason to question her original selection. The combination of a gap in or distortion of memory, combined with a reason to doubt her original selection means that responding to a request for reasons will require further consideration of features that speak in favour of the presented selection. And one method of acquiring these reasons would be to *re-open* (or persist with) the inquiry based upon those features. An appropriate variety of inquiry—one that would provide the 'direction or purpose' to the 'stream of thought' (Shah 2003; see §8) and would allow for a feeling of ownership over the outcome[270]—would be one characteristic of deliberation as described above. The characteristics of factual memory—and in particular its phenomenological paucity—also lend credence to the claim that a move between a process of recall and a process of deliberation can go undetected.

Over the last two sections, we have removed a number of obstacles to thinking that deliberation could—at least in part—play a role in subjects' responses to requests for their reasons. We have also seen that deliberation is the right kind of process in terms of shaping the direction of our thought processes in attempting to respond to a gap in or distortion of memory. And in

[269] Some version of this position is supported by, for example, Tyler Burge (1993); Michael Dummett (1994); David Owens (1999), but is not without its detractors (see e.g. Lackey 2007).

[270] If we are to take seriously the idea that the subjects form an attitude in favour of the new selection (as evidenced by their apparent willingness to argue 'unequivocally' for it), then merely listing some features of the presented selection will not do. The subject will need to feel that they are the author of thoughts, or at feel able to endorse them (in other words to be able to accept the reasons as her own). However, space prohibits extensive discussion of this issue.

§8 we noted that deploying deliberation can result in the distortion or substitution of attitudes, most obviously where there is has been some change in the individual or environment (with a change in the environment being a key element of the Choice Blindness methodology). What remains is to further the positive case for the deployment of deliberation in subjects questioned about their own attitudes.

12. Transparency and self-knowledge

In this section I will further the case for the proposed model by considering the question of why a subject in conditions such as those present in Choice Blindness experiments might engage in deliberation in response to an inquiry into her own attitudes. The answer I propose is that engaging in deliberation is a commonplace, though not always intended, occurrence in ordinary subjects' attempts to essay their attitudes, and that conditions such as those experienced in Choice Blindness experiments simply its likelihood.

A good deal of self-knowledge literature over the last decade or so[271] has been preoccupied with a case of transparency that is related to, but distinct from, the transparency of doxastic deliberation discussed above (see §8). The literature follows remarks by Gareth Evans (1982), among others, with regard to the possibility that we can come to know our own minds via world-directed inquiry. Evans's own remarks refer largely to self-knowledge of belief,[272] although more recent attempts (e.g. Boyle 2009; Byrne 2011) discuss the possible application of the idea to a variety of states and attitudes, such as desire, intention, and even pain (see e.g. Byrne 2011, p. 213).[273] Transparency

[271] See, for example, Moran (2001); Byrne (2005, 2011); Boyle (2011); and Fernandez (2013).

[272] A frequently cited remark is: 'In making a self-ascription of belief, one's eyes are, so to speak … directed outward—upon the world. If someone asks me "Do you think there is going to be a third world war?," I must attend, in answering him, to precisely the same outward phenomena as I would attend to if I were answering the question "Will there be a third world war?" (Evans 1982, p. 225).

[273] 'I know that I feel a pain in my elbow, not by attending to myself, or my own mind, but by attending to the painful disturbance in my elbow. (My elbow hurts: hence, I feel a pain in my elbow.)' (Byrne 2011, p.

accounts of self-knowledge come in markedly different styles but, by way of example, Alex Byrne's (2011) 'Gallois-style' doxastic schema sees a would-be self-knower engaged in an '*inference from world to mind*' (p. 203):[274]

p

I believe that *p*

Transparency accounts are taken to explain at least one important feature of self-knowledge in that they mark a contrast between the first-person case of coming to know one's mind and the standard way of coming to know the minds of others. For assuming one can and does approach questions of one what one believes via some transparency method—by considering the matter without 'essential reference' to oneself or one's belief (see Edgley 1969, p. 90; in Moran 2001)—one tends not to relate oneself in this way to the 'question of what someone else believes' (Moran 2001 p. 60), or at least approaching it in that way would often lead one astray (see Byrne 2011).

