UNIVERSITY OF SOUTHAMPTON

FACULTY OF HUMANITIES

Philosophy

Knowing Our Reasons

Distinctive Self-Knowledge of Why We Hold Attitudes and Perform Actions

by

Sophie Alexandria Keeling

Thesis for the degree of Doctor of Philosophy

July 2018

To Liz and Les Keeling

With love and gratitude

ABSTRACT

This thesis argues that we have a distinctive way of knowing why we have our attitudes and perform actions. It is often thought that we have no special access of this sort, even if we have privileged access to other facts about ourselves, like what attitudes we hold. I argue that this orthodoxy fails. Rather, we do have a privileged first-person access to a key explanation of our attitudes and actions – the reasons for which we hold/perform them (i.e. our motivating reasons). In providing an account of this, I draw on insights from cognitive science and appeal to both the personal level, where we can talk of the subject herself, and that of low-level processing. I argue that at the low level, self-knowledge indeed resembles other-knowledge. But regarding how the subject *herself* learns of her motivating reasons, I argue that we use a 'transparency method'. We learn of what our reason is for believing *p*, say, by considering what justifies believing *p*. We look out into the world and consider the good reasons in favour of having that belief. This then allows us to self-ascribe our motivating reason. The final substantive chapter builds on the foregoing to argue that self-knowledge of motivating reasons is distinctive in a further, perhaps surprising, way. Our motivating reasons self-intimate – if we have a motivating reason, then necessarily we will be in a position to know that we have it. Indeed, this is the case even though we can hold attitudes that we are not in a position to know that we hold. Therefore, self-knowledge of motivating reasons not only differs significantly from other-knowledge, but from self-knowledge of attitudes as well.

Table of Contents

Academic Thesis: Declaration of Authorship

I, Sophie Alexandria Keeling

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Knowing Our Reasons

Distinctive Self-Knowledge of Why We Hold Attitudes and Perform Actions

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

   Keeling, Sophie. 2018. Confabulation and Rational Requirements for Self-Knowledge, *Philosophical Psychology*

Signed: Sophie Keeling

Date: 24th of July 2018

Acknowledgements

Parts of chapters four, five, and six constitute the paper 'Confabulation and Rational Obligations for Self-Knowledge'. As such, I have special acknowledgements regarding that material. I would like to thank the two anonymous reviewers from *Philosophical Psychology*, Alfonso Anaya, Ryan Cox, Richard Gray, Conor McHugh, Ema Sullivan-Bissett, Lizzy Ventham, and Jonathan Way. I am also grateful to audiences in Cardiff, Southampton, the *Self-deception: what it is and what it is worth* conference in Basel 2017, the *Philosophy of Language and Mind* conference in Bochum 2017, the *Transparency in Belief and Self-Knowledge* workshop in Oviedo 2015, and the *Annual Meeting of the European Society for Philosophy and Psychology* in Tartu 2015.

Regarding the rest of the thesis, I am indebted to Teresa Baron, Suki Finn, Felix Hagenström, Christopher Little, James McGuiggan, Genia Schönbaumsfeld, Lizzy Ventham, Lee Walters, Jonathan Way, and the research community at the University of Southampton generally. I am also grateful to the many people I've spoken to about my research along the way and who are too numerous to count. Special thanks go to Eleanor Gwynne, for incredibly helpful conversations and feedback on written work, and above all, my supervisors Conor McHugh and Richard Gray. Immense gratitude of a different sort also goes to the AHRC for funding this doctoral research.

Chapter 1 Introduction

> Lizzy and I are in a supermarket discussing weekend plans. Lizzy asks me what I believe the weather will be like and, without missing a beat, I tell her that I believe it will be sunny. 'Huh,' Lizzy replies, having listened to my answer. She's surprised I think this – she knows that the forecast is for rain. (On my part, I had forgotten to check the forecast that morning.) 'Why do you believe that it will be sunny?' 'The clouds look fine at the moment,' I reply. Lizzy then tells me about the forecast. 'Ah,' I reply gloomily. I had been reaching for sunblock, but now withdraw my hand having changed my mind about the weekend weather. Lizzy sees this and apologises for being the bearer of bad news.

The above illustrates an important phenomenon: what appears to be a significant difference between at least some instances of self-knowledge and our knowledge of other people, i.e. other-knowledge.[1] 'Self-knowledge' for the purposes of this thesis concerns my knowledge of certain properties and states of mine, as opposed to my knowledge of a self (whatever a 'self' may be). In the example, I know that I initially believed that *it will be sunny*, and that I now believe that *it will rain*.[2] Lizzy also knows that I held one belief and then the other. But our knowledge seems to differ in various ways. It looks like I 'just know' my beliefs, while Lizzy must rely on my testimony and behaviour. Indeed, it is noteworthy that my testimony sufficed for Lizzy to ascribe the belief that *it will be sunny* to me. She trusted my self-ascriptions automatically: she didn't doubt that I initially believed that *it will be sunny* even though she doubted the belief's truth. In the opening example, then, my self-knowledge seems distinctive relative to other-knowledge. And it's not just my knowledge of belief that seems importantly special. We might observe something similar concerning my knowledge that I have a headache, that I am looking at a computer screen, that I want a cup of tea, that I intend to make a Bakewell tart ...

The extent of the difference between self- and other-knowledge, as well as what it amounts to, bears philosophical significance in numerous ways. Some fundamental consequences include the following. i) It looks like self-knowledge need not be based on evidence. As a result, the thought goes, our self-ascriptions serve as basic beliefs – beliefs which are not themselves inferential but

[1] I take it that we at least sometimes possess self-knowledge. I.e., it is not the case that we only make self-ascriptions that fall short of knowledge in some way, or only express mental states without self-ascribing them (see e.g. Wittgenstein (1953)). I shall not engage with Wittgensteinian challenges here, but the thesis as a whole forms an argument against them.

[2] I probably, however, will not report my belief in these terms – 'I believe that *it will be sunny*.' This sounds as if I have doubts about whether it will be sunny. But telling Lizzy that 'it will be sunny' in this context reports that I believe that *it will be sunny*.

from which we can infer further beliefs. If self-ascriptions constitute basic beliefs, they would be epistemically fundamental.[3] ii) We might think that our self-knowledge of certain mental features, such as our beliefs, is importantly bound up with our rationality and capacity for agency. iii) We might also think that our capacity for self-knowledge helps constitute what it is to have certain mental features in the first place – maybe part of what it is to experience redness just is to be in a position to know that I am experiencing redness. We may not accept all these three contentions, but they at least illustrate the broader significance of discussions of self-knowledge.

Most discussions of self-knowledge centre on our self-knowledge of belief, e.g. how I know that I believe that *it will be sunny*. But this thesis concerns another sort of self-knowledge: our knowledge of why we have our attitudes and perform actions. (At times I will just refer to why we have our attitudes, but only for brevity's sake. The following also applies to our self-knowledge of why we perform actions.) For instance, in the above example, I don't just know that I believe that *it will be sunny*. I also know *why* I believe that *it will be sunny* – for the reason that the sky is cloudless. *Prima facie*, this self-knowledge also seems importantly different from other-knowledge – I don't need to listen to my speech or observe my behaviour to learn of the relevant explanation. And yet philosophers and psychologists often deny that we have distinctive access to why we have our attitudes. This thesis argues to the contrary. Just as my knowledge of my beliefs importantly differs from Lizzy's knowledge of them, so does my knowledge of why I have those beliefs. In arguing for this claim, this thesis will also touch on a range of issues including: self-knowledge of belief; self-ignorance; confabulation; rational requirements; what it is to hold attitudes and perform actions on the basis of reasons.

Let me outline this introductory chapter before continuing. The aim of this chapter is to set up the thesis questions:

> ***Thesis questions:*** *Is self-knowledge of why we have our attitudes and actions a distinctive species of knowledge? In what ways is it/is it not?*

§1.1 further sets up questions concerning the scope of distinctive self-knowledge and what this 'distinctiveness' amounts to. §1.2 distinguishes several meanings of the phrase 'why we have our attitudes' which will prove crucial in what follows. §1.3 outlines the thesis as a whole and the route I will take in arguing that, contra orthodoxy, we have distinctive self-knowledge of why we hold our attitudes and perform actions.

[3] E.g., BonJour (2003) makes this move.

1.1 The difference between self- and other-knowledge

This section outlines the general claim that our knowledge of ourselves differs from others' knowledge of us. §1.1 considers the scope of distinctive self-knowledge.[4] What facts about ourselves can we learn about in a distinctive way? §1.2 considers what this distinctiveness consists in. Supposing that some instances of self-knowledge differ from other-knowledge, in what ways do they do so?

1.1.1 Which instances of self-knowledge are distinctive?

So far, I have introduced the thought that self-knowledge sometimes differs importantly from other-knowledge. (Or, at least, self-knowledge *seems to* sometimes differ importantly from other-knowledge; more on this later.)

This raises the question: which types of self-knowledge differ in this way? We can discount some sorts of self-knowledge immediately. Everyone agrees that nothing special distinguishes our knowledge of properties such as our particular height, eye colour, character traits, and star sign. For instance, I cannot 'just know' whether I bear one of these properties. I would have to get the tape measure out like anyone else to learn how tall I am. And I could only learn that I am shy by inferring it from various pieces of evidence such as the fact that I am afraid of social situations, that I don't linger at parties, and that I speak quietly. Regarding the features of ourselves that we *can* know about in a distinctive way, there are some popular contenders. These include the following:

- Sensations.
- Perceptual experiences.
- Mental images.
- Inner speech.
- Occurrent mental attitudes – e.g. our judgements and occurrent desires. These are events that occur 'in the moment', such as a subject's act of judging that *it will rain* and her sudden fancy for a Bakewell tart.
- Standing attitudes, such as standing beliefs and desires. These are dispositional states that subjects possess even when asleep, like a subject's belief that *Cardiff is the capital of Wales* and her desire to become a philosopher. It is worth emphasising that I take

[4] I talk in terms of 'distinctive' self-knowledge, as Gertler (2015) does. Often the notion is referred to as 'privileged access', but the former term captures the generality I want here.

> believing and judging to come apart, and that one can believe that *p* without being prepared to judge that *p*.[5]

The opening example featured instances of self-knowledge of standing belief; there it very much appeared that I had a distinctive way of learning that I first believe that *it will be sunny* and that I then believe that *it will rain*. This thesis concerns whether we can add another feature of ourselves to the list – why we hold our attitudes and perform actions.

I should note, though, that while we can learn of some features of ourselves in a special way, this is not to say that we always do so. For example, someone might infer that they are in pain from their medical records, just like an observer would have to. Or, when I was in the supermarket with Lizzy, I might have instead inferred that I believe that *it will be sunny* on the basis of my behaviour. I might have reasoned thus: 'I am picking up sunblock and I invited people to a barbeque. Only subjects who believe that it will be sunny engage in these behaviours. Therefore, I believe it will be sunny.'

1.1.2 In what way are some instances of self-knowledge distinctive?

It looks like we have a capacity for distinctive self-knowledge of some facts about ourselves, although not all. But what exactly distinguishes this special self-knowledge from other self-knowledge and other-knowledge (i.e., knowledge of other people)?

Returning to the opening example, some key differences between Lizzy's and my own knowledge of my beliefs should stand out. Recall that I didn't need to observe my behaviour or listen to my speech to learn of my beliefs, unlike Lizzy; rather, I seem to use a different, first-personal, method. Further, when I told Lizzy that I believe that *it will be sunny*, she assumed my self-ascription was right. Even though Lizzy doubted that it would actually be sunny, she didn't doubt I had that belief. Rather she trusted my word, and indeed seemed to take my ascription to be an especially good guide to what I believe. And, when I learnt that I believed that *it will be sunny* and that it actually would rain, I automatically revised my belief. But, when Lizzy learnt that I believe that *it will be sunny*, she had to inform me of relevant evidence for me to alter the belief.

So there appear to be a range of ways in which self-knowledge of some properties might be said to differ from other-knowledge. Indeed, this distinctive self-knowledge will also differ from non-distinctive self-knowledge in these ways, although I do not focus on this distinction here. The following lists the main suggestions regarding the ways in which self-knowledge is distinctive:

---

[5] In taking belief to come apart from judgement, I follow, e.g., Moran (2001) over Boyle (2011a).

1. Self-knowledge is epistemically special in one or both of the following ways:

    1.a. Self-knowledge is especially reliable.

    1.b. Certain features self-intimate.[6]

2. We are in a position to use a distinctive method and warrant to learn of certain features of ourselves.

3. We have first-person authority regarding certain features of ourselves.

4. Some self-knowledge is grounded in our position of agency concerning our attitudes.

Let me elaborate on each of these. Both the list and the following elaboration is heavily indebted to Gertler (2015, 2011).[7]

1. *Self-knowledge is epistemically special in one or both of the following ways*:

1.a. Self-knowledge is especially reliable. Most philosophers accept that, as a minimum, our self-ascriptions are more reliably true than other-ascriptions. At the claim's strongest, though, one might say that certain self-ascriptions will always be correct if formed using the distinctive method, i.e., they will be infallible.[8] This is most plausible in the context of experiential states. So, for example, we might think that when I believe that *I am experiencing redness*, I will always in fact be experiencing redness. Infallibility is more debatable than the weaker claim.

1.b Certain features self-intimate. To say that a feature 'self-intimates' is to say that some sort of entailment relation holds between possessing the feature and knowing that one possesses the feature.[9] Self-intimation claims come in various strengths. At its extreme, the thought is that I will always know that I bear a particular mental feature when I in fact bear that feature. Again, this is most plausible for features other than belief. So one might say that, e.g., I cannot feel pain without knowing that I feel pain (maybe there would be a relevant bodily state, but it would not be pain). There will never be cases in which I do not know about the feature in the distinctive

[6] Philosophers sometimes refer to feature 1.b by saying that certain features are *luminous* (Williamson 2002), or that they are *self-presenting* e.g., Chisholm (1982).
[7] See also Alston (1971) and Jongepier and Strijbos (2015).
[8] E.g., Descartes in the *Meditations* advocates something like this concerning our thoughts, provided one employs the distinctive method with 'great care' Descartes (1644: I.66). And Chalmers (2003: 241-246) and Horgan (2012) argue that a certain sort of belief about our experiential states is infallible.
[9] A related notion is *incorrigibility*. Roughly, an incorrigible self-ascription is one we cannot doubt. See Reed (2011) and Aydede (2013). for varying uses. For the purposes of simplicity, I won't discuss this further. My discussions of reliability and first-personal authority should capture what I would say about incorrigibility.

way.[10] More plausibly we might say that I will always *be in a position* to know that I bear a certain feature (even if I do not currently possess this knowledge). E.g., even if I happen to infer that I believe that *it will be sunny* from my behaviour as Lizzy must, I am still in a position to gain knowledge in the distinctive way. In the following, I will term this 'strong self-intimation' – it is still a weighty claim even if it isn't the strongest possible self-intimation thesis.[11]

We can contrast strong self-intimation with a much weaker version. The 'weak self-intimation' thesis states that we will be in a position to know that we bear a certain feature *if we are rational/if the relevant attitude is rational*.[12] One way of putting this is to claim that (as Moran 2001 does): if my belief that *p* is rational, then necessarily I will be in a position to learn that I believe that *p* in a distinctive way. According to this position, if I have to rely on detective work, then something is rationally wrong with that belief.

We can think that self-knowledge of a given feature is especially reliable without thinking that the feature self-intimates, and vice versa. It is often thought that self-knowledge is especially reliable, but self-intimation tends to be more controversial. I go the other way: I will argue that self-knowledge of why we have our attitudes is strongly self-intimating, but am happy to accept that it is not especially reliable.

2. We are in a position to use a distinctive method and warrant to learn of certain features of ourselves. The vast majority of philosophers take there to be a distinctive method and warrant underpinning at least some instances of self-knowledge. This special method and warrant fundamentally differs from the sort you must use in order to learn of the same features of mine. In the opening example, I don't seem to learn of my beliefs by, say, observing my behaviour, listening to my speech, and/or engaging in inference like Lizzy must. And accordingly, it doesn't seem that my self-ascriptions are perceptually or inferentially warranted. It's controversial, though, as to what the special method and warrant actually amounts to, other than being non-inferential. Chapter two discusses the options.

3. We have first-person authority regarding certain features of ourselves. We are authoritative regarding certain properties of ours, e.g., our beliefs. There are two ways of cashing out first-person authority, which relate to two senses of 'authority' in English. First, one might say that a scholar is an authority in their field, meaning that they know a lot about it and are generally

[10] E.g., Locke claims that '[w]hen we see, hear, smell, taste, feel, meditate, or will any thing, we know that we do so' (1689: II.27.ix).
[11] E.g., Chisholm claims this about thoughts, feelings, and beliefs (1982: 9-11, 25), Siewert about conscious experience (1998: 197-8), and Smithies about mental states including belief and pain (2012).
[12] E.g., Bilgrami (2006), Boyle (2011a), Burge (1999, 1996), Moran (2001), Shoemaker (2012, 1994).

reliable. Second, we might say that a parent is an authority figure, meaning that the parent has a degree of control and responsibility regarding their child. These two uses *roughly* map onto two ways of characterising first-person authority.[13] Perhaps I am more authoritative than Lizzy about my belief because I am especially reliable about my mental states. Alternatively, perhaps I am authoritative concerning my belief because I am in a unique position to change my mind in light of new evidence, and indeed ought to do so.[14]

4. Some self-knowledge is grounded in our position of agency concerning our attitudes. The thought goes that we bear agency over our attitudes; that is, believing, desiring, and so on are all things we *do*. In this way, attitudes differ from other mental states such as pains and perceptual experiences which merely happen to us. Further, our rational agency grounds our self-knowledge of our attitudes.[15]

Let me say a bit more about this so-called rational agency.[16] Some philosophers argue that subjects have a distinctive agential relation to their own attitudes that others do not. Subjects make it the case they have that attitude, and they occupy a position of responsibility concerning the attitude whereby they must ensure it matches with the evidence. Importantly, subjects can form and alter their attitudes directly – it is sufficient that they grasp the relevant evidence. Recall the opening example. As soon as I realised that the weather forecast said it would rain, I revised my belief, and indeed, I would have been irrational not to. In contrast, others can only alter our minds indirectly. Lizzy caused me to revise my belief by informing me of relevant evidence. And indeed, I can also change my mind in this indirect way, such as if I go to therapy. But this isn't the sort of thing we see in the initial example when I revised my belief in light of Lizzy's testimony.

According to some philosophers, this agential relation grounds self-knowledge of our attitudes. The thought is that self-knowledge is distinctive in the other ways we have just discussed in virtue of our agential relation to the target mental state. Recall the opening example again. One might

---

[13] Understanding 'authority' as 'control' is limited, and I will be more precise as the thesis progresses. Moran in particular contrasts two notions of 'first-person authority', e.g., (2001: 113).

[14] E.g., An understanding along these rough lines is either endorsed, or would be endorsed, by Bilgrami (2006), Boyle (2011a, 2011b, 2009a, n.d.), Burge (1999, 1996), Moran (2012, 2004, 2003, 2001), O'Brien (2007, 2005) and Parrott (2015). A related account argues that we have first-person authority in virtue of the fact that our self-ascriptions *express* the mental state they concern, e.g., Bar-On (2004) and Finklestein (2003).

[15] E.g., Bilgrami (2006) who notably grounds self-knowledge of belief in practical as well as rational agency, Boyle (2011a, 2011b, 2009a, n.d.), Burge (1999, 1996), Moran (2012, 2004, 2003, 2001), O'Brien (2007, 2005) and Parrott (2017, 2015). This approach can be distinguished from others that ground self-knowledge in rationality where this rationality is not construed in agentive terms, e.g. Gallois (1996) and Shoemaker (2012, 1994).

[16] Rational agency and control are especially discussed in Boyle (2009b, 2011b). Also, outside of discussions of self-knowledge, see e.g., Hieronymi (2009, 2008, 2006) and McHugh (2017, 2013).

say that I come to know what my new belief is *in virtue of* changing my mind in this way, i.e., exercising rational agency.

Suggestions (1)–(4) are highly debated. As noted, we might dispute what each claim amounts to. Further, we might deny that all the claims capture the difference between self- and other-knowledge, or that all bear equal importance. For example, until Moran (2001), (4) was traditionally ignored, and many still deny it. At any rate, these issues should become clearer when we explore the debates more fully in chapter two. For now, let us note that at least some instances of self-knowledge seem to importantly differ from other-knowledge in various ways. Further, I can be more precise regarding the term 'distinctive self-knowledge.' Such self-knowledge differs significantly from other-knowledge, and will do so in virtue of possessing at least some of the features from the above list.

Before continuing, though, let me note that one might deny something I have supposed so far – that self-knowledge actually *does* differ from other-knowledge. Cassam (2017, 2014, 2011, 2010a), Carruthers (2013, 2010), and Ryle (2009) take this route. While they do seem to allow that some self-knowledge might be distinctive, such as that of sensations, they curtail the scope of distinctive self-knowledge to an extreme degree. Regarding our opening example, they would go so far as to say that my knowledge of my beliefs only *looks* special: actually, it resembles Lizzy's in all important respects. Cassam, Carruthers, and Ryle think that both I and Lizzy use the same method in learning of my beliefs – some form of inference. (What I mean by 'some form' of inference will become clear later on.) Further, my ascriptions are no more reliable or authoritative than Lizzy's other-ascriptions and are not more distinctive in any other way. Cassam, Carruthers, and Ryle's position is controversial. But, as we will see, one argument of theirs in particular raises real problems for distinctive access, including my main focus here – distinctive access to why we hold our attitudes. As such, chapter two returns to their account in detail.

1.2 How we know our reasons: some preliminaries

So far, I have discussed the distinctiveness of self-knowledge. I'm interested in one type of self-knowledge in particular – our self-knowledge of why we hold our attitudes and perform actions. In addition to knowing *that* we have a given attitude, we can also know *why* we have it. So, in the opening example, I know that *I believe that it will be sunny for the reason that the sky is cloudless*. And a subject might also know, for example that: she prefers apples to bananas because of the taste; that she wants a fancy yogurt because of the clever marketing; that she believes that *God doesn't exist* because of her upbringing; that she is going to the shop for the reason that she's run

out of hummus. This comes down to knowing what explains an attitude or action. I take knowing an explanation to amount to knowing the fact that a given *explanans* explains a given *explanandum*. (An *explanandum* is what is to be explained, and an *explanans* is what explains the explanandum. E.g., when a subject knows why she prefers apples to bananas, she knows a fact: her preference for apples to bananas (*explanandum*) is explained by the taste (explanans).) Indeed, we should agree that knowledge of an explanation takes this rough form regardless of the finer details concerning the nature of explanation.[17]

The example explanations that I opened this section with are a diverse bunch, and we can say 'why' one has an attitude in different ways. §1.2.1 clarifies the ways in which we can explain an attitude, and §1.2.2, the different senses in which we can talk of the 'reason' for an attitude. It is worth bearing these in mind. The distinctions are crucial: I will only argue that we have distinctive access to one type of explanation. Further, I will use these concepts throughout the thesis.

1.2.1 Types of explanation

We can explain someone's attitude or action in two ways. One group consists of the following sorts of explanation: I believe that *it will be sunny* because there currently aren't any clouds; Suki prefers takeaway pizza to frozen because it tends to be cheesier; Felix goes to see a film because it looks interesting from the trailer; Sally is going to the shop because she's run out of hummus. These explanations cite the subject's reasons. I will take these explanations to be causal (contra Anscombe (2000)). The subject's reason causes them to adopt a given attitude, but the explanation is not *purely* causal.[18] To use Davidson's (1963) term, this is a type of *rationalising explanation* in making it intelligible that the subject would have the attitude or perform the action. I will call explanations citing the reason(s) for which someone holds an attitude or performs an action *reason explanations*. These only apply to attitudes and actions – there is no reason explanation for a sensation of pain, say, since one cannot feel pain *for a reason*.

Reason explanations can be contrasted with another sort of explanation. Examples of this alternative group include: I believe that *it will be sunny* because I didn't read the weather forecast; I believe that *it will be sunny* because my light receptors identified patterns and processed them in a certain way; I believe that *it will rain* because I am rational. These are all *non-reason explanations* in that they don't cite the reason for which the subject holds the belief. Non-

[17] For a good idea of the debates and options concerning the nature of explanation, see Woodward (2017) Achinstein, (1985: ch1), and Reutlinger (2017).
[18] Chapter seven further discuss the nature of motivating reasons in the context of belief.

reason explanations are diverse, but at the moment I can just say that they are generally causal ones.[19] Indeed, these explanations are *purely causal*, unlike reason explanations.[20]

1.2.2 Types of reason

I have talked of *reason* and *non-reason explanations*, but we must clarify the sort of reason I have in mind; I take there to be (at least) three different types.[21]

First, we can return to the reason explanations above: I believe it that *it will be sunny* for the reason that there are no clouds; Suki prefers takeaway pizza because of the cheese. These cite a *motivating reason*, the reason for which the subject actually holds an attitude.[22] We should not confuse these with the following two other types of reason.

Second, we might say: the reason it is raining is the air pressure; the reason that I believe *it will be sunny* is that I didn't look at the weather forecast; the reason that Carl believes *the table is dirty* is his OCD. I class these as *purely causal explanatory reasons* – these cite one state/event which explains another simply in virtue of having caused it.[23] Purely causal reasons are at work in what I have been calling *non-reason explanations*. Such explanations do, then, feature one sort of reason, but they are not motivating reasons.

Third, we sometimes say things such as: I believed *it will be sunny* because the sky was cloudless at the time, but that's not really a reason – clouds come and go with speed; you should prefer takeaway pizza to frozen – one reason is the greater amount of cheese; a reason for going to the seminar is that you will learn a lot. These cite *normative* reasons, i.e. good reasons for having an attitude or performing an action. Roughly, we can say that a normative reason is a consideration

---

[19] We can also give non-causal explanations. For a good overview of these approaches, see Reutlinger (2017). One important type of non-causal explanation is a grounding explanation, which cite what makes it the case that the explananda obtain. For example, we can explain my belief that *it will rain* by citing the neurophysiological structure underpinning my belief. On grounding explanations see for example deRosset (2013), and this sort of approach in psychological explanation, Bermúdez (2005). I return to some of these issues later in the thesis. But, often reason explanations are contrasted specifically with (purely) causal explanations, e.g., Hornsby (1997: ch. 8).

[20] Regarding this distinction concerning action, see e.g., Alvarez (2017), Malle et al., (2007), Hornsby (1997), and in the context of belief, e.g., Jones (2002). I use the term 'reason explanation' as opposed to 'rationalising explanation', although one could use either.

[21] Alvarez (2010), though, would dispute this, and say that there is only one kind of reason which occupies different roles.

[22] My use of the term 'motivating reason' thus differs from some understandings, see e.g., Hieronymi (2011).

[23] I term these 'purely causal explanatory reasons' as opposed to just 'explanatory reasons', unlike, say, Alvarez (2017). This is because I take motivating reasons to also be explanatory.

that speaks in favour of having an attitude or performing an action.[24] There can be normative reasons for a subject to do something whether or not the subject recognises these, and whether or not they are the subject's motivating reasons.[25]

Unless otherwise specified, when I talk of 'reasons', I am referring to a subject's motivating reasons. So, this thesis' title – Knowing Our Reasons – refers to one sort of reason in particular. To anticipate my conclusion, I argue that we have distinctive access to our motivating reasons for our attitudes and actions.

1.3 Overview

Now that we have covered some preliminaries, the thesis questions should be clearer. To recall:

> ***Thesis questions:*** *Is self-knowledge of why we have our attitudes and actions a distinctive species of knowledge? In what ways is it/is it not?*

That is, is self-knowledge of why we have our attitudes like that of, say, self-knowledge of belief in being importantly different from other-knowledge? And if so, in virtue of which of the possible characteristics (1)–(4) is it distinctive, and how should we cash these characteristics out? E.g., what method do we use when acquiring this self-knowledge? And I can note that if we do have distinctive access to why we have our attitudes, it will be in virtue of a distinctive access to reason and/or non-reason explanations for our attitudes.

I answer the first thesis question in the affirmative by answering the second. I argue that we have distinctive self-knowledge of why we have our attitudes by proposing a specific account of this self-knowledge under which it is distinctive. More specifically, we have distinctive self-knowledge of why we have our attitudes in virtue of distinctive self-knowledge of one type of explanation in particular: reason explanations. I accept that we learn of purely causal factors in the same way as observers but I argue that we have distinctive access to the reasons for which we hold attitudes and perform actions.

---

[24] I happily accept, though, that the precise details will differ between practical and epistemic normative reasons.

[25] Regarding the distinction between three kinds of reasons regarding action, see, e.g.: Alvarez, who discusses *normative, motivating,* and *explanatory* reasons (2017, but see (2010) for an important qualification in Alvarez's own view); Baier who discusses 'reason' in *deliberation, justification,* and *explanation* (1965: ch. 6). Regarding the distinction between reasons for belief, see, e.g., Sylvan, who distinguishes between *normative*, *explanatory,* and *operative* reasons (2016a, 2016b).

In arguing for my answers to the thesis questions, I argue against the orthodox position in the literature: that we lack distinctive access to why we have our attitudes in a blanket sense. But, I argue that while there is reason to accept this orthodox view, the position faces a range of problems. In place of this orthodoxy, I propose a *two explanations* account. This alternative uses insights from cognitive science and appeals to both the personal level where we can talk of the subject herself, and the subpersonal level, where we cannot. I argue that at the subpersonal level, self-knowledge is indeed akin to other-knowledge. Self-knowledge requires a similar sort of quasi-inferential process using representations about the subject. But self-knowledge is nevertheless distinctive in virtue of the personal level account. I argue that self-knowledge of motivating reasons is rooted in our rational agency. As such, we acquire this self-knowledge using a distinctive method and warrant, and have first-person authority regarding our motivating reasons. Further, our motivating reasons strongly self-intimate.

In the thesis, I will focus on our explanations of attitudes, as opposed to explanations of action. And indeed, some discussions centre on certain attitudes over others. For example, chapter seven focuses on the case of belief since it's the most straightforward in the context. But I take what I say to extend to motivating reasons for action as well. I return to motivating reasons for action in the conclusion.

The remainder of the thesis proceeds as follows:

Chapter Two. *From Self-Knowledge of Belief to Self-Knowledge of Why We Have Our Attitudes: The Options*

Here I introduce what seem to be the options on the table for answering the thesis questions. I introduce the questions by first explaining the major positions regarding self-knowledge of belief. This helps me make concrete suggestions for cashing out our self-knowledge of why we hold our attitudes; the literature rarely discusses this specific issue.

I then set out what appear to be the two main possible answers to the questions. The mainstream position, which I call the Orthodoxy, says that we lack distinctive access to why we have our attitudes. Indeed, even those who think that we have distinctive access to many other mental features deny that the same holds for why we have our attitudes. There are two compatible ways of cashing out what this self-knowledge looks like under the Orthodoxy. We might think that such self-knowledge is inferential proper, where the subject herself engages in inference. Let's call this *inferentialism* about self-knowledge of reasons. Or we might think that that the relevant process is (just) comprised by subpersonal mechanisms that transition between various representations to reach the conclusion. This is the sort of processing that even underlies perception and is not

something the subject herself can be said to engage in. I do not take self-knowledge at the subpersonal level to be properly inferential at all, but rather computational. As such, *computationalism* about self-knowledge of reasons argues that self-knowledge of reasons is acquired using computational mechanisms. That said, the Orthodoxy faces resistance. In particular, some philosophers seem to assume that our agential relation to our attitudes grounds a distinctive self-knowledge of reason explanations. But the literature fails to set out what precisely this looks like, so I propose a way of doing so. I introduce what I dub the *reasons transparency method* (RTM): we learn why we hold an attitude by considering the question 'why hold that attitude?'. I should note that while one could develop a quasi-perceptual account of self-knowledge of motivating reasons, I do not do so here; it would face similar problems to the Orthodoxy.

To sum, the chapter will end with the main options set on the table. We might say that self-knowledge of why we have our attitudes isn't distinctive, but rather inferential/computational. Or we might say that it is, and that it is somehow grounded in our rational agency.

Chapter Three. *Self-knowledge of why we have our attitudes – taking the question seriously*

Before proceeding any further, this brief chapter argues that the two options from chapter two are indeed live options. After all, one might have thought that the answers to the thesis questions are obvious: the Orthodoxy is, well, orthodox. As such, one may not see the need to read or write an entire thesis on the topic. Here, I briefly reject three arguments in favour of the Orthodoxy; it is not clearly right.

Chapter Four. *The Orthodoxy's appeal: Inference to the best explanation*

One argument for the Orthodoxy still remains, though, and I take this to be the best. Considering this argument will ultimately strengthen my own case for the distinctiveness of self-knowledge of why we hold our attitudes, and indeed will shape my own positive account of the way in which this self-knowledge is acquired.

The strongest argument is an inference to the best explanation (IBE). Arguments of the IBE form show that *x* is true on the grounds that accepting *x* provides the best explanation of a given phenomenon. In the Orthodoxy's inference, self-ignorance and error about why we have our attitudes form the explananda. We might think that subjects only have a third-personal method for ascribing their reasons because this account of self-ascription provides the best explanation of self-ignorance and error (i.e., confabulation). Indeed, the Orthodoxy contends, alternative accounts of self-knowledge cannot even provide a good explanation at all. So, I end the chapter

having argued for the following claim: there is reason on explanatory grounds to accept either computationalism or inferentialism about self-knowledge of motivating reasons.

Chapter Five. *Problems with the Orthodoxy*

Here I start the case for my answer to the thesis questions. This chapter advances various problems the Orthodoxy faces. I first argue that the Orthodoxy's explanation of self-ignorance and error is weaker than it may seem, even if it is the best on the table so far. I then argue that the Orthodoxy is generally unappealing. I conclude that while we have (*pro tanto*) reason to accept either computationalism or inferentialism, there is also (*pro tanto*) reason not to accept either.

Despite the progress, we now seem to be left in a troubling impasse – how do we reconcile the advantages and disadvantages of the Orthodox account of self-knowledge of reasons?

Chapter Six*. The two* explanations *account of self-knowledge*

This chapter motivates a better account of how we know why we have our attitudes and starts spelling out my answer to the thesis questions. Contra the Orthodoxy, I argue that we have distinctive self-knowledge of why we hold our attitudes. I also propose a positive account of the way in which such self-knowledge is distinctive.

To be precise, I propose a *two explanations* account of self-knowledge of motivating reasons (and indeed other mental features). The resulting account of self-knowledge retains what is appealing about the Orthodoxy but avoids what isn't. We should explain self-knowledge in two ways, like how we can provide two compatible explanations of perception. Regarding self-knowledge I say the following. At one level, it is computational and akin to other-knowledge. At another level, the agentialist picture holds. We use the reasons transparency method, and our resulting self-ascription is warranted as a result of our rational agency. In a way, therefore, I take aspects from both options in chapter two but maintain that self-knowledge is importantly distinctive under this picture.

I then set out various advantages of my view, such as that it provides an even better explanation of confabulation than the Orthodoxy, and I counter some objections one might have at this point.

Chapter Seven*. Motivating reasons as strongly self-intimating*

This chapter develops my picture of the ways in which our self-knowledge of motivating reasons is distinctive. I argue that our motivating reasons strongly self-intimate. That is, necessarily, if a subject has a motivating reason, they will be in a position to learn that they have it. We should think that motivating reasons strongly self-intimate even though attitudes do not. We will have

come a long way, then, from the Orthodox position: that our self-knowledge of why we have our attitudes is not distinctive even if other sorts of self-knowledge are.

Chapter Eight. *Conclusions*

This chapter summarises the thesis as a whole and then does two things.

First, I clarify how the preceding discussions support the thesis' main claim – that we have distinctive self-knowledge of our motivating reasons. This includes spelling out the four ways in which self-knowledge of our reasons is distinctive under my model. (My full account of first-person authority relies on chapter seven.) And, having focussed on reasons for attitudes, I widen my scope and clarify that my arguments also apply to reasons for action.

Second, I emphasise the thesis' contributions aside from this central contention concerning self-knowledge of reasons. For example, we can and should extend the *two explanations* approach to self-knowledge of attitudes. Also, I will have said something important about the nature of motivating reasons and argued for a novel explanation of confabulation.

Chapter 2 From Self-Knowledge of Belief to Self-Knowledge of Why We Have Our Attitudes: The Options

This chapter introduces two options for answering my thesis questions (*is self-knowledge of why we have our attitudes and actions a distinctive species of knowledge? In what ways is it/is it not?*). The following chapter then discusses why we should seriously consider both options; contra popular assumption, it is not that one is clearly better than the other.

I start by introducing accounts of self-knowledge of belief. I do so because little has been said about self-knowledge of why we have our attitudes. Considering another mental feature will therefore allow me to fully introduce the self-knowledge debate and provide details I can use in fleshing out the options regarding why we have our attitudes. I outline three main accounts of how we learn that we have a mental state like belief: quasi-perception (which I discuss in §2.1), agentialism (§2.2), and neo-Ryleanism (§2.3). I then consider self-knowledge of why we have our attitudes. I start with the orthodox position that many, including quasi-perceptual and neo-Rylean theorists, accept – that such self-knowledge fundamentally resembles other-knowledge (§2.4). I then introduce an agentialist alternative and spell out what I take to be the best way of articulating this other option (§2.5). I will end with what seem to be the best two options on the table for answering the thesis questions: either the Orthodoxy or agentialism concerning self-knowledge of why we hold our attitudes.

2.1 Self-knowledge of belief: Quasi-perception

This section first outlines one particular quasi-perceptual account of self-knowledge, that of David Armstrong, before introducing the view more generally.

Here is an illustration of how self-knowledge looks under Armstrong's (2001) view:

> NICARAGUA ONE. I ask you whether you believe that *the capital city of Nicaragua is Managua*. The way you learn of your belief parallels how you learn of the surrounding world using vision. You can come to know that there is an apple by detecting that one is present, that is, by seeing it. Similarly, you can learn that you believe that *Managua is the*

> *capital of Nicaragua* by simply quasi-perceiving that you have the belief. In all this, you need not consider evidence for your having the belief (e.g., how you answered a question about Nicaragua's capital in a quiz last week).

For Armstrong, we learn that we have a belief by 'quasi-perceiving' that we have it. Armstrong argues that self-knowledge resembles perceptual knowledge and that we acquire both using a detection mechanism. In the case of self-knowledge, we use a 'self-scanning process' (2001: 324). The lower-order state, such as your belief that *Managua is the capital of Nicaragua*, is metaphysically distinct from one's awareness of it. The relation between the two is purely causal and the lower-order state causes its self-ascription by way of relevant processes. This causal link could conceivably be replicated such that we would be able to detect others' mental states in a similar way, although this isn't the case as things stand. The causal link also means that self-knowledge resembles visual perception in the potential for mistakes since the causal mechanism might fail. Unlike perception, though, the processes underpinning self-knowledge do not operate via a sense organ. As such, we cannot direct the mechanism as we wish, unlike how we can move our eyes to better see certain things (*Ibid*. 325-6).

Armstrong also thinks that both self- and perceptual-knowledge are warranted non-inferentially, namely, along reliabilist lines (e.g. *Ibid*. 325, 238). Armstrong writes that one's belief that *p* can be non-inferentially warranted in virtue of its being sensitive to *p*. This sensitivity obtains when as a matter of 'empirical necess[ity]' (*Ibid* 189) one would not have the belief that *p* had *p* not been the case. If the self-ascriptions formed by the self-scanning process are reliably accurate, as Armstrong thinks they are, then our self-ascriptions will be thusly warranted.[1]

Armstrong's account represents a broader class of views claiming that self-knowledge significantly (although not completely) resembles perceptual knowledge. We see this in thinkers such as Russell (1917), Gertler (2012, 2001), Chalmers (2003), Pitt (2004), BonJour (2003), Locke (1689), Lycan (1996, 1995) and Macdonald (2014, 1998). Such accounts agree with Armstrong that the target mental state directly causes our self-ascription of it without the subject having to infer. According to the above thinkers, our attention or other such mechanism focusses on the belief that *p* or other mental state. The object of attention is not, say, that *p* (contra Moran, whom I discuss in §2.2.1) or evidence that we believe that *p* (contra Cassam whom I introduce in section in §2.3.1). And, like Armstrong, quasi-perceptual accounts in general think self-knowledge is warranted non-inferentially, as perceptual knowledge is usually thought to be. There are various

[1] Armstrong's theory of self-knowledge falls out of his account of consciousness. Armstrong argues that a mental state is conscious if one is conscious of it, where he construes being conscious of the state quasi-perceptually. See (*Ibid*. Ch. 6 §9).

controversies beyond these, e.g., concerning the nature of both perceptual and introspective warrant, and whether the mental state partly constitutes as well as causes the self-ascription.[2] These debates, however, do not matter for the present project.

***

To summarise, we can return to our list of ways in which self-knowledge might be distinctive and observe what the quasi-perceptual account says concerning them. This section on quasi-perception can be brief for my purposes, so I will simply fill in the gaps here.