But while transparency accounts are successful in at least this respect,[275] a number have run into a common difficulty: questions of the variety *Do you think that p?* can be read in more than one way. They can be read as an invitation to deliberate or 'make up one's mind'[276] about *p*, or as a question about whether one *already thinks that p*. In the former case, treating the question *Do you think that p?* much as one would treat the question *Is p true?* is well and good. However, clearly we do not wish to *make up our minds* upon each attempt to respond to inquiries about what we think. More to the point it seems clear that we don't. And such a strategy would be risky for cases in which one wishes to gauge what one *already thinks*. For if there is a tendency for one question to

213).

[274] The step is unpopular among non-inferentialists (e.g. Boyle 2011) and inferentialists (e.g. Cassam, forthcoming) alike. Matthew Boyle suggests 'only a madman could draw such inferences' (2011, p. 227) while Quassim Cassam suggests they are 'patently invalid' (forthcoming). Byrne recognizes the problem and offers a 'partial' response (see e.g. 2011, pp. 204ff.).

[275] Alex Byrne (2011) takes the method to explain what he calls 'peculiar access', but also argues that it explains why self-ascriptions of belief place us in an epistemically 'privileged' position.

[276] Cassam (forthcoming)

collapse into the other,[277] then the subjects response must be 'brute' or 'spontaneous' (Shah and Velleman 2005, p. 506) to be reliable in that respect, since 'reasoning aimed at answering the question *whether p* … would contaminate the result by possibly altering the state that one is trying to assay' (p. 507).

The implications of this difficulty for transparency views of self-knowledge, in general, are debatable. Some argue that the disruptive potential of the deliberative process for attitudes in place prior to the initiation of a self-knowledge procedure leave an important feature of self-knowledge—a kind of epistemic security—unexplained (see Gertler 2011),[278] while others are less concerned with explaining the supposed epistemic features of self-knowledge in the first place (see e.g. Moran 2001). For the current discussion, the issue of primary concern is that if transparency accounts of self-knowledge are broadly descriptive of how we come to know our minds, we have good reason to think that deliberation is a commonplace, if not always intended or detected, occurrence in ordinary attempts to respond to inquiries into our attitudes.[279] And this can present a challenge whenever one attempts to assay an attitude in place prior to the self-knowledge procedure. For, unless one's responses are brute or spontaneous, the process will risk contamination by becoming, even in part, a case of making up one's mind. It is a short distance from here to imagining how one might come to offer factors or features that surface during deliberation among the reasons for a previously expressed attitude.

In the discussion so far, we can see how a subject might engage in deliberation even when ordinarily attempting to answer questions about her own attitudes. We can also see a number of ways in which the risk of slipping to into a (partially) deliberative process—as opposed to a purely recollective

[277] Edgley (1969) takes the questions 'Do I think that P?' and 'Is it the case that P?' to be first-personally indistinguishable, but this is stronger than what is required for our present purposes.
[278] Brie Gertler (2011) offers a more in-depth discussion of the problem that has been possible here, which applies to both Alex Byrne's and Jordi Fernandez's transparency accounts of self-knowledge.
[279] This is not to suggest it is ubiquitous, although the question of exactly how reliable transparency procedures are lies outside of the scope of this paper.

process—will be exacerbated by the Choice Blindness set-up and result in the attribution of self-knowledge failure. Firstly, responding to requests for the decisive factors or features in one's decision-making process assumes of the subject a degree of access to that process that is at best controversial—there is a significant risk of a memory gap or failure in that respect. Secondly, the presentation of a choice other than the one originally selected is enough to defeat the prima facie authority of memory, thus reopening what may well have been a previously shut case. Thirdly, a change in salient features of the environment (via 'manipulated' choice) will register as self-knowledge failure whenever the subject is drawn into considering (and refers to) the features of the presented, rather than original, selection.

Conclusion

Existing explanations of a specific variety of self-knowledge failure—exemplified in Choice Blindness research—fail in a number of respects: they fail to explain the broad spectrum phenomena that appear to fall under the category, and are often maximally mutilating to our conception of humans as introspectively competent rational decision makers. Focusing on Choice Blindness research, I have proposed and defended an alternative model that highlights the roles of memory and deliberation in the production of confabulatory reports. While the model relies on an assumption, it is one that few have challenged, and even optimists about introspective reliability accept: that our access to our cognitive processes is incomplete. As long as we are willing to accept that assumption, we have a model of Choice Blindness, and potentially other varieties of self-knowledge failure, that allows us to concede vulnerability to self-knowledge failure while avoiding more alarming conclusions about our cognitive abilities and rational status.

Bibliography

Alloy, L. B. and Abramson, L. Y. (1988) 'Depressive Realism: Four Theoretical Perspectives', in L. B. Alloy (ed.), *Cognitive Processes in Depression*, New York: Guilforf.

Aristotle (1985) *Nichomachean Ethics*, (trans.) Erwin, T., Indianapolis, IN: Hackett.

Armstrong, D. M. (1993) *A Materialist Theory of Mind*, New York: Routledge.

————(1981) *The Nature of Mind and Other Essays*, New York: Cornell University Press.

Ayer (1956) *The Problem of Knowledge*, Harmondsworth: Penguin Books

Ball, P. (2016) 'The Tyranny of Simple Explanations' [Online], *The Atlantic*, 11 August 2018. Available from: <http://www.theatlantic.com/science/archive/2016/08/occams-razor/495332/> [14 October 2016].

Bar-on, D. (2004) *Speaking My Mind: Expression and Self-Knowledge*, Oxford: Clarendon Press.

Benson, B. (2016) 'Cognitive Bias Cheat Sheet: Because Thinking is Hard', *Better Humans* [Online], 1st September 2016. Available from: <https://betterhumans.coach.me/cognitive-bias-cheat-sheet-55a472476b18#.df5op2i9a> [18 March 2017].

Berlyne, N. (1972) 'Confabulation', *British Journal of Psychiatry*, 120, 31–39.

Bermúdez, J. L. (2013) 'Immunity to Error Through Misidentification and Past-Tense Judgements', *Analysis*, 73 (2), 211–20.

————(forthcoming) 'Memory and Self-Consciousness', in Bernecker, S. Michaelian, K. (eds.) *Routledge Handbook to the Philosophy of Memory*. Routledge.

————(2012) 'Memory Judgments and Immunity to Error Through

Misidentification', *Grazer Philosophische Studien*, 84, 123–42.

Boghossian, P. (2014) 'What is Inference?' *Philosophical Studies* 169 (2014), 1–18.

Bonhoeffer, K. (1901) *Die akuten Geisteskrankzeiten der Gewohnheitstrinker*. Jena: Gustav Fischer.

Boyle, M. (2011) 'Transparent Self-Knowledge' [Online], *Proceedings of the Aristotelian Society, Supplementary Volume*, 85 (1). Available at: <http://nrs.harvard.edu/urn-3:HUL.InstRepos:4879167> [31 October 2012].

————(2009) 'Two Kinds of Self-Knowledge', *Philosophy and Phenomenological Research*, 78 (2009), 133–64.

Brown, J. (2000) 'Critical Reasoning, Understanding, and Self-Knowledge', *Philosophy and Phenomenology Research*, 61 (3), 659–76.

Brownstein, M. (2015) 'Implicit Bias', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition). Available at: <http://plato.stanford.edu/archives/spr2015/entries/implicit-bias/> (24 November 2015).

Burge, T. (1996) Our Entitlement to Self-Knowledge: I. in Tyler Burge and Christopher Peacocke 'Our Entitlement to Self-Knowledge'. *Proceedings of the Aristotelian Society, New Series*, 96 (1996), 91–116.