	Quasi-perception
Extra-reliable	Yes
Self-intimation	Depends
Distinctive method and warrant	Yes (at least in practice)
First-person authority	Yes – based in our reliability
Role of rational agency	No

Quasi-perceptual views face attack from two directions. Some philosophers argue that quasi-perceptual accounts take self- and other knowledge to be too similar, others, that they take self- and other-knowledge to be too different. The next position (agentialism) I will outline has the first worry.

[2] Descartes, Russell, Gertler, Chalmers, Pitt, and BonJour think that the lower-order state partly constitutes the self-ascriptive belief. Lycan and MacDonald, on the other hand, do not. We can call these two subtypes 'acquaintance' and 'inner sense' views respectively (as e.g. Gertler (2015) does). See Gertler (2015) for overviews of these two different, but related, accounts.

Additionally, Shoemaker's distinction between *object perception* and *broad perceptual* models of quasi-perception captures another set of differences. See Shoemaker (1994 esp. lectures I and II), and on this Cassam (2014).

2.2 Self-knowledge of belief: Agentialism

Agentialists claim that we exercise agency towards our attitudes, and that this agency grounds our self-knowledge of those attitudes. Here in this section, I first introduce a key agentialist account – that of Richard Moran (§2.2.1) – before considering the position more broadly (§2.2.2).

2.2.1 Moran's agentialist account of self-knowledge

The following exemplifies self-knowledge of belief under Moran's (2001) account:

> NICARAGUA TWO. Say that I ask you whether you believe that *the capital of Nicaragua is Managua*. Considering what you *believe* the capital to be consists in thinking about what it actually *is*. This might involve going on the internet, for example, or asking a knowledgeable friend. You will realise that Managua is Nicaragua's capital and can therefore reply that you believe that *Managua is the capital city of Nicaragua*.

Moran starts by observing that we bear agency regarding our attitudes, and that this constrains one's account of self-knowledge. Quasi-perceptual accounts, Moran contends, have overlooked this (§2.2.1.1). He also provides a positive picture, whereby our rational agency grounds various aspects of distinctive self-knowledge, such as the method we use to acquire it (§2.2.1.2).

2.2.1.1 Moran and rational agency

Moran takes it, as do I, that we bear an agential relation to our attitudes (e.g., beliefs, desires, and hopes). Like drinking tea and eating hummus, believing, desiring, etc. are things that we do. Indeed, it is not even that we simply exercise practical agency in gathering evidence, say. Rather, we possess a distinctly rational agency concerning our attitudes and bear epistemic responsibility for them. In this way, attitudes differ from sensations – I do not *do* something in experiencing pain. Moran argues that, as a result, although we may well learn of our sensations quasi-perceptually, this cannot be the case for our attitudes.[3] Moran gives us the following:

> The special features of first-person awareness cannot be understood by thinking of it purely in terms of epistemic access (whether quasi-perceptual or not) to a special realm to which only one person has entry. Rather, we must think of it in terms of the special responsibilities the person has in virtue of the mental life in question being *his own*. In much the same way that his actions cannot be for him just part of the passing show, so his

[3] Moran makes clear he is only concerned with our attitudes on (2001: 9-10).

> beliefs and other attitudes must be seen by him as expressive of his various and evolving relations to his environment, and not as a mere succession of representations (2001: 32).

For Moran, we are the agents of our attitudes, and therefore 'see' our attitudes in particular ways. Moran therefore rejects accounts of self-knowledge like quasi-perception whereby we treat our attitudes as just a 'passing show.'

Moran's thoughts in the passage can be clarified by considering Moore paradoxical (MP) statements.[4] MP statements follow one of two forms: '(1) "P, and I don't believe it," or (2) "I believe that P, but P is not true."' (Moran 2001: 69). *Prima facie*, such statements seem odd while the interpersonal correlates do not. For example, there is something odd about saying that 'the gig is sold out, but I don't believe it' or 'I believe that *the gig is sold out*, but it isn't.' And yet it is fine to say that 'the gig is sold out, but Beth doesn't believe it' or 'Beth believes that *the gig is sold out*, but it isn't.' Subjects who utter or believe MP statements do not see their attitudes as expressing their take on the environment. Rather, these subjects think that their lower-order belief is false. And yet we do not normally have such a relation to our mental life.

For Moran, the irrationality of MP statements precludes quasi-perception as an account of how we learn that we have an attitude. Moran writes that using quasi-perception or inference to learn of an attitude would be to occupy a 'theoretical stance' towards it (*Ibid.* 65). The subject would be simply trying to discover what her attitude happens to be, similar to how she might try to learn whether the gig is sold out, or whether Beth believes that *the gig is sold out*. Yet, not only does saying that we acquire all self-knowledge from this stance fail to explain the irrationality of MP statements; further, it is incompatible with it (*Ibid.* 83-84). From the theoretical stance, a belief's truth and justification bear no relevance when trying to learn of it. When the subject employs the 'theoretical stance' to learn whether she believes that *p*, whether *p* is true and whether she would be justified in believing that *p* bear no significance; the processes she employs do not take into account these facts at all. It is like how, uncontroversially, the question of whether the gig is sold out does not directly bear on the question of whether Beth believes that it is. It only helps the subject to learn of Beth's belief if she happens to know that Beth regularly checks the availability of tickets. Under the theoretical stance, 'the thought expressed in a Moore-type sentence would describe a perfectly coherent empirical possibility on which one could sensibly report' (*Ibid.* 84). But while it is sensible to say that 'Beth believes that *the gig is sold out*, but it isn't,' it is *never* sensible to say that 'I believe that *the gig is sold out*, but it isn't.' Yet quasi-perceptual accounts allow that MP statements are at least sometimes rational. Unless we sacrifice

[4] So-called after Moore (1942).

our plausible intuitions about MP statements, Moran argues, we should reject quasi-perceptual accounts. This problem also applies to neo-Ryleanism about self-knowledge, which we shall encounter in §3.

2.2.1.2 Moran's account of self-knowledge

For Moran, it is not just that our rational agency and responsibility preclude quasi-perceptual accounts; our rational agency grounds self-knowledge of our attitudes and its distinctive features. I will set out Moran's account of the method, warrant, and first-person authority, and also an extra component of his picture.

Moran argues that we acquire distinctive self-knowledge using the *Transparency Method* (TM). Our self-ascriptions normally treat our mental life as more than simply a 'passing show' in part because we use TM when forming these self-ascriptions. Recall NICARAGUA TWO, in which you learn whether you believe that *Managua is the capital of Nicaragua* by considering whether Managua actually *is* the capital. NICARAGUA TWO illustrates the transparency method, which was introduced by Gareth Evans (1982):

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

When using TM to learn whether we believe that *p*, we do not consider anything concerning ourselves, but rather, whether *p* is true. That is, the question of whether we believe that *p* is *transparent* to the world-directed question of whether *p* is true. In answering the question concerning ourselves, we turn our attention outwards, to the world itself, and reach a conclusion about the world. In NICARAGUA TWO, for instance, you learn whether you believe that *Managua is the capital of Nicaragua* by researching Nicaragua and ascertaining that the capital is indeed Managua. This procedure sharply contrasts with the quasi-perceptual theorist's talk of mechanisms directed at the mental state itself.

Moran cashes out the transparency method in the following way. When learning that we have a belief using TM, we form the lower-order belief using deliberation (*Ibid.* §2.5). We can bring about a belief in ourselves in other ways, e.g., hypnotism, but these do not constitute the norm (*Ibid.* 117)). For example, Moran would say the following about NICARAGUA TWO: you consider the evidence concerning Nicaragua and make up your mind on the basis of that evidence. That is, you weigh the reasons on the matter and reach a conclusion. That conclusion then tells you what your

belief is. This 'deliberative stance' is one of rational agency. You do something in forming a belief on the basis of evidence, and you hold responsibility for it.

For Moran, our rational agency also grounds our warrant for self-ascription. His account makes use of a transcendental entitlement. Transcendental entitlements are entitlements secured by a transcendental argument. Arguments of this form show that *x* must be the case because *x* is a prerequisite for a certain fact to obtain that would be hard to reject.[5] So, to give Moran's account, I read him as follows.[6] Our self-ascriptions are warranted because we possess a transcendental entitlement rooted in the nature of deliberation; our self-ascriptions formed using TM must be warranted because we require this warrant to deliberate. If we are to deliberate at all, we must fulfil various norms. One relevant norm is that we must think that the conclusion we reach will change our beliefs. This is because deliberating is to decide what to believe, and not in the sense of simply forming an intention to believe something dependent on the ability to do so. But we can only *assume* that our conclusion will change our belief. This is because of another norm of deliberation: we cannot form our belief that *p* on the basis of our judgement that *p* and the premise that our judgements regularly lead to attitude change. The fact that we are rational agents is extraneous to deliberation since it does not help us decide what to believe (*Ibid.* 95). Therefore, in order to fulfil the norms of deliberation and actually deliberate at all, we must tacitly assume that when we conclude that *p*, we indeed believe that *p*. This assumption entitles us to self-ascribe the belief that *p*.

Moran also thinks that rational agency grounds our first-person authority regarding our attitudes. Specifically, this is a special sort of first-person authority. Recall from the introduction that 'authority' has several meanings in English. Someone might be an 'authority' on a matter in virtue of knowing a lot about it or be an 'authority' over something in having control over it. Quasi-perceptual accounts think we have first-person authority concerning our attitudes in the first sense – our testimony is very reliable. Indeed, Moran does not question this. But he also thinks that we have first-person authority in a way closer to the second sense as well: we have rational control and responsibility concerning our beliefs. Note, though, that this is specifically a *rational* control. On all this, see e.g. (*Ibid.* 92).

The final feature of self-knowledge under Moran's account is that our rational agency grounds obligations for self-knowledge. As a result of the position of responsibility we bear towards the lower-order belief, we bear the obligation to be in a position to learn of it using TM (2001: e.g.

[5] See Gertler (2011: 183, §6.4.1 and §6.4.2) and Stern (2015).
[6] From Moran (2001) I especially use p. 94-5. Moran (2003) and Gertler (2011) are also helpful in understanding his view.

xxix-xxx, 84, 127). If we cannot, there is 'something wrong with [us]' and we will be in a state of 'alienation' concerning the belief (*Ibid*. 68). After all, we will be in a position to learn of our belief that *p* using TM when we are prepared to judge that *p* is the case. (To clarify, I understand judgement as an occurrent event of taking *p* to be true which can come apart from a standing state belief that *p*.) If we are not prepared to judge that *p*, then our belief that *p* will be irrational by our lights – our beliefs should match up with what we judge to be the case. In cases in which we are not prepared to judge that *p*, we have to rely on an alternative method such as inference to learn of our belief. Moran, then, endorses a weak version of self-intimation: necessarily, if we have a *rational* belief then we will be in a position to know whether we have it using TM.

As hopefully should be clear, Moran's positive account avoids his worry with the quasi-perceptual approach. Moran can accept that Moore paradoxical statements are always irrational and can indeed explain this in the following way (*Ibid.* 83-4). MP statements are irrational because, Moran takes it, we rationally ought to use TM and deliberate when forming self-ascriptions. When using TM, we learn of our belief by forming it in line with our best estimation of the evidence. A subject therefore would not be using TM if they conclude that 'I believe *p*, but *p* is false,' or that '*p* is true, but I don't believe it.'

2.2.2 Agentialism in general

I class Moran's account as a form of 'agentialism' about self-knowledge.[7] Agentialism grounds self-knowledge and its distinctive features in our agential relation to our attitudes. As such, agentialism pertains minimally to our self-knowledge of attitudes (and, I will argue, our motivating reasons), but not our sensations. The view intellectually descends from Kant (1958) and other proponents include Bilgrami (2006), Boyle (Boyle, 2011a, 2009a, n.d.; Burge, 1999, 1996; Moran, 2012, 2004, 2003, 2001; O'Brien, 2007, 2005), and Parrott (2017, 2015). This section formulates a paradigm agentialist account which I take to be the best way of understanding these agentialist ideas; indeed, I endorse the account. I will then use this account as my base when formulating an agentialist account of how we know why we hold our attitudes in §2.5. In this section, I first say something about rational agency and how it precludes quasi-perception (§2.2.2.1) and then provide a positive account of self-knowledge of belief (§2.2.2.2).

[7] Some have tried to capture the account under the label 'rationalism' about self-knowledge (e.g., Cassam (2014) and Gertler (2011)). But this confuses matters due to inconsistent use of the term.

2.2.2.1 Agentialism and rational agency

Agentialists, as do I, locate various key features of self-knowledge in our capacity for rational agency and our resulting responsibility. Now is therefore a good time to say more about these concepts.[8] Being responsible for an attitude is to be 'accountable' (McHugh 2013: 132), and potentially 'epistemically praise- or blameworthy for it' (Hieronymi 2008: 363). This is similar to the way in which we hold each other morally accountable. For example, if you were to kick someone, I would tell you that you're acting immorally; in this case, you have failed to fulfil the obligation not to kick people. Similarly, we bear various epistemic obligations and can be criticised if we fail to fulfil them. This would include a defeasible obligation to believe only if we have normative reasons. E.g., I may well say that you're being irrational if you believe that *p* on the basis of what is in fact bad evidence. I also take it that we bear an obligation to believe on the basis of reasons even if they are in fact bad. After all, even if a belief does not cohere with what are in fact good reasons, we still do something right in believing on the basis of reasons at all. Believing on the basis of the reason that *p* rationalises our belief in our eyes, and makes *p* seem, at least to a degree, like a sensible thing to believe. And this is because believing for the reason that *p* in some way involves taking *p* to be a good reason for the belief. Further, we bear this responsibility because believing is something we do. This is comparable to how we are responsible for actions we perform (intentionally kicking someone) but not for movements we lack agency over (if your leg spasmed so that kicked someone). I discuss these issues further in chapter seven.

Like Moran, the agentialist project in general rejects accounts of self-knowledge that conflict with our rational agency.[9] I understand these ideas in the following way (I found Burge 1996 especially helpful). Our self-ascriptions bear direct rational relations to the attitudes they concern.[10] As a result, our attitudes can, and should, seem rational from our perspective (I further discuss this notion in chapter seven). If I believe that *p*, it should make sense to me to believe that *p.* This will generally be a matter of having a motivating reason for the belief. So, for my belief that *p* to be rational by my lights, I only need to have a motivating reason for believing that *p*. And for my belief that *Beth believes that p* to be rational by my lights, I only need a motivating reason for believing that Beth has this belief. But, my belief that *p* directly affects whether my belief that *I*

[8] See Boyle (2011b, 2009b), Hieronymi (2008), and McHugh (2013).
[9] For indicative comments see especially Boyle (2015: 344-5) and Burge (1996: 110; 1999: 32).
[10] See especially (Boyle 2011a: 236) and Burge (1996: 114).

believe that p seems rational from my perspective, and vice versa.[11] Our account of self-knowledge must allow for these direct rational relations. Yet views like quasi perception do not. According to the norms of the quasi-perceptual process, the rationality of the target mental state does not affect the rationality of the self-ascription. After all, the norms of quasi-perception will be relevantly similar to the norms governing perception. Suppose for a moment that one could perceive other people's beliefs. My perceptually-formed belief that *Beth believes that p* would be equally rational when I believe that *p is false* as when I believe that *p is true*. Indeed, under an account like Armstrong's, where self- and perceptual knowledge alike are warranted on reliabilist grounds, even the self-ascription would not always seem rational from the one's own point of view.

2.2.2.2 The agentialist account of self-knowledge

Regarding the method involved, my version of a paradigm agentialist view says that we acquire self-knowledge using TM (contra Burge but agreeing with Boyle and O'Brien).[12] This indeed seems like an intuitive picture of what goes on – I do look outward when learning of my attitudes.

But I should emphasise that we can use TM to learn of attitudes that we already possess; so far, I have been considering cases where the subject both learns that they have the belief and acquires the belief itself. Recall that in NICARAGUA TWO, you had not yet made up your mind about the capital of Nicaragua and had to investigate what it is. But we also seem to be able to use TM to learn of pre-existing attitudes (see Byrne 2011: 208-9; Shah and Velleman 2005: 506-8). For example, it may well be that you have known for a long time that the capital of Nicaragua is Managua. Yet, as Boyle (2011a, 2011b) argues, you can still learn of this belief by considering whether the capital of Nicaragua is Managua and concluding 'yes.' Even though you formed the belief in the past, when you answer the world-directed question now:

> What [you] call to mind must be not merely [your] past assessment of the question, but [your] present assessment of it — the answer to the question whether P that presently strikes [you] as correct (Boyle 2011b: 10).

Despite already believing that *the capital of Nicaragua is Managua*, you re-open the world-directed question so to speak, since it is up to you to change your mind and revise your belief in

[11] Regarding the thought that the rationality of the belief that *p* affects the rationality of its self-ascription see Boyle (2015: 344-5). Concerning the inverse – that the self-ascription affects the rationality of the belief that *p* – see e.g., Burge (1996). I take rational relations to hold in both these directions.

[12] Indeed, Burge may well take the method to be quasi-perceptual. See Burge (1996: f.n.12) and Gertler (2011: 185 inc. f.n. 7) for discussion. I should note that TM can also be formulated in non-agentialist ways, e.g. see Evans (1982), Byrne (2011, 2005) and Fernández (2013, 2003).

line with the evidence. It is not that you make up your mind and then leave it be; rather, believing, Boyle takes it, is a continuing activity: you still exercise your rational agency in virtue of a basing relation between the attitude and your motivating reasons for it.[13] This is not to say that you always explicitly think about the considerations, but still, you can access this justification if needed, and your belief is sensitive to it. You would change your mind if you recognise that the considerations are weak and you would judge instead that Managua is not the capital of Nicaragua. I have sympathy with Boyle's position, but need not argue for it here due to the thesis' focus.

At this point I should emphasise that TM is non-inferential. This is the case even though employing TM involves transitioning from one content – that *p* – to another content – I believe that *p*. Still, TM cannot be inferential since it follows different norms to inference. We can see this in how the transitions involved in TM are rational; the transitions satisfy whatever norms govern TM. But if TM was an inference it would be a bad one. To infer rationally, the content of the premises must in some way support the conclusion. For example, I can rationally infer as follows: 'the dessert is a Bakewell tart; Bakewell tarts contain ground almonds; therefore, the dessert contains ground almonds.' The fact that the dessert is a Bakewell tart, and the fact that Bakewell tarts contain ground almonds, together provide evidence that the dessert contains ground almonds. The process means that it is rational to hold the concluding belief. But the content of the representations involved in TM do not appropriately support the conclusion. I transition from *p* to *I believe that p*. Yet the fact that *p* is not good evidence that I believe that *p* – people frequently are ignorant.

What about the warrant for self-knowledge acquired using TM? Agentialist accounts ground our warrant in rational agency, but we need not, and should not, follow Moran's account. Recall that Moran appealed to a transcendental entitlement.[14] But transcendental entitlements are highly controversial. Also, it is unclear that something of this nature could capture the way in which the self-ascription is sensible to believe by the subject's own lights. Just because the nature of deliberation entitles one to assume that one's conclusions are one's beliefs, it does not mean that the subject herself is aware of this. Indeed, under such an account, the subject need not be aware of anything relevant at all.[15]

[13] I take this to be the case from (2011: 236) and (*Ibid.* 236 f.n.15) in which Boyle links the point to Byrne's criticism. See also (2011b: esp. 9-10).
[14] Burge also thinks that self-knowledge is transcendentally warranted (1996: esp. 98-103).
[15] On these points, see O'Brien's criticism of 'top-down' theories of rational entitlement (2005: 593) and O'Brien (2003).

Instead, I endorse an approach whereby rational agency provides subjects with some sort of awareness of their attitudes, where we should understand this awareness in experiential terms and not purely epistemically.[16] So, I have in mind the way in which a subject can be aware of something in undergoing an experience, e.g., when James is aware of the laptop in front of him. Subjects can also be 'aware of' something in a colloquial sense according to which the subject has access to a fact. E.g., when James is aware that Bakewell tarts contain ground almonds, or when James is aware that his friend is busy. Subjects' epistemic awareness is sometimes underpinned by experiential awareness, e.g., James will have access to the fact that there is a laptop in front of him in virtue of undergoing a perceptual experience of the laptop. But, epistemic and experiential awareness come apart – James is not aware that his friend is busy in virtue of a perceptual experience of his friend's hectic schedule. Turning to the matter at hand, I think that subjects ultimately have epistemic awareness of their attitudes in virtue of some sort of experiential awareness of those attitudes. I say 'some sort of awareness', though, because subjects do not relate to their attitudes as they would a laptop.