————(2011) 'Self and Self-Understanding Lecture II: Self and Constitutive Norms', *The Journal of Philosophy*, CVIII (6/7), pp. 316–38.

————(1993). 'Content Preservation'. *Philosophical Review*, 102, 457–88.

Burgess, P. W. (1996) 'Confabulation and the Control of Recollection'. *Memory*. 4 (4), pp. 359–411.

Burt, C. D. B., Kemp, S. and Conway, M. (2001) 'What Happens if You Retest Memory 10 Years On', *Memory and Cognition*, 29 (1), 127–36.

————(2003) 'Themes, Events, and Episodes in Autobiographical Memory', *Memory and Cognition*, 31 (2), 317–25.

Byrne, A. (2012a) Review of 'The Opacity of Mind: An Integrative Theory of Self-Knowledge'. Notre Dame Philosophical Reviews [Online]. Available at: <https://ndpr.nd.edu/news/30799-the-opacity-of-mind-an-integrative-theory-of-self-knowledge/> [18 September 2014].

————(2012b) 'Knowing What I See', in *Introspection and Consciousness*. eds. D. Smithies and D. Stoljar. Oxford: Oxford University Press.

————(2011a) 'Transparency, Belief, Intention', *Proceedings of the Aristotelian Society Supplementary Volume* 85 (2011), 201–21.

————(2011b) 'Knowing What I Want', in J. Liu and J. Perry (eds.) *Consciousness and the Self: New Essays,* Cambridge: Cambridge University Press.

————(2011c) 'Knowing That I am Thinking', in A. Hatzimoysis (ed.) *Self-Knowledge,* Oxford: Oxford University Press.

————(2010) 'Recollection, Perception, Imagination', *Philosophical Studies*. 148, 15–26.

————(2005) 'Introspection', *Philosophical Topics*, 33 (1), 79–104.

Carruthers, P. (2011) *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, Oxford: Oxford University Press.

————(2015) *The Centred Mind*, Oxford: Oxford University Press.

Cassam, Q. (2014) *Self-Knowledge for Humans*, Oxford: Oxford University Press

————(2009) 'The Basis of Self-Knowledge', *Erkenntnis*, 71 (1), 3–18.

————(2007a) 'XIV – Ways of Knowing', *Proceedings of the Aristotelian Society*, 107 (2007), 339–58.

————(2007b) *The Possibility of Knowledge*, Oxford: Oxford University Press.

————(2003) 'Can Transcendental Epistemology be Naturalised', Philosophy, 78 (2), 181–203.

————(forthcoming) 'Evans on Self-Knowledge', in Misselhorn, C. (ed.) *Sprache, Wahrnehmung und Objektivität: Neue Perpektiven auf die Philosophie von Gareth Evans*.

Chisholm, R. (1981) *The First Person*. Minneapolis: University of Minnesota Press.

Conway, M. (2005) 'Memory and the Self', *Journal of Memory and Language*, 5 (2005), 594–628.

Conway, M. and Loveday, C. (2015) Remembering, Imagining, False Memories and Personal Meanings, *Consciousness and Cognition*, 33 (2015), 574–81.

Clarke, A. and Chalmers, D. J. (1998) 'The Extended Mind'. *Analysis*. 58, 10–23.

Crane, T. (2014) 'Unconscious Belief and Conscious Thought', *Aspects of Psychologism*. Cambridge, MA: Harvard University Press.

Davidson, D. (1973) 'Radical Interpretation', *Dialectica*, 27 (3–4), 313–28.

Davies, R. A. (2015) 'Refining Our Understanding of Choice Blindness'. Epistemic Innocence. Blog post: 19 May 2015. Available at: <http://imperfectcognitions.blogspot.co.uk/2015/05/refining-our-understanding-of-choice.html> [23rd September 2015].

Descartes, R. (1998) *Discourse on Method*, (trans.) Cress, D. A., Indianapolis, IN: Hackett.

Descartes, R. (n.d.) *The Passions of the Soul*, Bennett, J. (trans.)

Douglas, H. (2013) 'The Value of Cognitive Values', *Philosophy of Science*, 80 (5).

Dummett, M. (1993). 'Testimony and Memory', *In The Seas of Language*, Oxford: Oxford University Press.

Dunning, D. (2014) 'We Are All Confident Idiots', *Pacific Standard* [Online], 27nd October 2014. Available at: <https://psmag.com/we-are-all-confident-idiots-56a60eb7febc#.4f5nbbtsm> [2nd January 2017].

Evans, G. (1982) *Varieties of Reference*, Oxford: Oxford University Press.

Fernández, J. (2013) *Transparent Minds: A Study of Self-Knowledge*, Oxford: Oxford University Press.

Fotopoulou, A., Jenkinson, P. M., Tsakiris, M., Haggard, P., Rudd, A. and Kopelman, M. D. (2011) 'Mirror-view Reverses Somatoparaphrenia: Dissociation between First- and Third-Person Perspectives on Body Ownership', *Neuropsychologia*, 49 (14), 3946–55.

Fricker, E. (2000), 'Special Access Versus Artefact of Grammar—A Dichotomy Rejected', in Wright, C., Smith, B., and Macdonald, C. (eds.) *Knowing Our Own Minds*, Oxford: Oxford University Press.

Garfinkel, S. N. and Critchley, H. D. (2013) 'Interoception, Emotion and Brain: New Insights Link Internal Physiology to Social Behaviour', *Social Cognitive and Affective Neuroscience*, 8 (3), 231–4.

Gertler, B. (2015) 'Self-Knowledge', in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition). Available at: <http://plato.stanford.edu/archives/sum2015/entries/self-knowledge/> [28nd July 2016].

————(2011a) 'Self-Knowledge and the Transparency of Belief', in Hatzimoysis, A. (ed.) *Self-Knowledge*, Oxford: Oxford University Press.

————(2011b) *Self-Knowledge*, Oxford: Oxford University Press.

————(2011c) 'Self-Knowledge', The Stanford Encyclopedia of Philosophy (Spring 2011 Edition), Edward N. Zalta (ed.). Available at: <http://plato.stanford.edu/archives/spr2011/entries/self-knowledge/> [2nd October 2013].

Ghiselin, B. (1952) *The Creative Process: A Symposium* (2nd Ed.) Berkeley and Los Angeles: University of California Press.

Greenough, P. (2012) 'Discrimination in Self-Knowledge', in Smithies, D. and Stoljar, D. (eds.), *Introspection and Consciousness*, Oxford: Oxford University Press.

Griswold, C. L. (1996) *Self-Knowledge in Plato's Phaedrus*, Pennsylvania: Pennsylvania State Press.

Gutchess, A. H., Kensinger, E. A., Yoon, C. Schacter, D. L. (2007) 'Ageing and the Self-Referencing Effect in Memory', *Memory*, 15 (8), 822–37.

Hall, L., Johansson, p., Tärning, B. Sikström, S. and Deutgen, T. (2010) 'Magic at the Marketplace: Choice Blindness for the Taste of Jam and the Smell of Tea', *Cognition*, 117, 54–61.

Hall, L., Johansson, P. and Strandberg, T. (2012) 'Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey'. *PLoS one.* 7 (9).

Hall, L., Johansson, P., Sikström, S., Tärning, B. and Lind, A. (2006) 'How something can be said about Telling More Than We Can Know: Reply to Moore and Haggard', *Consciousness and Cognition*, 15, 697–99.

Harman, G. (1986), *Change in View: Principles of Reasoning*, Cambridge, MA: MIT.

Hasher, L., Goldstein, D. and Toppino, T. (1977) 'Frequency and the Conference of Referential Validity', *Journal of Verbal Learning and Verbal Behaviour*, 16, 107–22.

Hitz, Z. (2011) 'Aristotle on Self-Knowledge and Friendship', *Philosophers' Imprint*, 11 (12), 1–28.

Jacoby, L. L. (1978) 'On Interpreting the Effects of Repetition: Solving a Problem Versus Remembering a Solution', *Journal of Verbal Learning and Verbal Behaviour*, 17, 649–67.