I'll sketch out the most promising way of grounding the warrant for self-knowledge in experiential awareness.[17] O'Brien (2007) argues that subjects have a unique 'agent's awareness' of attitudes such as judgements (which O'Brien takes to be mental actions) and that this is because of the agency subjects have in relation to them. We might also term this awareness 'practical awareness' O'Brien (2003: 381). This awareness is most obvious when subjects form a judgement through deliberation. S is aware of the judgement that *p* and of the judgement as *her* judgement. S is aware of the judgement as *hers* because considering the options regarding what action to perform provides her with an awareness that she is the agent in question. This is because considering the options, in O'Brien's words, 'carries with it the idea of an assessment by an agent of actions *for her*. For a subject to engage in an assessment of what to do is for a subject to determine what *she* should do' (O'Brien 2007: 117). And why is the subject warranted in ascribing a *judgement* with a *particular content* to herself? The subject has an agent's awareness of judging that *p* (as opposed to, say, wondering that *p* or judging that *q*) because she has the agent's awareness in virtue of deciding whether to judge that *p.* S is aware of judging that *p* as an option for her, and so in concluding that *p,* S is aware of her conclusion as the formation of a judgement that *p*. As O'Brien writes: 'X is warranted in taking herself to be *judging* that *P* because X has

[16] O'Brien's distinction between *top-down* and *bottom-up* accounts of rational entitlement is also relevant when considering the differences between rational agency views (2005).
[17] Boyle (n.d.) suggests another way, which draws on Sartre's (1956) notion of *non-positional consciousness*. This picture is less concrete compared to O'Brien (2007) though. Relatedly, see the Roessler and Eilan (2003) edited collection *Agency and Self-Awareness*, which contains relevant discussions of agent's awareness regarding intentional action.

concluded that *P* is true which is, in the context, equivalent to her realising the practically known possibility of judging that *P* on the basis of a consideration of whether *P*' (O'Brien 2005: 594).[18] So, the thought goes, subjects have a special agent's awareness of their judging that *p* in virtue of forming the judgement. This awareness then warrants the subject in self-ascribing the judgement.

I've set out the warrant and method under agentialism; let me highlight two other features of Moran's picture we should preserve. My paradigm agentialist account takes it that our rational agency grounds first-person authority. Specifically, is a special rational authority – it is not just that self-ascriptions are more reliable than other-ascriptions. And, further, we bear rational requirements to be able to acquire self-knowledge using TM. If we are unable to do so, then the attitude in question is irrational.[19]

***

To summarise, we can return to the possible ways in which we might think that self-knowledge is distinctive, and contrast agentialism with quasi-perceptual accounts in the following table.

[18] I think this suffices to account for our warrant for ascribing an attitude with the particular *content*, although O'Brien suggests something else (2005: 581, 594). Also, I should note that O'Brien in (2005) is less explicit in cashing things out in terms of agent's awareness, but we can use her discussion in this way.
[19] E.g., (Burge 1996: 103).

	Agentialism	**Quasi-perception**
Extra-reliable	Yes	Yes
Self-intimation	Necessarily, we will be in a position to know our attitudes if we are rational/the attitudes are rational	Depends
Distinctive method and warrant	Yes (in practice)	Yes
First-person authority	Yes – based in our responsibility	Yes – based in our reliability
Role of rational agency	Yes. This grounds the other features	No
Extra features	Rational obligations for self-knowledge	No

We have seen that quasi-perceptual and agentialist theories disagree about a lot. For the agentialist, self-knowledge fundamentally differs from other-knowledge in virtue of the agential relation we bear to our attitudes. And for the quasi-perceptual theorist, it is because we acquire self-knowledge using an inwardly-directed detection mechanism. But both views have something in common. They both agree that self-knowledge significantly differs from other-knowledge, even though they diverge on the way in which it does. In this way both theories disagree with our next account which takes self- and other-knowledge to fundamentally resemble each other.

2.3 Self-knowledge of belief: Neo-Ryleanism

In contrast to quasi-perceptual and agentialist accounts, neo-Ryleanism contends that self- and other-knowledge are fundamentally the same. Neo-Ryleanism offer a newer and more plausible

version of ideas proposed by Gilbert Ryle. This section introduces Quassim Cassam's account (§2.3.1), before exploring the position more broadly (§2.3.2).

2.3.1 Cassam's Neo-Rylean account of self-knowledge

For Cassam, the following exemplifies the way in which we acquire self-knowledge:

> NICARAGUA THREE. Say I ask you whether you believe that *the capital of Nicaragua is Managua*. You might think about how you have been mentally saying to yourself that 'Managua is the capital of Nicaragua,' how you gave that answer in a pub quiz the other day, and your feeling of certainty when you consider the proposition that *Managua is the capital*. You then infer from these facts about yourself that you believe that *the capital of Nicaragua is Managua*.

Cassam thinks that self-knowledge is fundamentally akin to our knowledge of other people. His picture applies to a wide range of instances of self-knowledge, encompassing that of our attitudes as well as features such as 'one's character, values, emotions, and abilities' (2014: 171).[20] Cassam allows that 'simple feelings or sensations like nausea and pain' may be exceptions (*Ibid*. 164), but maintains that the vast majority of self-knowledge resembles other-knowledge.

Cassam contends that both self- and other-knowledge are inferential in virtue of the method and warrant involved.[21] To consider the method first, the subject forms the self-ascriptive belief by engaging in inference. For Cassam, this occurs in one's psychology at either the conscious or unconscious level (2014: 138-9). The evidential base includes, among other things, feelings, mental images, judgements, and other mental goings on (e.g., *Ibid*. 138, 162).

Second, self-knowledge is also warranted inferentially (e.g., *Ibid.* 139). I take it that a belief is inferentially justified in virtue of the subject's justification for other true beliefs on which they base it.[22] It is worth clarifying that, as Cassam also notes, a belief can be inferentially justified without being formed by inference. It is enough to hold the belief on the basis of a justified belief (I further discuss basing relations in chapter seven). Cassam takes it that the beliefs constituting the evidential base are also acquired and warranted inferentially. Indeed, he also takes it that the beliefs constituting the evidential base for these supporting beliefs will themselves also be acquired and warranted inferentially, and may even include facts concerning one's attitudes. In

[20] Even proponents of distinctive access accept that such features exceed its scope. Cassam takes it to be a benefit of his account, though, that it applies to a broad range of features.
[21] We need not understand other-knowledge inferentially, but Cassam argues for this in (2017: §4).
[22] Cassam seems to have this formulation of inferential justification in mind in (*Ibid.* 166) although not in (*Ibid.* 139, 153, and 165). I take the formulation I use in the main body to be the most charitable.

claiming that the supporting beliefs are inferential *all the way back* in this way, Cassam therefore commits to taking the warrant to be circular. But Cassam denies that the warrant would be viciously so, provided 'the interpretive circle is wide enough' (*Ibid*. 165, 169). In this way, he draws on coherentism about justification.[23]

Further, Cassam thinks that self-knowledge is akin to other-knowledge in other ways too, as well as its method and warrant. One such way is that self-knowledge also isn't any more reliable. Cassam writes that subjects often lack self-knowledge and that his inferential picture accommodates this (*Ibid.* esp. chapter 11). Indeed, it is not just that Cassam thinks that as a matter of fact individuals lack distinctive access to their mental states. Further, he holds that subjects are not *required* to have such access. He takes it that self-ignorance and error do not always represent a rational failure on the part of the subject, in this way disagreeing with the agentialists (*Ibid*. 197).

2.3.2 Neo-Rylean accounts in general

I class Cassam's account, along with Peter Carruthers' (2013, 2010) as *neo-Rylean*.[24] I take the central claim of neo-Ryleanism to be that the vast majority of self-knowledge (such as self-knowledge of belief) fundamentally resembles other-knowledge.[25] Under the account, both self- and other-knowledge are inferential (well, inferential broadly construed, but I'll elaborate shortly). Further, the evidential base includes various mental features (Carruthers places a lot of importance on sensory evidence, including mental images and inner speech (2001:69)). In this way, neo-Rylean views differ from Gilbert Ryle's. Ryle (2009) also denies that we have distinctive self-knowledge and is generally interpreted as thinking that all self-knowledge is inferential. But Ryle restricts the evidence to inner speech and facts about our behaviour.

Core similarities aside, I take there to be two main types of neo-Ryleanism. The difference concerns the subpersonal/personal distinction. Cassam locates his account at the personal level of explanation.[26] But we could also locate this sort of picture at the subpersonal level, as Carruthers

[23] Indeed, coherentists generally are committed to self-knowledge being inferential (thanks to Kurt Sylvan for pointing this out). See for example BonJour (1985).
[24] There are other accounts in the vicinity but which are not as extreme, e.g., Wilson (2002) and Schwitzgebel (2012, 2009, 2008, n.d.; Hurlburt and Schwitzgebel 2007).
[25] For exceptions in Carruthers account, see Carruthers (2013: 378).
[26] Cassam's defence against the charge that inferentialism over-intellectualises self-knowledge is that the inferences in question can be instances of 'fast' thinking, which is 'automatic, effortless, and barely conscious' (2014: 140). Further, he clarifies elsewhere that he does not see fast thinking as subpersonal (2014: 17 f.n. 2). I happen to think that fast thinking would be subpersonal, but at any rate, Cassam cannot think that the processing would be subpersonal if he wants it to give rise to inferential justification, as he does.

does. The subpersonal/personal distinction bears on both the method and warrant for self-knowledge. This subsection introduces the subpersonal/personal distinction (§2.3.2.1) which indeed bears great significance for the thesis and proves crucial to my own view. The following subsection then draws on this distinction to delineate the two versions of neo-Ryleanism (§2.3.2.2).

2.3.2.1 The subpersonal/personal distinction

Let me start with an analogy.[27] We can talk of an orchestra in several ways. Firstly, we can talk of the orchestra itself, and how it plays a Bach suite, moves from the allegro to the adagio, and crescendoes at the end. Alternatively, we might talk in terms of the players themselves. For example, we can say that the lead violinist came in late, even though the orchestra itself did not. Or we might say that each individual player is good, where this is not to say the same about the orchestra as a whole. We can distinguish, then, between orchestra and sub-orchestra ways of talking. And indeed, this is not just a matter using of different descriptions – we can offer different explanations. In explaining why music filled the hall, we might say that the orchestra played Bach with a crescendo at the end. But, at the level of players, perhaps we might say that each player played a part (apart from the lead violinist who didn't even play the whole piece). And when explaining why the performance was good, we would do so in terms of what makes an orchestra good, e.g., consistent timings.

Similarly, we can explain subjects at both the personal and subpersonal levels (though this is not to say that the orchestra case is exactly analogous).[28] For example, we can say that Ben bought a Godzilla film because he enjoys them, and that he is happy as a result of the film. These movements and states are all attributable to Ben, the person, and therefore occur at the personal level. Or alternatively we might say that Ben bought the film because of certain neural reactions which then caused his joints and tendons to work in specific ways. As a result, serotonin was released leading to chemical changes in the brain. Just as we do not say that the whole orchestra comes in late, similarly, *Ben* did not contract or work in these specific ways – his neurons and joints did. His neurons and joints therefore reside at the subpersonal level.[29]

---

[27] On the subpersonal/personal distinction, see especially Bermúdez (2005, 2000) Drayson (2014, 2012), and Hornsby (1997).

[28] I talk of personal/subpersonal levels for ease of presentation. I am open to the possibility that, while we can talk of personal/subpersonal explanations, we cannot make a metaphysical claim that there are personal and subpersonal states (see Drayson (2012)). Even then, though, this is not to say that we abandon making any metaphysical claims when making separate personal and subpersonal level explanations. The explanations appeal to different sorts of states – doxastic and subdoxastic states respectively.

[29] Unlike personal explanations, subpersonal explanations are often grounding explanations. See Bermúdez (2005: 31-33).

2.3.2.2 Subpersonal/personal neo-Rylean accounts

Neo-Rylean accounts can pertain to the personal or subpersonal level.[30] The account we've looked at so far, Cassam's, operates at the personal level. For Cassam, self-ascriptions are formed inferentially or at the very least based on supporting beliefs. This can be attributed to the subject herself. Accordingly, the resulting self-ascription is inferentially warranted.

Alternatively, there is also a subpersonal version of neo-Ryleanism. I take Carruthers as providing an account like this. He thinks that both self- and other-knowledge are acquired by a 'mindreading' module (e.g. 2011: 260), and therefore a distinct mechanism with its 'own neural realisation' (*Ibid.* 227). This is distinct from the idea that mindreading results from 'domain-general theorising' (*Ibid.* 227) in which the subject acquires theories that can then be used by multiple mechanisms. The mindreading module reaches its conclusion via *computation*. Computation involves transitioning between contents, like inference, but is not something that the subject herself engages in. As such, the process is subpersonal.[31]

Accordingly, the warrant for self-knowledge under this subpersonal version differs from Cassam's – it is reliabilist. Carruthers can't say that the ascriptions are warranted inferentially (although Carruthers himself does not discuss this). While there are many debates to be had regarding the nature of inferential justification, it seems very reasonable to think that it requires the subject to have personal level access to the propositions that the belief is based on. Yet if the process in question is subpersonal then this may not be the case. The subject may not have access to the representations from which the mindreading module transitioned to the self-ascriptive belief. And even if the subject did, they may not see the representation as at all relevant to the matter at hand. I return to the question of the warrant for self-knowledge in chapter five.

The thesis, then, uses the following distinction. We might think that self-knowledge is acquired in the same way as other knowledge because it results from *computation* which occurs at the subpersonal level. Or we might think that we acquire it using *inference*, at the personal level. These are not wholly distinct options though – one can, and should, think that computation underpins the personal-level inference. I will term an account relying on inference (even if

[30] For a helpful discussion of the subpersonal/personal distinction regarding self-knowledge and immediacy, see Jongepier and Strijbos (2015).

[31] While not in (2013), Carruthers does seem to deny that the process is subpersonal in (2010: 93-4). Yet I question this for the reasons stated and for others. At any rate, even if Carruthers himself does not endorse it, the account I have attributed to him occupies logical space. Having this position in mind will be vital for the thesis. I will continue terming it his view for simplicity's sake.

computation underpins it) an *inferentialist* one, and one appealing just to computation, a *computationalist* one.

***

We've now considered three main positions in the self-knowledge literature and can express their commitments in the following table. Agentialism conflicts with quasi-perception and neo-Ryleansim in emphasising a fundamental role for rational agency. But both agentialist and quasi-perceptual accounts nevertheless agree that self-knowledge differs fundamentally from other-knowledge, which is something the neo-Ryleans reject.

	Agentialism about belief	**Quasi-perception about belief**	**Neo-Ryleanism (Computationalism and Inferentialism) about belief**
Extra-reliable	Yes	Yes	No
Self-intimation	Necessarily, we will be in a position to know our attitudes if we are rational/the attitudes are rational	Depends	No
Distinctive method and warrant	Yes	Yes (in practice)	No
First-person authority	Yes – based in our responsibility	Yes – based in our reliability	No
Role of rational agency	Yes	No	No
Extra features	Rational obligations for self-knowledge	No	No

So far, I have considered the state of play regarding self-knowledge of belief. But what about the specific topic of the thesis – self-knowledge of why we have our attitudes and perform actions? I want to build on the foregoing discussion to set out two options. The first is the orthodox position. Here the neo-Ryleans and quasi-perceptual theorists tend to converge and think that self-and other-knowledge of why we have our attitudes fundamentally resemble each other. I set this view out in §2.4. Alternatively, agentialists seem to assume that we can learn of one particular explanation of our attitudes in a distinctive way – reason explanations. I.e., subjects can learn of their motivating reason in a distinctive way, although not any purely causal factors. The agentialists do not, though, really spell out what this would look like. So, I will formulate what I take to be the best version of an agentialist account regarding motivating reasons (§2.5). It is worth taking my time to set this out, since the thesis will go on to argue for it in chapter six (albeit combined with computationalism). We will end, then, with two main options on the table for understanding self-knowledge of why we have our attitudes and perform actions: the Orthodoxy and agentialism.

2.4 Self-knowledge of motivating reasons: The Orthodoxy

The orthodox position takes self-knowledge of why we have our attitudes and perform actions as fundamentally akin to other-knowledge. What I will term 'the Orthodoxy' claims that we lack distinctive access to 'why' we have our attitudes in a blanket sense, without differentiating between types of explanation. The following two chapters introduce arguments for this. (To anticipate, one thought is that reasons are causes and we can only learn of causal factors by inference.)

Even if one denies neo-Ryleanism about self-knowledge in general, it is common to agree with it about self-knowledge of why we hold our attitudes. If we recall the list from earlier, even if generally philosophers answer yes in the various rows, they standardly answer 'no' regarding why we have our attitudes. E.g., we get this explicitly in Gertler's (2011) seminal survey book on self-knowledge, who herself advocates a quasi-perceptual account in other contexts.[32] Gertler states that we lack a distinctive method or reliable access to why we hold our attitudes (2011: 72-5). We also see similar explicit denials in Nisbett and Wilson (1977), Rey (2008), Nichols and Stitch (2003), and Schwitzgebel (2016: §4.2.1).

[32] E.g., Gertler (2012).

And, while philosophers sometimes explicitly limit the scope of distinctive self-knowledge in this way, often it is implicit. The literature simply rarely discusses the issue. Broadly, it is notable that philosophers often argue that we have distinctive self-knowledge of 'mental states.' For example, Armstrong writes that 'in introspection, we have direct, non-inferential, awareness of our mental states' (2001: 124). And we get this in two introductions of edited collections – Smithies and Stoljar (2012: 4) and Coliva (2012: 1). But roughly, why it is we have an attitude won't be a mental state in the relevant sense. There might be a mental state such as a belief that happens to cause us to have an attitude. But the explanation of an attitude won't be a mental state per se. Rather, it will be a fact that connects an explanans (which may indeed be a mental state) to the explanandum.[33] Indeed, the fact that philosophers do not feel the need to explicitly limit the scope of distinctive access to exclude explanatory facts further illustrates the Orthodoxy's pervasiveness.

I should note that there are two versions of the Orthodoxy, although its proponents do not distinguish them. The thought goes that subjects use the same method to explain both their own attitudes and other peoples', but we could construe this method along either *inferentialist* or *computationalist* lines. I suspect Nisbett and Wilson (1977) at any rate take the method to be subpersonal, but it does not actually matter who subscribes to inferentialism and who subscribes to computationalism. Regardless of who holds the positions, there are two possible versions of the Orthodoxy and this distinction will prove important to my project.

The Orthodoxy and the shape it takes should become clearer throughout the next chapter. There I will argue that the question of self-knowledge of why we have our attitudes is an open one, despite the seeming force of the Orthodoxy. In doing so, I will introduce their arguments for the position. For now, let us note the following table. It will help to sketch out the view in stark terms to show its extreme nature.

[33] Perhaps, though, Armstrong would happily extend his account to reasons. One might object to extending self-knowledge to motivating reasons by saying that reasons have a causal element and we cannot have non-inferential knowledge of causes. (I discuss this in chapter three.) Yet, for Armstrong, we have introspective access to mental states, where even these are fundamentally causal features (2001: 326).

	The Orthodoxy about self-knowledge of why we have our attitudes and perform actions
Extra-reliable	No
Self-intimation	No
Distinctive method and warrant	No
First-person authority	No
Role of rational agency	No (not mentioned)
Extra features	No

2.5 Self-knowledge of motivating reasons: Agentialism

Although the Orthodoxy is, well, the orthodox position, not everyone endorses it. Agentialists specifically seem to reject it. This resistance concerns one type of explanation in particular – reason explanations. That is, agentialists assume that we have distinctive access to the reasons for which we have our attitudes, although not any purely causal factors. This view can plausibly be attributed to agentialists Boyle (2011a, 2011b), Burge (1999, 1996), Cox (2018), and Moran (2001), as well as Davidson (1963), Leite (2008, 2004), Sandis (2015), and Setiya (2013).[34]

The agentialist account of self-knowledge indeed provides the resources with which to contest the Orthodoxy. According to agentialism, our reasons for holding an attitude seem to play an important role in how we learn of those attitudes. After all, we learn of our attitudes in virtue of our rational agency, which involves holding attitudes on the basis of reasons.

That all said, the literature rarely discusses the prospect of distinctive self-knowledge of why we hold our attitudes. While there are promising ideas, often the discussions are brief (e.g. Boyle 2011a: 8 and Leite 2004: 226) or aren't sufficient for my purposes. I'll say a couple of things about what our account should look like, before proposing what I take to be the best agentialist picture of self-knowledge of motivating reasons.

[34] Cox (2018) does not use the term 'agentialism' and expressed concerns with it in correspondence. But since his view is sufficiently alike those of Moran and Boyle, I term it an agentialist one for simplicity.