Johansson, P., Hall, L. and Sikström, S. (2008) 'From Change Blindness to Choice Blindness', *Psychologia*, 51, 142–55.

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, 116–19.

Johansson, P., Hall, L., Sikström, S., Tärning, B. and Lind, A. (2006). 'How Something can be Said about Telling More Than We Can Know', *Consciousness and Cognition*, 15, 673–92.

Kahneman, D. (2011) *Thinking, Fast and Slow*, London: Penguin Books.

Kant, I. (1929) *Critique of Pure Reason*, Kemp-Smith, N. (trans.), Basingstoke: Palgrave Macmillan.

Kant, I. (1998) *Groundwork for the Metaphysics of Morals*, Gregor, M. (trans.), Timmermann, J. (eds.), Cambridge: Cambridge University Press.

Kind, A. (2003) 'Shoemaker, Self-Blindness, and Moore's Paradox', *Philosophical Quarterly*, 53, 39–48.

Klein, S. B. (2012) 'Self, Memory, and the Self-Referencing Effect: An Examination of Conceptual and Methodological Issues', *Personality and Social Psychology Review*, 16 (3), 283–300.

Koriat, A. (1995) 'Our Knowledge of Our Own Knowledge: Monitoring and Control Processes in Memory', in Pawlik, K. (ed.) *Bericht uber den Kongress der Deutschen Gesselschaft fur Psychologie in Hamburg 1994*, Gottingen: Hogrefe.

Kornblith, H. (2012) *On Reflection*. Oxford: Oxford University Press.

Lackey, J. (2007) 'Why Memory Really is an Epistemically Generative Source: A Reply to Senor', *Philosophy and Phenomenological Research*, 74 (1), 209–219.

Levinson, S. C. and Majid, A. (2014) 'Differential Ineffability and the Senses', *Mind and Language*, 29 (4), 407–27.

Lewin, K. (1951) 'Intention, Will and Need', in Rapaport, D. (trans. and comm.) *Organization and Pathology of Thought*, New York: Columbia University Press.

Lind, A., Hall, L., Breidegard, B., Balkenius, C., and Johansson, P. (2014). 'Speakers' Acceptance of Real-Time Speech Exchange Indicates that we use Auditory Feedback to Specify the Meaning of What we Say'.

Psychological Science. 28th April 2014, 1–8.

Loftus, E. F. (1975) 'Leading Questions and the Eye-Witness Report', *Cognitive Psychology*, 7 (4), 560–72.

Lopes, D. M. (2014) 'Feckless Reason', *Aesthetics and the Sciences of Mind*, Oxford: Oxford University Press.

Mace, J. H. (2006) 'Episodic Remembering Creates Access to Involuntary Conscious Memory', *Memory*, 14 (8), 917–24.

Mather, M. and Johnson, M. K. (2000) 'Choice-Supportive Source Monitoring: Do Our Decisions Seem Better to US as We Age?', *Psychology and Aging*, 15 (4), 596–606.

McGrath, M. (2007) 'Memory and Epistemic Conservatism', *Synthese*, 2007 (157), 1–24.

McNeill, W. E. S. (2014) 'Embodiment and the Perceptual Hypothesis', *The Philosophical Quarterly*, 62 (284), 569–91.

————(2012) 'On Seeing That Someone is Angry'. *European Journal of Philosophy*, 20 (4), 575–97. Available at: <http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0378.2010.00421.x/full> [29th March 2017].

McVittie, C., McKinlay, A., Della Sala, S. and MacPherson, S. E. (2014) 'The Dog that Didn't Growl: The Interactional Negotiations of Momentary Confabulations', *Memory*, 22 (7), 824–38.

Mill, J. S. (1882/2009) *A System of Logic, Ratiocinative and Inductive*, New York: Harper and Brothers, Gutenberg edition [Online]. Available from: <http://www.gutenberg.org/files/27942/27942-h/27942-h.html#toc5> [22nd February 2015].

Mitchell, T. R. and Thompson, L. (1994) 'A Theory of Temporal Adjustments of the Evaluation of Events: Rosy Prospection and Rosy Retrospection', *Advances in Managerial Cognition and Organizational Information Processing*, 5, 85–114.

Moore, G. E. (1903/1922) 'The Refutation of Idealism', *Philosophical Studies*, London: Kegan Paul (1922), 1–30 [in *Mind* 1903].

Moran, R. (2011) 'Self-Knowledge, "Transparency", and the Forms of Activity', in Smithies, D. and Stoljar, D. (ed.) *Introspection and Consciousness*, Oxford: Oxford University Press.

————(2003) 'Responses to Shoemaker and O'Brien', *European Journal of Philosophy*, 11 (3), 402–19.

————(2001) *Authority and Estrangement: An Essay on Self-Knowledge,* Princeton and Woodstock: Princeton University Press.

Moscovitch, M. and Melo, B. (1997) 'Strategic Retrieval and the Frontal Lobes: Evidence from Confabulation and Amnesia', *Neuropsychologia*, 35 (7), 1017–34.

Nichols and Stich (2003) *Mindreading*, Oxford: Oxford University Press.

Nisbett, R. E. and Wilson, T. D. (1977) 'Telling More Than We Can Know: Verbal Reports on Mental Processes'. *Psychological Review*. 84 (3), 231–59.

Owens, D. (2011) 'Deliberation and the First Person', in Anthony Hatzimoysis (ed.) *Self Knowledge*, Oxford: Oxford University Press, pp. 261–77.

————(1999) 'The Authority of Memory', *European Journal of Philosophy*, 7 (3), 312–29.

Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., J. Spivey, M. J. and C. Richardson, D. C. (2015) 'Biasing Moral Decisions by Exploiting the Dynamics of Eye Gaze', *Proceedings of the National Academy of Sciences of the United States of America* (early edition), 1–6. Available from: <www.pnas.org/lookup/suppl/doi:10.1073/pnas.1415250112/-/DCSupplemental> [20th March 2015]

Reid, T. (1785/1983) *Inquiry and Essays*, Beanblossom, R. E. and Keith Lehrer (eds.), Indianapolis: Hackett.

Roese, N. J. and Vohs K. D. (2012) 'Hindsight Bias', *Perspectives on Psychological Science*, 7 (5), 411–26.

Ross, M. and Sicoly, F. (1979) 'Egocentric Biases in Availability and Attribution', Journal of Personality and Social Psychology, 37 (3), 322–36.

Ryle, G. (1949) *The Concept of Mind* (2009 Ed.), Abingdon: Routledge.

Schwitzgebel, E. (2009) 'Knowing Your Own Beliefs', in Hunter, D. (ed.), *Belief and Agency*: *Supplementary Volume of The Canadian Journal of Philosophy*, 35 (2009), 41–62.

Schwitzgebel, E. (2008a) 'The Unreliability of Naive Introspection', *Philosophical Review*, 117, 245–73.

————(2008b) 'Self-Blindness?', *The Splintered Mind* (Online), 4th June 2008. Available at: <http://schwitzsplinters.blogspot.co.uk/2008/06/self-blindness.html> [24th September 2016].

————(2006) 'The Nisbett–Wilson Myth', *The Splintered Mind*, 11th October 2006. Available at: <http://schwitzsplinters.blogspot.co.uk/2006/10/nisbett-wilson-myth.html> [29th March 2015].

Senor, T. D. (2013) 'Epistemological Problems of Memory', Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition) [Online]. Available at: <http://plato.stanford.edu/archives/fall2013/entries/memory-episprob/> [30th December 2013].

Seth, A. K. (2013) 'Interoceptive Inference, Emotion, and the Embodied Self', *Trends in Cognitive Sciences*, 17 (11), 565–73.

Shah, N. (2003) 'How Truth Governs Belief', *The Philosophical Review*, 112 (4), 447–82.