Our account requires (at least) two things. First, we need to make sure to explain self-knowledge of motivating reasons as opposed to related phenomena. This thesis concerns how a subject knows that she, say, *believes that it will rain for the reason that there are no clouds in the sky*. It is not that in answering the question 'why do you believe that it will rain?' the subject simply *provides* the absence of clouds as a normative reason. And neither is it just that she can *express* her motivating reason in providing that justification (as we get in Cassam's version of what an agentialist might say in 2014: 199). Rather, the subject can also *know that it is her reason*. And indeed, I am interested in what it is to know one's motivating reason where I take this to be a subtype of explanatory reason (contra Anscombe (2000) and Moran (2001, esp. p. 128)).

Second, our account should explain how the subject learns what her motivating reason is *simpliciter*, and not just how she learns whether or not she has a given motivating reason. Cox (2018) provides a systematic account of how subjects learn of their motivating reasons, and he advocates a transparency method. Yet Cox cashes it out in terms of learning whether *p* is one's motivating reason or not – the subject does so by considering whether to treat *p* as a normative reason (e.g., Cox 2018: 193). But it won't always be the case that subjects want to know whether or not a given consideration is their reason; often they just want to learn the more general fact of what their reason is. We see this in the canonical example of self-knowledge of reasons – when someone answers the question 'why?'. Here, both the questioner and the subject herself want to know why the subject has an attitude *simpliciter*, not the specific fact of whether or not a given consideration is her motivating reason.

I now want to motivate my own agentialist picture of the matter at hand. A foil will help this. To form one, I will charitably flesh out one brief suggestion from the literature into a fully-fledged view but argue that even this charitable version faces worries (§2.5.1). By contrast, my own account is plausible and intuitive (§2.5.2). To be upfront, I take the best alternative to the Orthodoxy to be a specific version of a transparency method. According to this, we learn of our reasons for having an attitude by answering the question 'why have that attitude?' where this amounts to the question 'what are the normative reasons for having that attitude?'.

Before continuing, though, I should note two things about the following discussion and its purpose. First, I extend the agentialist account of self-knowledge of belief to self-knowledge of motivating reasons, but I do not do the same for the quasi-perceptual account. A quasi-perceptual account of self-knowledge of motivating reasons does indeed occupy logical space, and furthermore, the position is more plausible than some might think (see the next chapter). The quasi-perceptual theorist might say that S learns that she believes that *q* for the reason that *p* because S quasi-perceives that she believes that *q* for the reason that *p*. But a quasi-perceptual

account is relevantly similar to the Orthodoxy in denying that rational agency grounds self-knowledge of motivating reasons, and so it faces some of the same problems. (I criticise the Orthodoxy in chapter five.) Therefore, in arguing for my position against the Orthodoxy, I will also be arguing against a quasi-perceptual account. This thesis will focus on the Orthodoxy and agentialism due to space. Second, the following takes some time in arguing for the best way of cashing out agentialist insights into an account of self-knowledge of motivating reasons. Indeed, it does so before I have even argued that we should draw on agentialist insights at all. The reader therefore may well wonder why I undertake this argumentative work and/or why I undertake it so early on. But the following discussion will prove vital for this thesis' overall position; I will endorse a qualified version of the account I set out below. Further, understanding the way in which the Orthodoxy troubles agentialism will be easier if we have a concrete and specific agentialist account on the table. Having in mind the account I set out below should make the dialectic easier to grasp in chapter four.

2.5.1 Boyle's account of self-knowledge of motivating reasons and a problem

Boyle gives us the beginnings of an agentialist account of self-knowledge of motivating reasons. In a discussion of deliberation, he tells us that:

> [I]f I reason "P, so Q", this must normally put me in a position, not merely to know that I believe Q, but to know something about why I believe Q , namely, because I believe that P and that P shows that Q (Boyle 2011a: 8).

Boyle also spells out what being 'in a position to know' amounts to. Boyle claims that if one deliberates in this way, one 'normally needs no further grounds in order knowledgeably to judge *I believe P because I believe Q*' (*Ibid.* 8).[35] So, to learn that I have the motivating reason that Q, I don't require any evidence about myself – reasoning with Q suffices. For example, say I'm deliberating about the weather and judge that 'there are grey clouds, so it'll rain.' In order to know that the grey clouds are my motivating reason, I don't require any further premises about me being rational or the like.

I will flesh out Boyle's suggestion into a more substantial picture (§2.5.2.1) before raising a worry (§5.2.2).

[35] We can contrast this with Williamson (2002).

2.5.1.1 A reconstructed Boylean account

I want to put some meat on the bones in two ways.

First, how would Boyle's suggestions apply to subjects' motivating reasons for attitudes other than belief? After all, subjects also seem to have a special way of learning the reason for which they, say, intend to *φ*. Boyle's account here depends on what he takes to be the reasoning subjects employ when forming these other attitudes. When forming a belief that *p,* subjects consider whether *p* is true, but when forming an intention to *φ*, they don't consider whether *φ*ing is true.[36]

In constructing a Boylean account of self-knowledge of our reasons for attitudes other than belief, I can draw on what Boyle writes elsewhere. For Boyle, we should think of a range of 'mental state[s] as constituted by the subject's knowingly evaluating a certain content in a certain way' and that:

> [T]here appears to be a connection between intending to do *A* and regarding *A* as to be done, desiring some object *O* and regarding one's having *O* as desirable, hoping that *P* and regarding *P* as a possibility whose realization would be good, etc. (2011b: 237).

So, Boyle thinks that there is an important relation between holding an attitude and evaluating the attitudes' object. For ease of presentation, I will use *Fa* as a generic placeholder for the evaluation appropriate to the attitude at hand, e.g., that the attitude's object is true, to be done, or desirable.[37] Boyle could say, then, that in considering whether to hold an attitude, we consider whether the object of the attitude is Fa. We can now put Boyle's claim about reasons for belief more generally and characterise the Boylean transparency method as:

> BTM: if S reasons '*p*, so the object is *Fa*,' this must put S in a position to know that she has attitude A for the reason that *p*.

For example, if I reason 'the seminar will be interesting, so going to the seminar is to be done,' this must put me in a position to know that *I intend to go to the seminar for the reason that it will be interesting.*

To clarify, in the interests of charity I understand 'Fa' to be a *pro tanto* notion as opposed to an *all things considered* one. This is because subjects can sometimes hold an attitude on the basis of reasons even though they do not take *o* to be *Fa* all things considered. For example, consider the

[36] C.f. Way (2007).
[37] The evaluation '*p* is true' will often be truncated to simply '*p*.'

following everyday scenario, put in rough terms: I decide to get a Bakewell tart because they are so delicious, even though I know I shouldn't – I had one yesterday, and there's celery in the fridge. There is a sense in which I form the intention by concluding that getting the tart is to be done because it is delicious. It is just that I do not conclude that it is to be done *all things considered*. But still, Boyle might say that forming my intention on the basis of this reason – the tart's deliciousness – puts me in a position to ascribe that motivating reason.[38]

Second, how would the subject capitalise on her epistemic position and actually acquire knowledge? I.e., in what way does she form the belief self-ascribing her motivating reasons? In extending the paradigm agentialist account from earlier, we are trying to formulate a transparency *method* after all. And as it stands, Boyle's comments only apply to self-knowledge of reasons for attitudes that subjects have just deliberated over. It doesn't yet tell us what goes on when you ask me why I intend to go to the seminar and I tell you.[39] This is akin to the broader worry we encountered earlier: how does TM extend to beliefs the subject has held for some time? Recall Boyle's response – the subject nevertheless learns of a pre-existing belief that *p* by considering whether *p* is true. The subject still makes up her mind in concluding that *p* is true, since she could have changed her mind in line with the evidence and stopped believing that *p*. For Boyle, the subject evaluates the object of the attitude afresh to learn of what her attitude currently is, even if she initially formed the attitude at an earlier date.

So, my way of cashing out Boyle's idea is as follows:

> BOYLEAN TM FOR MOTIVATING REASONS (BTM): To learn of her reason for having attitude A, S considers whether the object of A is Fa. In concluding that '*p*, so the object is Fa,' S then then ascribes *p* as her motivating reason.

Say I intend to go to the seminar and you ask me why. I reconsider whether going to the seminar is to be done and conclude that the seminar will be interesting so going to the seminar is to be done. I can thereby tell you that I intend to go to the seminar for the reason that it will be interesting.

BTM is comparable to one of two options which Leite (2004) runs together. He writes that a subject learns of their reasons for a belief by considering 'whether […] [they] should hold it' and thereby '[reconsider] the issue at hand' (Leite 2004: 226). We may or may not think that reconsidering whether o is F is a way of reconsidering whether one should hold the attitude. But

[38] For related discussion see Boyle (2015)
[39] Cassam raises this issue in (2014).

we can nevertheless note that BTM falls under one general strategy for self-knowledge of motivating reasons: subjects learn of their reasons for an attitude by reconsidering the relevant issue as a whole.

2.5.1.2 The Boylean TM for motivating reasons (BTM) rejected

Intuitively it is simply implausible to say that subjects answer the question 'why?' using BTM. I present this objection as an empirical claim about what happens. Perhaps BTM could provide subjects with knowledge, but they do not standardly use it. To start, consider how unusual the reasoning process looks. I hope it already seemed alien to the reader when I was setting it out. Using BTM would involve something like the following, at least unconsciously:

> Why do I believe that it will rain? Is it true that it will rain? There are grey clouds, so it is true it will rain. Therefore, I believe that it will rain for the reason that there are grey clouds.

According to the BTM account, subjects answer the question 'why?' concerning, say, a belief, by considering whether the belief is true; the subjects somehow hope that knowledge of their reasons for the belief falls out of this whole process. This seems very odd. Appeal to brute intuition alone, though, rarely passes philosophical muster. In the following, I will outline three ways of cashing out this intuition that subjects do not in fact use anything approaching BTM when answering the question 'why?'.

i. If subjects answered the question 'why?' by reopening the question 'is *q* true?', there would be occasions in which they change their mind. Subjects would at least sometimes conclude that *q* is false and revise their belief. As a result, subjects would not ascribe reasons for the original belief that *q* at all. And yet, individuals rarely answer the question regarding why they believe that *q* with 'no, that's wrong. I don't believe that *q*.'

ii. One need not explicitly deliberate when using BTM, which means that using BTM would not always issue in a self-ascription of one's motivating reasons. After all, considering whether the object is *Fa* does not always involve explicit deliberation.[40] Sometimes deciding whether the object is *Fa* will be a simple matter of judging either *yes* or *no*. Accordingly, subjects using BTM would consider whether the object is *Fa* and frequently conclude *yes* or *no* without consciously considering normative reasons. As a result, subjects would often answer the question 'why?' with just 'yes, I have the belief' or 'no, I do not have the belief.' To give an example, let us return to my

[40] After all, Boyle acknowledges that we do not always consciously deliberate when forming attitudes (Boyle 2011a: 236). On this, see also O'Brien (2007: 90-2).

belief that it will rain. I believe that it will rain for the reason that there are grey clouds. Yet when answering the question 'is it true that *it will rain*?', I may well simply confirm that it will rain, and reply to the questioner 'yes, I believe that *it will rain*.'

So, it cannot be that subjects use BTM to answer the question 'why?'. This is because subjects generally *do* cite a motivating reason (even if they do not have it) or occasionally admit that they lack one. In the above example, my reply does not seem to answer the questioner at all.

iii. The process of BTM is somewhat convoluted. At any rate, BTM is more complicated than the standard TM for belief. E.g., to learn why I believe that *it will rain*, I consider whether it will rain, and hope that an awareness of my motivating reason falls out of this. Indeed, the subject reconsiders whether *o* is *Fa* even though she has not been asked whether *o* is *Fa* or whether she holds the relevant attitude. She has just been asked *why* she has the particular attitude. But this seems at odds with general trends in cognition. Subjects often tend to reason in a way that is as straightforward as possible (sometimes at the expense of accuracy!) (e.g., see Kahneman (2012) for a thorough overview). Subjects frequently use shortcuts and 'substitute' complicated questions for easy ones. It seems implausible and *ad hoc* then to insist that individuals would perform unnecessary steps when learning of their reasons.

As a result, there is further cause to doubt that we use BTM. I hope then to have shown that BTM proposes an odd picture of how we learn our reasons. In the following section I introduce a better alternative.

2.5.2 My account

The transparency method can be extended to motivating reasons in a way other than BTM. I will put my cards on the table in §2.5.3.1 and set out my account briefly, and then provide some more details. According to my overall agentialist picture, we acquire self-knowledge using what I term the *reasons transparency method*. As such, rational agency grounds the method, warrant, first-person authority, and obligations for self-knowledge of why we have our attitudes. Having set this all out, I will then in (§2.5.3.2) clarify my accounts' advantages, qua an agentialist account of self-knowledge of motivating reasons.

2.5.2.1 My agentialist account introduced

Let me start with examples:

> RAIN: I believe that *it will rain* and you ask me 'why?'. I consider what justifies believing that *it will rain*, i.e. what the normative reasons are in favour of the belief. I conclude that a

> normative reason is that there are grey clouds. I can then tell you that my motivating reason is that there are grey clouds.
>
> SEMINAR: I intend to go to the seminar today and you ask me why. I consider what justifies going to it. I conclude that a normative reason is the fact that the seminar will be interesting. I can then tell you that my motivating reason is that the seminar will be interesting.

The above cases illustrate what I call the *reasons transparency method* (RTM). I take RTM to be a way of learning of what are now in fact our motivating reasons for an attitude; it does not tell us about the reason for which we originally formed it. According to this picture, we do not learn of our (current) reason for an attitude by deliberating about whether the object of the attitude is *Fa*, e.g., whether a proposition is true or an action is to be done. And nor do we reconsider the matter at hand in any other way. Rather, we answer the question of why we have an attitude by considering what the normative reasons are for having that attitude.[41] For example, we treat the question of 'why do I believe that *q*?' as transparent to the question 'what are the normative reasons for believing *q*?'. We then conclude that *p* is a normative reason for having the attitude at hand and can reply that *p* is our (motivating) reason for it. I should clarify that answering the 'world-directed' question in RTM does not just involve considering the outside world. This is especially the case for attitudes other than belief. E.g., when learning why I intend to go to the seminar I may well take into account the fact that I find seminars interesting even though others may not. That I find seminars interesting is not a fact about the world external to me. But importantly, when using RTM, subjects only turn their attention inward in considering what are good reasons for them. It is not that subjects use evidence such as their behaviour to conclude that they have a given motivating reason.

To help bring home the difference between my account and the Boylean TM, we can note Leite (2004). Recall that I earlier mentioned Leite runs together something resembling RTM with another option. Leite writes that:

> Someone challenges you: ''On what do you base that belief? Why do you think it is true?'' To answer this question you do not consider facts about yourself or your psychology, such as how you came to hold the belief, but instead what there is to be said in favor of the belief—*whether and why you should hold it*. So in many cases, deliberating about whether a consideration represents one of your reasons is a matter of evaluating possible reasons for

[41] I think this will amount to considering the normative reasons for taking *o* to be *Fa*, but nothing rests on this.

> holding the belief. It is a matter of looking outward, as it were, *considering or reconsidering* the issue at hand and taking a stand on particular grounds (2004: 226) (my italics).

This fails to fully distinguish between BTM and RTM which, I contest, are importantly distinct. The BTM account takes it that we learn of our reason by considering whether we should hold the attitude. But I have been arguing that subjects don't employ BTM, but instead RTM. We instead consider what normative reasons favour holding the attitude, i.e. why one should hold it. We don't reconsider the issue, and whether something is to be done, or true.

Let me now provide more details concerning RTM and what I present as the agentialist picture on the table. I will outline the method's reliability and warrant, and the place of first-person authority and obligations for self-knowledge under this picture.

I can provide a plausible story about how RTM is reliable enough to result in knowledge. This is important. For a belief to constitute knowledge, it must be formed via a reliable mechanism, as well as being true and warranted in some way. The answer to the question 'why have attitude *A*?' must correspond reasonably frequently with our motivating reason. This is not to say that subjects won't sometimes make mistakes, but these won't be sufficiently frequent to render RTM unreliable. And I can explain why it is that that our self-ascriptions formed using RTM would be sufficiently reliable. It seems plausible that, as an empirical matter, our answer to the question 'why have that attitude?' reflects which reasons come to mind most strongly and easily, i.e. which are most *vivid* and *available*. Our conclusions are often influenced by such factors (e.g., Mele 2000). And in these cases, our motivating reasons will tend to be the most vivid and available facts.[42] This is because of the following three reasons:

- We may well have formed the attitude by conscious deliberation. We would have, then, consciously reasoned with the consideration constituting our motivating reason, and entertained various associated mental images. It will accordingly become salient and one that comes to mind when considering normative reasons.
- There might be lots of considerations we would take to be normative reasons, but only a few we take to be weighty. Often our attitudes will be based on what we take to be weightier reasons. For example, I want to go to the seminar because it will be interesting, not because there is a *very* slim chance that there will be cake. And what we take to be the best reasons will come to mind most strongly when answering the question 'why?'.

[42] While I talk of motivating reasons in as being facts for convenience, I can be neutral regarding their ontology.

This is because we are also engaged in the activity of justifying our attitude to others/persuading them.

- Being one's motivating reason will further serve to increase the salience of a consideration. Playing an active role in one's cognition will presumably make the consideration easier to access and more 'present' to the subject.

It is an empirical claim that our motivating reasons will tend to be most salient and available. At any rate, though, the cognitive biases used in this explanation – salience and availability – are robust. As such, given that we do at least sometimes have knowledge of our motivating reasons, my RTM account renders this explicable.

What about the warrant for self-knowledge acquired using RTM? I endorse the following (although not dogmatically). It draws on O'Brien's account of self-knowledge of attitudes. Recall that for O'Brien, I am entitled to self-ascribe judgements because I have an *agent's awareness* of my judgement in forming it. In deliberating, I take there to be various possibilities – judging that it is raining or judging that it is not – and I indeed take them to be possibilities for *me*.

We might apply O'Brien's picture to RTM in the following way. In concluding that *p* is a normative reason for believing that *q*, I take *p* to be a normative reason. In doing so, I exercise my rational agency, and I therefore have first-personal agent's awareness that the consideration is my (motivating) reason. This is because in considering possibilities regarding what to take as a normative reason, I consider possibilities for me regarding what to take as a normative reason. And further, in doing this, I am considering possible *motivating reasons* for me. So, in concluding that *p* is a normative reason, I have an implicit awareness that *p* is my motivating reason, and this awareness warrants my self-ascription.

Why am I aware of considering possible *motivating reasons* for me? One suggestion is as follows. This paragraph states it baldly; chapter seven further discusses the picture it stems from. Having a motivating reason that *p* for my belief that *q* requires being prepared to take *p* as a normative reason for believing that *q*. (There will also be a further element, such as that the belief that *p* sustains my belief that *q*.) But further, our understanding of the nature of motivating reasons involves a recognition that the subject must be prepared to take the consideration as a normative reason. So, in considering whether something is a normative reason, I am aware that this partly constitutively determines what my motivating reasons are and as such, that I am considering possible motivating reasons for me. It is not just that I happen to be considering motivating reasons for me without realising it. I am therefore warranted in ascribing the consideration as my motivating reason.

So far, this section has proposed a method and warrant for self-knowledge grounded in rational agency. As part of this agentialist picture, agency also grounds first-person authority regarding one's reasons. First, I should briefly say why subjects have first-person authority regarding their reasons; after all, some may doubt that they possess it. We should think subjects have such authority because of the oddity of cases where observers fail to accord someone the requisite deference. Consider the following:

> SEMINAR 1
>
> Suki: Why do you want to go to the seminar?
>
> Felix: Because it will be interesting.
>
> Suki: No, you want to go for the reason that it will help your general philosophical education.

> SEMINAR 2
>
> Suki: Why do you want to go to the seminar?
>
> Felix: No reason.
>
> Suki: No, you want to go for the reason that it will be interesting.

Suki's responses seem most peculiar in both these cases. Standardly we would accept what Felix says without question. We might criticise him in SEMINAR 2 for lacking a motivating reason, but we would not doubt that he lacks one. My agentialist picture says the following. We normally wouldn't contradict Felix because he bears responsibility for his attitudes having formed them. *He* is therefore 'accountable' when asked the question 'why?' and can control his attitudes in basing them on reasons. Felix therefore occupies a distinctive position of authority regarding his motivating reasons. This, though, is very brief, and the thesis will end with a fuller picture.

Finally, recall another key aspect of agentialism: we bear obligations for self-knowledge. In the case of reasons, my agentialist picture says that we bear the:

> *Knowledgeable reason explanation* (*KRE*) obligation: The obligation to knowledgeably self-ascribe motivating reasons when explaining one's own attitude/action.[43]

[43] We find the KRE obligation, and views in its vicinity, in Anscombe (2000), Boyle (2011a, §3; 2011b), and Moran (2001, esp. p. 124-9).

That is, we ought to use RTM when explaining our attitudes, and RTM ought to result in knowledge. The KRE obligation specifically pertains to reason-sensitive attitudes. Some attitudes are not reason-sensitive, such as cravings. There is nothing wrong with our craving *qua* craving if it proves resistant to what we take to be normative reasons. As a result, it may well be acceptable to say 'no reason' when asked why I desire a large pizza, insofar as that desire is in fact a craving.[44]

To give an example concerning the KRE obligation, take my belief that *it will rain*. I ought to explain this belief in terms of my motivating reasons, and not purely causal explanatory ones. So, I ought to explain my belief by reference to, say, the fact that there are grey clouds. I ought not explain it in terms of how I really want it to rain for the sake of my garden, even if this explanation is valid. Further, this self-ascription – that *I believe that it will rain for the reason that there are grey clouds* – ought to constitute knowledge. That is, it ought to be true and not just a lucky guess. We can say more about this obligation's structure and grounds, but this must wait until the conclusion – discussions from chapter seven prove vital.

To make clear how the agentialist picture is plausible, I will now quickly motivate thinking that subjects bear the KRE obligation: our interpersonal interactions indicate that this is the case.[45] For example, you would *expect* me to knowledgeably offer my motivating reason for my belief that *it will rain* when asked why I have the belief. You would see me as open to criticism if I replied to the question 'why?' with one of many alternatives. These include: 'I don't know'; 'no reason'; 'the grey clouds are a good reason, although that's not *my* reason'; 'I'm generally rational'; 'the perceptual mechanism detected patterns in the sky and processed them so as to result in a state of belief'.[46] Or, to consider a different attitude, say that I tried two yogurts and preferred the branded one to the supermarket offering. Again, you would think that something was wrong with my preference if I answered the question 'why?' by saying something like 'the advertising made it look like the sort of thing sophisticated people eat, and I want to be sophisticated.' Even if this is indeed true, you would still expect me to talk about the (supposed) rich and sophisticated flavours, and so on.[47]

---

[44] On judgement-sensitive desires, see Scanlon (1998).

[45] In this argument, I am following Boyle's thoughts in (2011a, p. 236; 2011b, p. 10; 2009, p. 4-5). In defence of this general sort of claim, see also Moran (2001) and, regarding action, Anscombe (2000).

[46] This latter response might be appropriate in certain situations (e.g., scientific discussions). Still, automatically offering it in a normal context seems to express a peculiar relationship to your belief. Jones (2002) is relevant here.

[47] It is worth noting that the motivating reasons we expect people to self-ascribe can be very minimal. In the case of perceptual belief, say, it might be enough simply to give replies shaped by Pryor's dogmatism or Huemer's phenomenal conservatism – one might self-ascribe the motivating reason that it 'seems to [me] as if *p* is the case' Pryor (2000: 519) or 'seems to [me] that *p*' Huemer (2007: 30). My point is simply that, defeaters non-withstanding, we expect people to knowledgeably self-ascribe at least *some* motivating reasons.

Here one might deny that we bear the KRE obligation by citing various counter examples. There are occasions where 'no reason' or 'I don't know' responses seem acceptable even concerning attitudes that are normally reasons-responsive. Perhaps 'I don't know' forms an appropriate response when enough time has passed that one might assume the subject has forgotten their original reason (e.g., 'why do you believe the Battle of Hastings was in 1066?' 'I don't know'). Perhaps also it is acceptable when the subject is asked why they intend to perform one trivial movement over another (e.g., 'why do you intend to stir your tea counter-clockwise?' 'No reason').

Yet we should think that the KRE obligation is simply defeated in such examples since 'no reason' and 'I don't know' is unacceptable in enough instances. The obligation will be defeated, for instance, when the subject can assume that the listener already knows their motivating reason or it can be easily worked out. In that case, the question 'why?' asks for certain details about a motivating reason that the questioner already knows about. In this conversational context, 'I don't know' or 'no reason' does not *actually* mean that the subject does not know the reason for which they hold the attitude or that they lack a reason. Let's return to the examples. In the case of attitudes formed a long time ago, the listener may already gather that such attitudes will often be based on the memory that the subject had evidence, which in itself is still a motivating reason. In that case, the question 'why?' asks for more information about the motivating reason, e.g., what the initial evidence itself was. When I tell you that 'I don't know why I believe that the Battle of Hastings was in 1066,' in this context I am only telling you that I don't remember the precise evidence that first lead me to form the belief; you are not interested in whether or not the belief is based on the memory of evidence, say.[48] My utterance in this context doesn't imply that I don't know at all what my reason is – you and I both know that it is based on the memory of evidence. And we can say something similar regarding trivial action cases, such as when I answer the question 'why do you intend to stir you tea clockwise?' with 'no reason.' In this case, there's an implicit comparison. It's not that you ask me why I have that intention *simpliciter*, but why I intend to stir the tea clockwise as opposed to counter clockwise. There, 'no reason' in this context seems to function as saying 'no particular reason, I just had to pick one.' Again, the listener will recognise this, and that the need to pick an option itself constitutes the subject's motivating reason. As such then, this thesis takes it that we bear the defeasible KRE obligation. I discuss the obligation's grounds in the conclusion of the thesis.

[48] Davidson also usefully observes that sometimes

> it is easy to answer the question 'Why did you do it?' with 'For no reason', meaning not that there is no reason but that there is no further reason, no reason that cannot be inferred from the fact that the action was done intentionally; no reason, in other words, besides wanting to do it (Davidson 1963: 688).

2.5.2.2 Why is this the best agentialist account on the table?

I take the RTM account to be the best way of extending agentialist insights to self-knowledge of motivating reasons. For starters, it satisfies the requirements I opened the section with: accounting for the way in which we self-ascribe motivating reasons in response to the open-ended question 'why?'. We look outwards and reach a conclusion about the normative reasons in favour of the attitude, but we are then warranted in transitioning to a conclusion about our *motivating reason* itself. And, in answering the open-question 'why hold that attitude?' we are not restricted to just learning whether or not we have a specified motivating reason.

Further, RTM improves on BTM in providing an intuitive account of the way in which we answer the question 'why?'. RTM captures how we don't reconsider the issue per-se, or whether *o* is *Fa*. Rather, the world-directed question specifically relates to our motivating reasons – we learn what our motivating reason is by considering the normative reasons. Let me return to the specific concerns I raised with BTM.

i. When asked why one has an attitude, subjects do not revise the attitude itself. Yet individuals would sometimes do this if they were reconsidering whether *o* is *Fa* since they may well conclude that *o* is not *Fa* after all.

RESPONSE: This stability in subjects' attitudes is exactly what we would expect if subjects answered the question by considering the question 'why hold that attitude?'. Answering it only involves marshalling reasons for the attitude and not against. Whether to hold the attitude itself is not one's focus.

ii. Subjects generally respond to the question 'why?' with a motivating reason (or occasionally the acknowledgement that they lack a reason). But one can consider whether *o* is *F* without marshalling reasons at all, and by simply concluding 'yes, *o* is *Fa*.' One need not conclude that '*p*, so *o* is *Fa*.'

RESPONSE: But according to BTM, the world directed question specifically concerns the justification for the attitude. Answering the question 'why hold that attitude?' appropriately will always involve forming a conclusion about the normative reasons (or occasionally concluding that there is no justification for the attitude). Subjects can then transition from this answer to an answer to the question 'why do I hold that attitude?'

iii. Employing BTM is somewhat convoluted and unreliable, which conflicts with our general tendency to reason as simply as possible.

RESPONSE: RTM is a simpler process than BTM. Subjects do not have to reconsider the issue at hand, and they do not run the risk of not actually forming an explanation of the attitude. Rather, BTM operates like a 'substitution heuristic'.[49] Individuals can reliably answer one question by substituting in another, without having to engage in unnecessary deliberative work.

***

Now I have set out the best agentialist picture for self-knowledge of motivating reasons, we can return to our table.

	The Orthodoxy about self-knowledge of why we have our attitudes and perform actions	**Agentialism (specifically RTM)**
Extra-reliable	No	No (more on this later)
Self-intimation	No	Yes (more on this later)
Distinctive method and warrant	No	Yes
First-person authority	No	Yes
Role of rational agency	No (not mentioned)	Yes
Extra features	No	We bear the KRE obligation

2.6 Conclusion

This chapter built on accounts of self-knowledge of belief to present what seem to be the two main options concerning self-knowledge of why we have our attitudes and perform actions. The Orthodox position takes self- and other-knowledge about why we have our attitudes to fundamentally resemble each other. Both are acquired using computation and/or inference. I also developed an agentialist alternative according to which our rational agency grounds distinctive self-knowledge of our motivating reasons. We learn that we have a motivating reason using RTM.

[49] See Kahneman on the substitution heuristic (2012: 97-105). I therefore reject Cassam's criticism of TM accounts in (2014: 7, 104).

I reject the Orthodoxy and endorse a qualified version of agentialism. In the following, I present the strongest argument for the Orthodoxy (chapter four), and then a range of problems the Orthodoxy faces even in light of this strong case (chapter five). In the Orthodoxy's place I advocate what I call the *two explanations* account. This captures the advantages of both computationalism and agentialism while construing self-knowledge as distinctive (chapters six and seven). But before doing this, I should say why we should take the thesis questions seriously at all, and the options on the table as genuine contenders. Many hold the Orthodoxy to be an obviously foregone conclusion – why bother spilling a thesis' worth of ink on the topic? The next chapter clears some ground. We can reject most of the arguments for the Orthodoxy fairly quickly, and chapter three does so.

Chapter 3 Why These Options are Live Options

Where do we stand in the thesis? The introduction presented my thesis questions:

> ***Thesis questions:*** *Is self-knowledge of why we have our attitudes and actions a distinctive species of knowledge? In what ways is it/is it not?*

Chapter two then introduced what seem to be the best options for answering the questions. According to the Orthodoxy, self-knowledge of why we hold our attitudes fundamentally resembles other-knowledge and requires inference and/or computation. I also developed a nascent agentialist alternative whereby we have distinctive access to our motivating reasons using the reasons transparency method (RTM). According to RTM, we learn why we have our attitude by considering the world-directed question 'why have that attitude?'. So, while agentialism accepts that we lack distinctive access to purely causal explanations, such as our biases, the position as I understand it nevertheless maintains that we have distinctive access to the reasons for which we hold our attitudes. I will go on to conclude that we indeed have distinctive self-knowledge of why we have our attitudes. To anticipate my view, I endorse an agentialist picture with an important concession to the Orthodoxy. I argue for my account by critically assessing the Orthodoxy in some depth. But before doing so, I should at least offer something to the Orthodoxy's proponents. After all, many think the issue is already cut and dried – why should we bother thinking seriously about the thesis questions more than I have already? Can we not just dismiss the agentialist alternative and accept the Orthodoxy? No, for we can counter most of the Orthodoxy's arguments fairly quickly. This chapter presents and refutes three (in §3.1, §3.2, and §3.3). The following chapter presents the strongest argument for the Orthodoxy, which requires more consideration.

3.1 Reason explanations are trivial

3.1.1 Argument

If we have distinctive self-knowledge of why we hold our attitudes, it will be in virtue of distinctive self-knowledge of reason explanations in particular. Cassam argues for the Orthodoxy by denying that reason explanations bear enough explanatory significance to constitute knowing why we have an attitude (Cassam 2014: 198-204). Cassam makes these claims regarding Boyle's remarks on the self-knowledge of motivating reasons. Recall that Boyle writes that:

> [I]f I reason "P, so Q" this must normally put me in a position, not merely to know *that* I believe that Q, but to know something about *why* I believe Q, namely, because I believe that P and that P shows that Q … successful deliberation normally gives us knowledge of what we believe and why we believe it (Boyle 2011: 8, quoted in Cassam 2014: 198).

I'll quote part of Cassam's reply before offering a charitable interpretation. Cassam thinks his criticism applies to most cases. These include that of 'Oliver' – a highly gullible conspiracy theorist who believes that *9/11 was an inside job* – and most everyday subjects whose cognitions are affected by a range of unconscious heuristics. Sceptics about our capacity for self-knowledge like Cassam:

> [W]ill insist that once you grasp what *they* mean by 'knowing why you believe that Q', it will be apparent that even if you have reasoned your way from P to Q you might *still* not be in a position to know why, in the relevant sense, you believe that Q. When it comes to knowing why your attitudes are as they are, there are different levels of explanation, some more superficial than others. In some cases only reflection on reasoning that is external from your reasoning from P to Q can tell you why, in the deepest sense, you believe that Q (*Ibid.* 199).
>
> For example:
>
> The explanation in terms of [Oliver's] intellectual character [i.e., citing Oliver's gullibility as opposed to his reasons] gives us an insight into the person that Oliver is, whereas merely talking about his inferential transitions in isolation doesn't do that; it doesn't explain why a particular claim or transition which in reality has little going for it is appealing to Oliver (*Ibid.* 203).

The most charitable argument to be had in these thoughts is as follows. Let's call it the Explanatory Importance Argument:

> PREMISE ONE. A subject only knows why *x* is the case in knowing an explanation of *x* if the explanation is significant (i.e. not trivial).
>
> PREMISE TWO. Reason explanations of our attitudes are generally insignificant.
>
> CONCLUSION. Subjects generally do not know why they have an attitude in knowing a reason explanation.

The thought is that reason explanations are too trivial to constitute knowledge at all. For example, say I believe that *the Sun shines because it is hot*. This belief is in some sense true. Yet the fact –

that *the Sun shines because it is hot* – seems too uninformative to count as an explanation. And, as a result, I don't seem to know why the sun shines at all. It is not that I just don't know everything about why the Sun shines, or that I don't know why it shines as well as I could.

As presented, the Explanatory Importance Argument shows that subjects rarely know why they have their attitudes in the way agentialists suggest. But Cassam thinks we can go further to say that subjects *never* can (*Ibid*. 204). That is, it's not the case that on the rare occasion in which a subject's reason explanation constitutes knowledge of why they have an attitude, they have acquired that knowledge in a distinctive way. Cassam seems to think that to know why something is the case, subjects don't just have to possess the best/full explanation of it. They also need to know that the explanation is the best/full one, which takes the form of a premise. The resulting self-ascription is thus rendered inferential.

3.1.2 Reply

I contest PREMISE TWO from the Explanatory Importance Argument (that reason explanations of our attitudes are generally insignificant). As a result, I maintain that one can know why one has an attitude in knowing reason explanations. Reason explanations do seem to bear explanatory significance and indeed count as explanations. I will say two things.

i. It is highly implausible to deny that reason explanations constitute explanations. This is because we treat reason explanations as explanations in many contexts. This is the case whether the subject herself provides the explanation or an observer does. For example, it seems at least somewhat informative to say that Julia crossed the road for the reason that a cat was on the other side. This is the case even if it is also informative to say that Julia has a cat-seeking disposition, and that the cat was especially noticeable to her since it was mewing loudly. And further, we take reason explanations to be informative in many contexts. We especially want to learn of someone's motivating reasons in cases such as the following: when determining whether someone's belief is justified; when determining whether someone is acting morally; considering whether a belief is rational by the subject's eyes (this can be especially important in relation to mental illness). If reason explanations were not explanations at all, then we would be doing something epistemically wrong in all these contexts, but intuitively we aren't.

ii. We can note one specific way in which reason explanations bear explanatory significance. Recall Cassam's example for thinking that reason explanations are relatively insignificant: Oliver, the habitual conspiracy theorist, who believes that 9/11 was a hoax. When explaining Oliver's belief, we want to know 'why a particular claim or transition which in reality has little going for it is appealing to Oliver' (*Ibid.* 203). But we also need to explain the way in which the claim seems

appealing to Oliver. Just saying that he is gullible does not do this. The proposition [9/11 is an inside job] seems sensible to believe, by Oliver's lights, because he has motivating reasons for it. I say more about motivating reasons in chapter 6. But roughly, we might think that believing a proposition for some consideration makes holding the belief intelligible and appealing to the subject. Reason explanations, then, play a non-trivial role, and can be properly seen as explanations. I therefore reject Cassam's conclusion and maintain that we can know why we have an attitude in knowing the reason explanation for it.

3.2 Knowledge of causes

3.2.1 Argument

To know what explains one's attitude is to know what caused it. But causal features are such that subjects could only learn of them using a third-personal method (i.e., the same method subjects use when explaining another person's attitude). Let's call this the Knowledge of Causation Argument. We see this argument in Gertler (2011: 73-5), and can express it in the following way:

> PREMISE ONE. What explains an attitude/action is what caused it.
>
> PREMISE TWO. Learning of causes must involve inference.
>
> PREMISE THREE. Distinctive self-knowledge doesn't involve inference.
>
> CONCLUSION. Subjects lack distinctive self-knowledge of what explains their attitudes/actions.[1]

While Gertler mostly addresses one putative first-personal method in particular (inner observation), she does think the Knowledge of Causation Argument applies more generally. If it holds, the Argument rules out any sort of distinctive access to why we have our attitudes/perform actions.

[1] Gertler uses the term 'theorising' but I take her just to mean inference. I shall refer to inference for the purposes of continuity in the thesis.

Let us say a bit more about the three premises. PREMISE ONE seems to be assumed, and I will accept this too.[2] In arguing for PREMISE TWO, Gertler considers the case of getting up to acquire cereal, and cites Wilson who writes that:

> My decision to get up off the couch and get something to eat, for example, feels very much like a consciously willed action, because right before standing up I had the conscious thought "A bowl of cereal with strawberries sure would taste good right now." (Wilson 2002, cited in Gertler *Ibid*. 73).

So, a conscious thought proceeds our action. But, Gertler writes, even if (which is a big if) the thought caused the action, we would still have to infer that the thought did so. This is because we cannot have non-inferential knowledge of causal relations (*Ibid*. 74). Therefore, learning why we have an attitude or perform an action will always be to infer. And, following PREMISE THREE, inference is inimical to distinctive self-knowledge. After all, inferring is to use a method that others can employ to learn of our mental life too. As such, we lack distinctive self-knowledge of why we have our attitudes.

3.2.2 Reply

We could reject any of the three premises of the Knowledge of Causes Argument, but I will challenge PREMISE TWO.[3]

Baldly, the proponent of RTM should say the following. S can non-inferentially learn what her motivating reason for an attitude is by considering the world-directed question 'why have that attitude?'. S takes the consideration *p* to be a normative reason and can thereby self-ascribe *p* as her motivating reason. S is non-inferentially warranted in self-ascribing her motivating reason that *p* because of the agent's awareness she has of the motivating reason that *p*. She has that awareness in virtue of her agent's awareness of deliberating and taking *p* to be a normative reason.

I will say two things to motivate rejecting PREMISE TWO in this way. As Cox rightly notes, it is highly debatable whether we can only learn of causal properties or relations by engaging in inference, and even more so regarding motivating reasons in particular (2018: 181).[4]

---

[2] Another option would be to characterise motivating reasons as dispositions instead of causes. Ian Evans (2013) makes this move, which I discuss in chapter seven. But Gertler would probably, I think wrongly, give a similar argument in this case.

[3] Setiya (2013) and Anscombe (2000) would deny PREMISE ONE, for instance.

[4] We can also note Davidson (1963: 699-700).

3.2.2.1 A precedent

There is relevant precedent, which indeed helps the proponent of the RTM account. I take it that we can non-inferentially learn of motivating reasons. That is, we can non-inferentially learn that we have a belief with a certain causal property – that of being a motivating reason.[5] Instructive literature in the philosophy of perception argues that we can non-inferentially learn of related phenomena: causal relations and dispositional properties. The thought is that we can *directly perceive* Ellie cutting the brownies (i.e. causing the brownies to separate) or the fragility of the ornamental cat.