Shah, N. and Velleman, D. (2005), 'Doxastic Deliberation', *The Philosophical Review,* 114 (4), 497–534.

Shoemaker, S. (1994) 'Self-Knowledge and "Inner-Sense"', *Philosophy and Phenomenological Research,* 54, 249–314.

————(1990) 'First Person Access', *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, in Tomerlin, J. (ed.), Ridgeview Publishing Company.

————(1988) 'On Knowing One's Own Mind' [Online] in *Philosophical Perspectives, Vol. 2, Epistemology*, pp. 183–209 [12th November 2011].

Shultz, S. (2016) '*Our Annual Year*', *The Onion* [online]. Available at: <http://www.theonion.com/article/study-majority-new-marine-life-species-now-discove-52214?utm_content=Main&utm_campaign=SF&utm_source=Twitter&utm_medium=SocialMarketing> [30th December 2016].

Slamecka, N. J. and Graf, P. (1978) 'The Generation Effect: Delineation of a Phenomenon', *Journal of Experimental Psychology: Human Learning and Memory*, 4 (6), 592–604.

Smithies, D. (Forthcoming) 'Belief and Self-Knowledge: Lessons from Moore's Paradox', *Philosophical Issues*, 26.

Stoneham, T. (2004) 'Self-Knowledge', in Niiniluoto, I., Sintonen, M. and Wolenski, J. (eds.) *Handbook of Epistemology*, Dordrecht: Kluwer Academic Publishers, pp. 647–72.

————(MS) *Memory in Inference* (2006).

————(1998) 'On Believing That I Am Thinking', *Proceedings of the Aristotelian Society*, 98 (1998), 125–144.

Strick, M., Dijksterhuis, A., Bos, M. W., Sjoerdsma, A., van Baaren, R. B. and Nordgren, L. F. (2011) 'A Meta-Analysis of Unconscious Thought

Effects', *Social Cognition*, 29 (6), 738–62.

Sullivan-Bissett, E. (2015) 'Implicit Bias, Confabulation, and Epistemic Innocence', *Consciousness and Cognition* (*Special Issue on the Costs and Benefits of Imperfect Cognitions*), 33, 548–60.

Sutton, J. 'Memory' (2012), in Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy* (Winter 2012 Ed.). Available at: <http://plato.stanford.edu/archives/win2012/entries/memory/> [2nd October 2015].

Teroni, F. (2015) 'On Seeming to Remember', in *Colloque 'Memoire et Connaissance'*, Grenoble UPMF: Grenoble, France, 5th June 2015.

————(2005) *An Analysis of Memory*, PhD Thesis, University of Geneva.

Thomas, N. J. T. (2013) 'Mental Imagery', in Zalta E. N. (ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2013 Ed.) [Online], Available from: <http://plato.stanford.edu/archives/fall2013/entries/mental-imagery/> [1st September 2013].

Tsakiris, M. Tajadura-Jimanez, A. and Constantini, M. (2011) 'Just a Heartbeat Away from One's Body: Interoceptive Sensitivity Predicts Malleabilility of Body-Representations', *Proceedings of the Royal Society B*, 278 (1717), 2470–6.

Valaris, R. (2011) 'Transparency as Inference: Reply to Alex Byrne', *Proceedings of the Aristotelian Society*, 111 (2), 319–24.

White, P. A. (1987) 'Causal Report Accuracy: Retrospect and Prospect', *Journal of Experimental Social Psychology*, 23 (4), 311–15.

Wikforss, A. (2004) 'Direct Knowledge of Other Minds', *Theoria*, LXX (2–3), 271–93.

Wilson, J. (2014) 'The Regress Argument Against Cartesian Skepticism', *Analysis*, 72 (4), 668–73.

Wright, C. (2000) 'Self Knowledge: The Wittgensteinian Legacy' in Wright, C., Smith, B., Macdonald, C. (eds.) *Knowing Our Own Minds,* Oxford and New York: Oxford University Press.

Zamuner, E. (2013) 'The Role of the Visual System in Emotion Perception', *Acta Analytica*, 28, 179–87.