We should start by noting the relevant debate. There is a question as to what it is subjects can see, that is, what they perceptually represent (or the naïve realist equivalent). Uncontentiously, for example, most subjects can see the colour red – it features in their experience and they can non-inferentially learn that something in front of them is red. And we might add various other properties to the content of perception – perhaps subjects can see *an apple* or maybe even *a Royal Gala* apple. After all, individuals seem to undergo a different experience depending on whether they see the fruit in front of them as a *Royal Gala apple* or just as an *apple*.[6]

Some have plausibly argued that the content of perception includes features such as causal relations and dispositional properties. We *see* causal relations and dispositional properties themselves in the sense that our perception represents them. It might be that the relevant features are present in experience (Siegel 2009) and/or unconscious perceptual representations (Nanay (2011) argues that dispositions can be represented in perception both consciously and unconsciously). So, say Ellie puts a knife into the brownies causing the brownies to be cut. It is not that I see the knife going into the brownies, and then two halves. Rather, I see Ellie cutting the brownies, i.e., causing the brownies to be cut (Siegel 2009). And we can also be said to see dispositional properties (Nanay 2011), such as fragility. E.g., I can see the ornamental cat's fragility. Nanay observes that this claim is more plausible once we acknowledge that 'perceiving x as F does not imply that I perceive what makes this so, that is, what makes it the case that x is F' (2011: 205). We can represent, say, the ornamental cat as being fragile without representing that it would break in certain circumstances (2011: 205-6; Seigel 2009: 538 makes a similar point). And further, we can perceptually represent an ornamental cat as being fragile without fragility being what our perceptual system itself covaries with – that would be the shape and colour of the cat (Nanay 2011: e.g. 302-3, 305). Therefore, it is plausible that our perceptual experience represents

[5] Or perhaps it is instead a dispositional property; see chapter seven.
[6] See Siegel (2012, esp. 2006) on this topic.

causal relations and dispositional properties. And this perceptual content, provided it reliably links up with facts in the world, suffices to give us perceptual knowledge of those features.[7]

So, our knowledge of causal and dispositional features is sometimes non-inferential in the relevant sense; I say 'the relevant sense' because such cases do involve *some* sort of inference. My perception of the ornament's fragility may well be underpinned by inferential processing of the sort that underpins all perception. And yet the process is not inferential in the important sense for our purposes. The subject herself cannot be said to infer – just the subpersonal mechanism – and the resulting warrant is l non-inferential. Indeed, as I discussed in chapter two, I take this process to involve *computation* rather than inference.

Direct perception, then, provides precedent for self-knowledge of motivating reasons. The proponent of RTM can say that we have agent's awareness of our motivating reason but that this awareness does not directly track the motivating reason that *p* itself. Rather, we are aware of forming the motivating reason that *p* via our agent's awareness of taking *p* to be a normative reason. And we have agent's awareness of our motivating reason without being aware of what makes the belief that *p* our motivating reason, e.g., the causal relations between the belief that *p* and the attitude in question.

3.2.2.2 Some causal relations are special

Second, PREMISE 2 is very strong – it states that subjects can *only* ever learn of causal properties by inferring. But the proponent of RTM only needs to say that subjects can learn of *some* causal properties without inferring. And this seems plausible. Motivating reasons are a special sort of cause (Cox 2018:181). Contra Gertler, believing for a reason is not (just) a matter of having a thought that happens to cause the belief. As I argue in chapter seven, for the consideration that *p* to be the reason for which a subject believes that *q*, they must be prepared to take *p* to be a normative reason for believing that *q*. After all, when a subject believes something for the reason that *p*, *p* makes holding the belief intelligible to them. Their belief that *p* does not just cause their belief that *q.* In that sense, motivating reasons importantly differ from, say, a bias or character trait that might also explain the individual's attitude. So, while subjects use only inference when forming non-reason explanations, the proponent of RTM can say that reason explanations are special, and that subjects at least sometimes non-inferentially learn what their reasons are.

[7] Another relevant option is to appeal to displaced perception/secondary seeing. This is the phenomenon introduced by Dretske whereby one can be said to see one thing by seeing another (e.g. (Dretske 1969: 153-162). See also Millar's (2010) discussion of indicators, although the knowledge involved is not immediate.

3.3 Confabulation

3.3.1 Argument

One might think that empirical data concerning self-ignorance and error shows that we lack distinctive self-knowledge of why we have our attitudes. Chapter four discusses this data in depth but let me briefly introduce the relevant issue. The following study – Nisbett and Wilson's (1977) stockings experiment – exemplifies a reasonably commonplace phenomenon. Nisbett and Wilson arranged four pairs of identical stockings on a table and asked individuals which one they preferred. The majority picked those placed towards the right and so were influenced by what is termed the 'position effect.' At any rate, the subjects in general would not have formed their preference on the basis of a (perceived) reason – after all, the stockings were all the same. Yet, when asked why they prefer the pair they chose, the participants did not say that it was because of the position, or for no reason at all. Instead, they offered reasons like its 'knit, sheerness, and weave.'[8] In such cases, the subject fails to know why it is they have their attitude – in this instance, they were ignorant that their preference resulted from the position effect. But further, in providing the mistaken self-ascription, the subjects also *confabulate*.[9]

Self-ignorance and confabulation feature in two arguments for denying that we have distinctive access to why we have our attitudes. Chapter four discusses the best argument in some depth, but let me quickly note the other. Confabulation cases show that the causes of our attitudes do not self-intimate, and that our explanations are not especially reliable. And, so the thought goes, this means that our self-knowledge of why we have our attitudes cannot be distinctive (for something like this argument see e.g., Gertler 2011: 72-3).

This argument has a lot going for it. There plainly can be causes that we do not know about – the position effect caused subjects to prefer the stockings on the right unbeknownst to them. Further,

[8] See also Wilson (2002, p. 103) and Wilson and Nisbett (1978, p. 123-4).

[9] One might deny that the subjects were mistaken about their reason – perhaps the stockings actually seemed sheerer to them, and they preferred the pair on that basis. Sandis (2015), for example, responds to many such cases in this way. Yet here I can say several things. First, this interpretation seems less charitable. It appears odd to think that subjects would take identical items to differ – perhaps they eventually come to prefer the stockings on the basis of their (perceived) sheerness, but it seems less plausible that they would do so from the start. Second, even if the (perceived) sheerness was their motivating reason, the subject was still mistaken in attaching so much explanatory importance to it. Nisbett later allows that the subjects in the stockings experiment might indeed have the self-ascribed motivating reason, but that nevertheless 'by normal standards of discourse, [their] causal analysis is inadequate or incomplete' (Nisbett and Ross 1980: 217-9). And third, we could bite the bullet in this case, but maintain that subjects still provide false self-ascriptions in others. At any rate, Sandis allows that other experiments show confabulation – ones which are especially important in the literature (e.g., the choice blindness cases I discuss in chapter five, and Haidt (2001)).

subjects weren't even in a position to know that the stockings' position caused their preference – few people know about the position effect, and even then, struggle to recognise it in their own cognition. The stockings experiment (and other similar cases) also show that our reason explanations are not especially reliable. In the stockings experiment, subjects were mistaken in believing that they preferred the stockings on the basis of a reason.

3.3.2 Reply

Here I will say two things.

i. Importantly, if we have distinctive self-knowledge of why we have our attitudes, it would only be in virtue of a distinctive access to one type of explanation – reason explanations. No one thinks we have privileged access to purely causal explanations. Our ignorance of biasing factors only show that biases do not self-intimate. It does not show that motivating reasons fail to self-intimate (I argue in chapter seven that motivating reasons do self-intimate).[10]

ii. Self-knowledge does not have to be especially reliable to still be distinctive. I accept that self-ascriptions of motivating reasons are no more reliable than other-ascriptions. Still, the proponent of RTM can maintain that self-knowledge is distinctive in other ways, e.g., the method and warrant used. The agentialist only need the self-ascriptions to be reliable enough to constitute knowledge. And surely the self-ascriptions would be; at any rate, no denies on the grounds of confabulation data that we can at least *sometimes* know why we have our attitudes (more on this in chapter six).

3.4 Conclusion

I hope to have persuaded the reader that the Orthodoxy isn't obviously right, and that it is worth assessing further. I will now get to the meat of the issue. The following chapter presents what I take to be the strongest case in favour of the Orthodoxy. The chapter provides (*pro tanto*) reason to accept the position and reject the agentialist alternative.

[10] Relevantly, see Sandis (2015).

Chapter 4 The Orthodoxy's Appeal and Inference to the Best Explanation

This chapter introduces the best case for the Orthodoxy: inference to the best explanation of self-ignorance and confabulation.[1] An 'inference to the best explanation' argument (IBE) argues for a claim on the grounds that it provides the best explanation of a given phenomenon. Proponents of the Orthodoxy contend that we best explain self-ignorance and confabulation by accepting the following claim: subjects' explanations of their own attitudes and actions are underpinned by the same method that underpins subjects' explanations of other people's attitudes and actions. Recall that the Orthodoxy has two forms, which operate at the subpersonal and personal levels respectively. According to computationalism, the same subpersonal mechanism forms self- and other-ascriptions. The mechanism transitions between contents that indicate one bears a particular reason. According to inferentialism, the subject herself infers both self- and other-ascriptions. This inference may well be grounded by subpersonal processing, but still constitutes a different sort of method. So, the IBE provides reason on explanatory grounds to accept either computationalism or inferentialism about self-knowledge of motivating reasons.

This chapter presents the IBE as provided by Carruthers. I should note that Carruthers' IBE concerns self-knowledge of *what* our attitudes are as opposed to *why* we have them, unlike, say, Nisbett and Wilson (1977). But Carruthers's points can be extended to this issue. And indeed, the fact that Carruthers doesn't do so himself seems to be because he assumes that everyone already accepts that we lack distinctive access to why we have our attitudes (2013: 329-30). I focus on Carruthers for two reasons. First, Carruthers presents the IBE in a particularly strong way, and raises what I term the TWO METHODS PROBLEM for alternative accounts of self-knowledge. Second, Carruthers uses the IBE to make an extreme claim (as does Cassam 2014). Carruthers claims that we do not just lack distinctive access to *why* we have our attitudes, but also *what* our attitudes are as well. Many of this thesis' arguments also apply to Carruthers' bold claim. I take there to be reason for thinking that self-knowledge of attitudes is computational/inferential, but ultimately, we should still accept that subjects have distinctive access. I will continue to focus on self-knowledge of *why* we have our attitudes, but revisit self-knowledge of *what* they are in the conclusion.

[1] E.g. Nisbett and Wilson (1977), Wilson and Dunn (2004), and Wilson (2002). Cox (2018) helpfully systematises the IBE implicit in such works. And we can usefully extend the arguments of Cassam (2014) and Carruthers (2013) in this way.

This chapter proceeds as follows: §4.1 introduces the IBE's explananda and §4.2 outlines a useful preliminary – what makes a good explanation. §4.3 then outlines the IBE itself. §4.4 argues that this argument does indeed trouble agentialists. We will end, then, with reason to accept the Orthodoxy.

4.1 Self-ignorance and confabulation

In introducing the explananda – self-ignorance and confabulation – I will mention two examples and then pinpoint the broader class of cases. Similar instances abound in the literature, and indeed in everyday life.[2]

Chapter three outlined a key study: Nisbett and Wilson's (1977) stockings experiment. To recall, experimenters asked subjects which of four pairs of identical stockings they preferred. Subjects chose the stockings towards the right but did not cite the position when explaining their preference. Instead they gave false explanations in citing motivating reasons like the stockings' 'knit, sheerness, and weave' (*Ibid.* 249).

Haidt (2001) also provides another indicative example. Experimenters provided subjects with a scenario in which two siblings, Julie and Mark, engage in sexual intercourse because they think 'it would be interesting and fun' to do so (*Ibid.* 814). Julie and Mark use two forms of contraception. Despite becoming platonic from that point, the siblings are happy with the experience and it strengthens their relationship. The participants largely pronounce Julie and Mark's actions to be 'wrong', before:

> [T]hey then begin searching for reasons (Haidt, Bjorklund, & Murphy, 2000). They point out the dangers of inbreeding, only to remember that Julie and Mark used two forms of birth control. They argue that Julie and Mark will be hurt, perhaps emotionally, even though the story makes it clear that no harm befell them. Eventually, many people say something like, "I don't know, I can't explain it, I just know it's wrong." (Haidt 2001: 814).

[2] E.g., in the case of attitudes: Hall, Johannson, and Strandberg (2013), Johansson, Hall, Sikström, and Olsson (2005), Johansson, Hall, Sikström, Tärning, and Lind (2006) (all choice blindness studies), and Haidt (2001). See Scaife (2014, §2.4) for a discussion of choice blindness in relation to confabulation. Regarding action: see studies such as those involving split-brain patients (e.g., Gazzaniga 2000), and hypnotised subjects (as discussed, e.g., in Wegner (2002).

The subjects attempt to explain their judgement, and only after this has been unsuccessful do they admit their ignorance of why they take the scenario to be immoral.[3] The participants are initially mistaken about why they have an attitude, as with the stockings experiment.[4]

Both these cases illustrate self-ignorance and confabulation concerning why we have our attitudes. It should be clear what the relevant sort of self-ignorance amounts to: the subject's failure to know why they have an attitude. For example, in the stockings experiment, subjects are ignorant of the position effect. Further, in providing a mistaken self-ascription, subjects *confabulate*. I will be stipulative in how I understand confabulation. How best to define confabulation is contested, but here I am just interested in the mechanism underpinning (one particular subtype of) it.[5] I draw on aspects of Hirstein's (2005) definition, and hope it, and the examples discussed, suffice to illustrate what I have in mind.

Subjects confabulate in expressing an 'ill-grounded' belief.[6] A particular subtype of confabulation interests me here: the sort exemplified by the stockings experiment. Firstly, it is an instance of *provoked* confabulation. The participants form the mistaken belief, and express it to the listener, specifically once they have been asked why they have the attitude. We can contrast this with the *spontaneous* sort in which subjects confabulate of their own volition. Further, I focus on confabulation in non-clinical subjects, rather than confabulation resulting from neurophysiological disorder.[7] And finally, I will be concerned with confabulation about why subjects have their attitudes, which I take to be a substantial subtype – much of the literature surrounding non-clinical confabulation concerns the confabulation of explanations.[8] I have, then, this in mind when referring to confabulation in what follows.

---

[3] It seems reasonable to see this as a case of self-ignorance rather than as a case in which subjects cannot justify a position but nothing self-representational is at stake. Haidt later refers to how 'in a moral judgment interview, a participant is asked to decide whether an action is right or wrong and is then asked to explain why she thinks so' (2001: 822).

See also the experiment carried out in Haidt et al. (1993), with information relevant to our discussion in p. 626 of that paper, Haidt and Bjorklund (2008: 196), Haidt (2001: 817) and Haidt et al.(2000: 3).

[4] We can take it that the subjects did not actually form their belief on the basis of the reasons they later provide. It would be uncharitable to think that they did this – the reasons are obviously false in the context of what is a short extract.

[5] For overviews of how one might define confabulation, see Bortolotti and Cox (2009) and Hirstein (2009; 2005).

[6] For example, we can contrast this with definitions of confabulation that just concern mistakes in memory, e.g. (Fotopoulou (2009) and McKay and Kinsbourne (2010).

[7] For these distinctions, see Hirstein (2005).

[8] Indeed, the confabulation of motivating reasons even constitutes Scaife's (2014) definition.
I should note that a reasonable amount of the literature on confabulated explanations concerns actions rather than our attitudes, e.g. Hirstein's (2009) section on confabulated introspection. I take my explanation to extend to these as well.

In chapter three, we encountered one way in which some have used cases of self-ignorance and confabulation to argue for the Orthodoxy. Proponents of the Orthodoxy observe that our self-ascriptions are not especially reliable. I agreed that if self-knowledge of motivating reasons is distinctive, it cannot be in virtue of special reliability. But, I argued, self-knowledge could still be distinctive in virtue of a range of other features. In the following, I will consider another, better argument for the Orthodoxy from self-ignorance and confabulation. This different approach rests on how we explain such cases. It argues that subjects lack distinctive self-knowledge in lacking a distinctive method for acquiring it.

4.2 A helpful preliminary: explanatory virtues

Now I've set out the *explananda*, let me clarify what Carruthers means by explaining something *well*. After all, an inference to the best explanation involves assessing the merits of explanations. An explanation can be good or bad, and one good explanation can be better than other good explanations. This is all a matter of possessing various explanatory virtues. I'll note the virtues that Carruthers finds important, and I will follow him on these matters:[9]

i. Simplicity

This is a matter of being able to explain the data parsimoniously. E.g. Carruthers (2013: 366).

ii. Not being ad hoc

Ideally, explanations should only rely on additional claims if they are part of our prior view of the world, i.e., claims we would accept regardless of explanatory need. E.g., Carruthers (2013: 366).

iii. Explanatory scope

An explanation can be a good in virtue of being able to explain a range of data.[10] E.g., Carruthers (2013: 366). We should note that a theory should actually *explain* this data as opposed to just be 'consistent with' it (2013: 367).

[9] What explanatory virtues are and how they relate to each other raises controversy in the IBE literature. For discussions of IBE and the explanatory virtues, see, for example Thagard (1978), Mackonis (2013), Lipton (2004), and Harman (1965).

[10] That they explain a *range* of data is important. As Thagard observes, an explanation does not do much good if it just explains very similar data (1978: 82-3).

4.3 The Orthodoxy's argument: inference to the best explanation from confabulation

I'm now in a position to set out the IBE for computationalism/inferentialism about self-knowledge. Carruthers and other proponents of the Orthodoxy argue that we best explain self-ignorance and confabulation as resulting from normal failures in self-ascription construed computationally/inferentially. As we have seen, the Orthodoxy thinks that even knowledgeable self-ascriptions of why we hold our attitudes are formed by computation/inference. Such a method might easily result in a false belief instead of knowledge, as emphasised in e.g., Carruthers (2013: 324).

I understand Carruthers' argument as follows:

> CLAIM ONE. Other accounts of self-knowledge cannot explain self-ignorance and confabulation.
>
> CLAIM TWO. The Orthodoxy can explain self-ignorance and confabulation. Indeed, it can explain the data well. This is a positive claim about the Orthodoxy's explanation and holds even if another account could explain all the data.

These two claims support a third:

> CLAIM THREE. The Orthodoxy offers the best explanation of self-ignorance and error about why we have our attitudes.[11]

Provided we take IBE to be a valid argument form (as I do), CLAIM THREE provides the following conclusion:

> CONCLUSION. We should accept the Orthodoxy.

CLAIM TWO seems to do most of work in the argument, but CLAIM ONE also offers important support. In this section, I will consider both claims in turn (in §4.3.1 and §4.3.2 respectively).

---

[11] I have kept a distinction from the main text for simplicity. We can explain two interrelated aspects of a given phenomenon, such as confabulation. First, we can explain how the phenomenon is brought about (i.e., cite a mechanism by which it occurs). For example, when explaining plant growth, we can cite how the cells multiply in particular ways. And second, we can explain the patterning of cases (i.e., the explanation must give rise to predictions). For example, the requirements for cell-multiplication mean that plants grow in sunny conditions. Carruthers looks to explain both what brings about self-ignorance and error and the patterning of cases. See e.g., (Carruthers 2013: 365-6). I take it that by 'patterning,' Carruthers has in mind the conditions under which individuals will be self-ignorant or mistaken, and why they are self-ignorant or mistaken in specifically those cases. E.g., (*Ibid*. 6 and 163). On explanation and prediction, see Douglas (2009). Mackonis (2013) also appeals to this notion.

Before discussing the claims, I should note that I take the IBE to be an argument against agentialism and the reasons transparency method (RTM). This is the case even though Carruthers sees agentialism as compatible with his view (*Ibid.* 326). But Carruthers makes this move based on an uncharitable understanding of agentialism, or at least one that conflicts with my own reading. To Moran and others, he attributes the view that:

> By articulating a belief or decision (whether out loud, or to ourselves), we don't just *express* an attitude, nor do we just *assert* that we possess that attitude; rather, we *commit ourselves* to having it. It is this commitment, and the normative motivation that accompanies it, that insures that our future actions are of a sort appropriate for an attitude of that kind. Hence even if the initial statement of attitude is arrived at through the unconscious interpretive activity of the mindreading faculty – indeed, even if that statement is completely confabulated – the very act of articulating the attitude will often insure its own truth (*Ibid.* 96).

Carruthers seems to think that the agentialist picture only makes claims about the self-ascription after one has formed it. For him, agentialism does not commit to an account of how the self-ascription is initially formed or even why the ascription constitutes knowledge. But, the proponent of RTM makes a claim about how we form our self-ascriptions, and so I take the IBE to directly target it.

4.3.1 CLAIM ONE

First, why accept CLAIM ONE: the claim that other accounts of self-knowledge cannot offer a good explanation of self-ignorance and confabulation concerning why we have our attitudes?[12]

Carruthers' case rests on a particular claim: even if subjects have a distinctive method for forming self-ascriptions (I will call this the 'distinctive method'), errors must result from an alternative one. Specifically, errors result from inference/computation (2013: 39-42, 325-6). Indeed, I will argue in §4.4.1 that Carruthers is right to think that errors result from inference/computation. For Carruthers, a proponent of distinctive access will therefore take there to be two distinct ways of forming self-ascriptions – they will be a 'dual-method theorist.' A *dual method* theorist would therefore say the following: subjects sometimes employ a distinctive method (e.g. RTM) but rely on computation/inference in cases such as the stockings experiment and Haidt's study. According

[12] Later in the central chapter on confabulation, Carruthers also assesses the accounts' abilities to explain 'self-perception' and 'dissonance' data' (*Ibid.* ch.11, sections 3, 4, and 5). But for simplicity, I focus on confabulation here.

to this approach, the distinctive method and computation/inference are wholly distinct from each other.

Dual method theorists face an explanatory problem – what I term the TWO METHODS PROBLEM. Namely, to explain our ignorance and errors about why we have our attitudes, the *dual method* theorist must state when computation is used to form the ascription as opposed to the distinctive method (*Ibid*. 201, 326). *Dual method* theorists must do this in order to explain such instances of self-ignorance and error, as opposed to just accommodating them (*Ibid.* 324). Carruthers writes that *dual method* theorists cannot solve the PROBLEM in a principled manner that accounts for all relevant cases. This limits the explanatory scope of their theory and/or means they must depend upon additional *ad hoc* principles (*Ibid.* 366). Carruthers rejects various possible solutions to the PROBLEM that the *dual method* theorist might suggest. In the remainder of this subsection, I outline two of these options in §3.1.1 and §3.1.2. Indeed, I accept that these options fail.

4.3.1.1 Option one

The *dual method* theorist (e.g., the proponent of RTM) might say that we rely on confabulation/inference in confabulation cases because there is no explanation of the attitude that the distinctive method can access (*ibid.* 333-5). After all, as the introduction mentioned, distinctive access has limits to its scope – no one thinks that a distinctive method can tell us about every state and every explanation. To extend this to RTM, then, the proponent of RTM might give the following picture. An explanation of an attitude will normally be accessible if it is a motivating reason. If it is not, and instead a purely causal explanatory reason, then it will not be. No one thinks I have distinctive access to the fact that I like Bakewell tart because of childhood nostalgia. In the stockings experiment, then, the subjects cannot learn of the explanation of their preference – the position of the stockings – because it does not feature a motivating reason. The absence of motivating reasons causes subjects to rely on computation/inference instead of RTM and to confabulate.

We can draw on the worries Carruthers raises with quasi-perceptual accounts to say the following about RTM (see *Ibid.* 335).[13] The option in question – that the subject would use computation as

[13] As mentioned, Carruthers discusses this in terms of self-knowledge of what our what our attitudes are. But perhaps the most relevant case from the central chapter (Ch. 11) pertains to intentions (2013: 342-3). Here, Carruthers discusses a study in which subjects were hypnotised so as to perform an action, such as to put a book on a shelf. The subjects then explained the action with a rationalising intention, for example, 'to tidy the room.' Carruthers thinks this is problematic for certain accounts because the subjects had an intention which they ignored, such as that 'WHEN I SEE THE BOOK ON THE TABLE I SHALL PLACE IT ON THE SHELF.' This means we cannot appeal the absence of the relevant state to explain why the individual uses interpretation over the privileged method.

opposed to RTM because they lack a motivating reason – is ad hoc. This is because the subject could still use RTM to form an answer to the question 'why?': that they lack a motivating-reason. For example, take cases such as the stockings experiment and Haidt's study. Say that the subjects considered with a reasonable degree of attention the question 'why prefer this pair of stockings?' or 'why judge that the siblings behave immorally?' Presumably the subjects would answer the world-directed questions with 'no good reason.' After all, the stockings are obviously the same, and the details of the (very brief) incest case clearly preclude the sorts of answers the subjects confabulate. This would allow the subjects to know that they lack a motivating reason for the attitude. Since individuals could form *an* answer, it is not clear why they would instead use computation to form a self-ascription. Insisting that subjects would nevertheless use computation when they lack a reason is *ad hoc* and unappealing.

As such, we cannot appeal to limits in the distinctive method's scope to explain why subjects use one method over another in confabulation cases. I do, though, take it that this option succeeds in explaining self-ignorance: subjects fail to know the explanatory fact because it is a purely causal reason and therefore RTM cannot access it.

4.3.1.2 Option two

The *dual method* theorist might instead appeal to motivational factors to explain why subjects use one method over another, specifically concerning the pragmatic pressures in the situation (*Ibid.* 337-8).[14]

For Carruthers, these pragmatic pressures take the form of norms associated with interpersonal communication generally and psychology experiments in particular. One might claim that:

> [E]xperimenter questioning (especially by a person of authority) is apt to place pragmatic constraints on subjects to present themselves in a good light, to have something interesting and valuable to say, to offer explanations that go beyond what they can report, and so forth' (*Ibid*. 337).

These pressures lead the subjects to rely on computation/inference to have something suitable to report to the experimenter. Indeed, we can emphasise the force of the pragmatic pressures option to a greater extent than Carruthers does. It forms a key position in the confabulation

---

Carruthers' response, though, might not hold if we cast the discussion in terms of motivating reasons. While subjects may have an intention, they lack a motivating reason. They therefore confabulate a motivating reason because they lack one.

[14] Carruthers doesn't discuss this in motivational terms, but we should. Cassam also rejects a motivational account of self-ignorance and error in (2014: 193-5).

literature. Philosophers and psychologists often think that a desire to fulfil pressures to reply to others plays at least some role in explaining confabulation. For example, the thought is that subjects confabulate because they are motivated by 'simply the desire to avoid saying, "I don't know," especially when the provoking question touches on something people are normally expected to know' (Hirstein 2005: 17). Doing so would be 'socially rewarded' (Bortolotti and Cox 2009: 961) and would avoid 'embarrassment' (Sullivan-Bissett 2015: 555).[15]

But, as Carruthers observes, it's not clear that relevant pragmatic pressures operate in the given situations.[16] Experimenters often control for influences of this sort, making it unobvious what the relevant pressures would be:

> Wilson et al (1989) argue in some detail, for example, that demand-characteristics are unlikely to be the explanation for people's reported changes of attitude in confabulation experiments. This is because those reports are often (supposedly) made privately and anonymously, sometimes to be thrown in the trash, sometimes to be immediately aggregated by the computer. And subjects are generally given the impression that those reports are entirely incidental to the main purpose of the experiment (*Ibid.* 337).

As Wilson et al. (1989) write: 'We tell subjects in our studies that we want them to think about their reasons in order to organize their thoughts, and we explain that no one will ever read what they write '(1989: 297). There seem to be two strands to this such that little is at stake for the subjects. Studies often minimise interpersonal factors that might lead subjects to think that the questioner or general demands of communication require them to give a certain answer. Also, the experimenters trivialise the importance of the subjects' self-ascriptions – subjects would not feel pressured to provide one explanation over another when it is insignificant either way. So, it looks like pragmatic pressures don't help us explain why subjects rely on computation/inference rather than the privileged method.

***

We have, then, some reasons in favour of CLAIM ONE. To recall, CLAIM ONE states that accounts of self-knowledge other than the Orthodoxy, e.g. RTM, cannot offer a good explanation of our ignorance and error concerning why we have our attitudes. While RTM can offer a good explanation of self-ignorance, it looks like RTM cannot explain confabulation. This is because we

[15] See also McKay and Kinsbourne (2010: 291), and Fotopoulou (2009: 270-1) for discussions of this sort of view.

[16] Carruthers also argues that pragmatic pressures couldn't do the relevant work even if they were present (*Ibid.* 337). I disagree with his particular argument but will agree in §4.4.2 that pragmatic pressures couldn't do the work.

have rejected the most appealing options for why subjects would rely on computation/inference rather than the distinctive method in those cases. These options were: (1) subjects lack an explanation that the distinctive method could access; (2) subjects are motivated by pragmatic factors.

4.3.2 CLAIM TWO

Further, Carruthers argues that Orthodoxy offers the best explanation of self-ignorance and confabulation. This holds even if other accounts of self-knowledge could also explain it.

I briefly mentioned the Orthodoxy's explanation at the start but let me introduce it now in more depth. The Orthodoxy argues that self-ignorance and confabulation occurs because subjects always explain their attitudes using computation/inference. Relying on this method results in self-ignorance and error in various ways. Carruthers references the possibility of 'misleading behavioural or other sensory evidence' (2014: 365).[17] And Cassam provides some more details: problem cases can stem from the method itself and the evidence used. He writes that you can be ignorant of an explanation for your attitude if you 'lack the necessary evidence' or 'you haven't performed the necessary inference from the evidence you have' (*Ibid.* 194-5). And you might make a mistake if you 'reason poorly,' 'misinterpret the evidence,' or 'have a defective theory about the relationship between your evidence and your attitude' (*Ibid.* 195). In proposing this sort of explanation, the Orthodoxy commit to the claim that it is not just possible for our inferences to go wrong in these ways, but that they actually do. To give an example, let us return to the stockings experiment. Subjects mistakenly believe that *they prefer the stockings for the reason that they are sheerer* because of the inference/computation they standardly rely on. In this case, the reasoning process might take the following form: 'I have chosen a pair of stockings. A common reason for preferring a pair of stockings is that they are sheerer. Therefore, I chose the pair of stockings for the reason that they are sheerer' (See Nisbett and Wilson 1977: 248-9).[18]

Carruthers takes the Orthodoxy's explanation to be especially good because it is simple. As Carruthers observes, other accounts become less parsimonious in claiming that there are two methods for forming self-ascriptions (*Ibid.* e.g. 366). Sometimes simple explanations aren't always the best – they can be too simple – but are preferable all things being equal. In the case of self-knowledge, Carruthers argues that we lack independent reason to accept the more complicated

[17] See also (*Ibid*. 6).
[18] See also Wilson (2002). Carruthers explains the pantyhose experiment a bit differently (*Ibid.* 335-7). He takes the relevant *explanandum* to be that the subject mistakenly self-ascribes the judgement that the pair is 'the softest.'

picture: 'in order to warrant the extra complexity, it needs to be shown that [computationalism] on its own is inadequate, or else some positive evidence of an additional method should be provided' (*Ibid.* 366). This, Carruthers argues, has not been shown.

This second claim of the IBE, supported by the first, entails the third claim: that the Orthodoxy about self-knowledge provides the best explanation of self-ignorance and error about why we have our attitudes. This, then, provides reason to accept the Orthodoxy.

***

Let me summarise this section. The Orthodoxy's proponents think the IBE provides reason to accept their account of self-knowledge. That is, we have reason to accept that either computation or inference underpins all self-ascription. For Carruthers, the Orthodoxy provides a good explanation of self-ignorance and confabulation, but other accounts cannot do so at all. Carruthers argues that other accounts cannot overcome the TWO METHODS PROBLEM in a non-*ad hoc* way: assuming that mistaken self-ascriptions result from computation/inference, when is it that we employ computation/inference as opposed to the distinctive method?

4.4 Why the IBE gives us reason to accept computationalism/inferentialism

I agree that the IBE gives us reason to accept either computationalism or inferentialism about self-knowledge of motivating reasons. I have already said a few things to bolster the IBE while setting it out. I clarified that the IBE applies to agentialism and emphasised the prominence of the unsuccessful pragmatic pressures approach. Here I will say why I agree to an extent with the IBE's first premise. I take it that other accounts of self-knowledge, at least *as they stand*, cannot satisfactorily explain self-ignorance and confabulation (I revisit my hedge in chapter six). While the proponent of RTM can explain self-ignorance, the account as it stands cannot explain confabulation. As such, we need to accept either computationalism or inferentialism on explanatory grounds.

This section proceeds as follows. First, I argue that Carruthers is right that subjects use computation/inference in confabulation cases; i.e., I agree that proponents of RTM face the TWO METHODS PROBLEM (§4.4.1). Second, I agree that RTM's proponents are unable to say in a non-*ad hoc* way why it is subjects use computation/inference in confabulation cases (§4.4.2). As such,

it does indeed look like distinctive access accounts like RTM cannot explain self-ignorance and error; instead we need to accept that all self-ascriptions are formed computationally/inferentially.

Before continuing, though, I should clarify that I take the IBE to give us reason to accept either inferentialism *or* computation. Philosophers face the TWO METHODS PROBLEM when they take there to be two distinct methods for acquiring self-knowledge that operate at different times. So, to sidestep this, we only need to say that one method operates in all instances of self-ascription: either inference (which in turn may well be underpinned by computational processing) or just computation. We can then choose between inferentialism and computationalism on independent grounds. This is useful to note; I go on in chapter six to specifically accept a form of computationalism while rejecting inferentialism.

4.4.1 Why think we use computation/inference in confabulation cases?

I agree with Carruthers that the proponent of RTM faces the TWO METHODS PROBLEM: subjects use computation/inference in confabulation cases, so the proponent of RTM must explain why subjects rely on this as opposed to RTM. To make this point as convincingly as possible, I will set out the alternative – thinking that subjects form both knowledgeable and confabulatory self-ascriptions using RTM (§4.4.1.1). I then argue against this alternative in the following way: at least some confabulation will be motivated, and the best account of motivated confabulation appeals to inferential/computational processing (§4.4.1.2).

4.4.1.1 Denying the *two methods* assumption?

Carruthers takes it that even if subjects do possess a distinctive method for acquiring self-knowledge, they fail to use it in confabulation cases. Instead, the mistaken self-ascription is formed by computation/inference.

But we might reject this thought, and say that subjects use the special method, e.g., RTM, in confabulation cases as well as knowledgeable ones. We need not think that RTM is infallible, or even that it is especially reliable. We only need to maintain that RTM is reliable enough that it can still issue in knowledge. As a result, we could accept that using RTM issues in the mistakes we see in confabulation cases. Cox (2018) gives this sort of explanation of confabulation using his own account of the transparency method, and we could apply what he says to RTM as well to give the following picture. RTM will result in errors when the world-directed question doesn't match up with subjects' actual motivating reason. E.g., the subjects in the stockings experiment may well form their self-ascription by answering the question 'why have that preference?'. The subjects conclude that 'the sheerness of the stockings is a good reason' and then falsely conclude that they

prefer the stockings for the reason that the stockings are sheerer. This explanation of confabulation receives further support from the fact that confabulation tends to involve mistakenly self-ascribing motivating reasons as opposed to purely causal factors (Cox rightly notes this, and I further discuss it in chapter five). Using RTM issues in reason explanations, and so it seems plausible that subjects are simply using RTM in these non-veridical cases.

4.4.1.2 Accepting the *two methods* assumption

I argue, though, that the above move fails, and that confabulatory self-ascriptions are formed by computation/inference. This is because at least some instances of confabulation are motivated. And the best way to explain how motivational factors bring about confabulation presupposes that computation/inference forms the confabulatory self-ascription. Indeed, my argument is especially relevant. As I will argue in chapter six, a*ll* instances of confabulation are in fact motivated; even given that fact, the *dual methods* theorist still cannot explain confabulation.

To start, then, motivational factors non-controversially play a role in at least some confabulation cases, but they don't provide the full story. For example:

> ALICIA: Alicia believes that a new colleague Bernice is unpleasant. When asked why, Alicia replies that Bernice did not smile at her in the corridor. Say also that Bernice is black, and that Alicia habitually overlooks black candidates for jobs, and so on. This has been pointed out to Alicia time and time again, but she just shrugs and tries to explain it away. We would say in this case that Alicia was mistaken about why she believes that Bernice is unpleasant. Alicia is ignorant of her racism and instead mistakenly ascribes a motivating reason. Alicia does not actually believe that Bernice is unpleasant for the (supposed) reason that Bernice didn't smile at her. And further, an obvious explanation of this mistake is that Alicia wants to not be racist.

We need to provide a mechanism by which the subject's desire causes her to confabulate. That is, Alicia's desire not to be racist plays some role in bringing about her self-ignorance and confabulation, but in what way?

Let me briefly mention one unsuccessful strategy for accounting for the role of motivation in confabulation; this will highlight the advantages of the picture I accept. One might endorse an account of repression (which Cassam discusses and rejects in 2014). The thought goes that motivational factors lead subjects to *repress* the relevant mental states.[19] For example, under this

[19] Wilson and Dunn refer to the repressed metal states in question as being 'thought, feelings, or memories' (Wilson and Dunn 2004: 495, quoted in Cassam 2014: 194) in the context of self-ignorance of our attitudes.

approach, the subjects in the stockings experiment might be said to repress a belief/judgement that they lack a motivating reason for their preference. This involves preventing the belief from becoming conscious without being aware of doing so (Wilson and Dunn 2004: 495-6). Yet this is implausible. There is evidence indicating each of the component parts of the process of repression, but 'no single study has demonstrated all the necessary criteria to establish the existence of repression definitively' (2004: 498).

A better approach regarding these cases is to see confabulation as an instance of self-deception construed along the lines proposed by Alfred Mele (2001). This approach is appealing as it requires few additional commitments. We already have independent reason to think that self-deception takes place. And further, Mele's account of it is particularly economical since a lot of the work is performed by the operation of cognitive bias which we can uncontroversially accept.

Mele offers the following account. Our desires can motivate self-deception since they lead us to: underestimate and overestimate the importance of given pieces of evidence, pay more notice to certain pieces of evidence at the expense of others, and use particular methods of acquiring evidence, all in accordance with what would speak in favour of the result we want (*Ibid.* 26-7). And desires have this effect, Mele thinks, by interacting with cognitive biases (*Ibid.* 28-31). He mentions three such biases. The first concerns the 'vividness of information.' When forming a belief, we are more likely to take information into account if it is vivid. Desiring something to be the case makes relevant pieces of evidence more vivid and can influence the resulting belief in this way. Secondly, we follow the 'availability heuristic.' The ease with which we can recall tokens of a particular type (i.e., their availability) leads us to think they are disproportionately representative of that type. Since our motivations lead certain pieces of data to be more vivid, and vivid data is more available, our desires can influence belief formation by way of this heuristic as well. And thirdly, the 'confirmation bias' means that:

> People testing a hypothesis tend to search (in memory and the world) more often for confirming than for disconfirming instances and to recognise the former more readily […] even when the hypothesis is only tentative (as opposed, e.g., to a belief one has) (*Ibid.* 29).

And our desires influence the hypotheses we have, and therefore go on to confirm them, since 'favourable hypotheses are more pleasant to contemplate than unfavourable ones and tend to come more readily to mind' (*Ibid.* 30).

I think, though, that my formulation of the account I give in the main text would be a natural way of extending the account in relation to why we have our attitudes.

The thought, then, is that subjects confabulate because motivational factors bias their inferential processing and the sort of evidence they use. We could say the following regarding Alicia. Alicia confabulates because she is motivated by a desire not to be racist. This desire leads Alicia to overlook evidence that she is racist, e.g., that she doesn't hire black candidates, and that others tell her that she's biased. The desire also leads Alicia to place too much weight on other pieces of evidence, e.g., the fact that she takes Bernice's blank expression to be reason for disliking Bernice, and that Alicia has a black acquaintance. Processing the evidence in this way leads Alicia to conclude that she believes that Bernice is unpleasant for the reason that Bernice didn't smile at her.

I have argued that motivational factors influence confabulation by shaping how subjects process evidence in forming the self-ascription. That is, motivational factors work by influencing the transitions involved in computation/inference. To avoid sacrificing parsimony, then, we should think that all confabulatory self-ascriptions result from computation/inference.

Before ending this subsection, let me consider an objection. Perhaps the proponent of RTM might nevertheless deny that subjects use computation/inference in confabulation cases, by applying Mele's insights in the context of RTM. One might say the following. Subjects use RTM when confabulating, and motivational factors influence their use of the method. Subjects' desires lead them to take a given consideration to be a normative reason that they would not normally take to be a reason, and to overlook evidence that the consideration is not a reason. E.g., we might say the following about Alicia. To explain why she believes that Bernice is unpleasant, Alicia answers the world-directed question 'why believe that Bernice is unpleasant?'. She concludes that a good reason is that Bernice did not smile at her on one occasion. Alicia's desire not to be racist causes that consideration to be especially salient and causes her to overlook evidence that suggests that it is in fact a bad reason. These defeaters include the fact that someone can be generally friendly but fail to smile on a given occasion due to stress, and so on. Alicia then forms her confabulatory self-ascription accordingly.

I reply to this option in the following way. It seems unintuitive that Alicia's desire not to be racist would only influence her self-ascription in such a limited way – just leading her to take a consideration to be a reason that she wouldn't normally. Intuitively, Alicia's mistaken self-ascription also results from how she places undue importance on the fact that she has a black acquaintance and ignores the fact that she only hires white candidates. That is, Alicia's treatment of other evidence concerning why she dislikes her colleague also plays an important explanatory role. Plausibly, Alicia's desire not to be racist influences how she processes a range of facts, and

not just her use of RTM. As such, then, we should think that confabulatory self-ascriptions are formed by computational/inferential processing.

4.4.2 *Can* we overcome the *two methods* problem?

I have argued that the proponent of RTM indeed faces the TWO METHODS PROBLEM. Subjects use computation/inference in confabulation cases, and a *dual methods* account must state in a principled way when subjects do and do not use this alternative method. I agree with Carruthers that we cannot provide such an answer to the PROBLEM. This is because even if motivational factors were present in all cases, they still wouldn't explain why subjects rely on one method over another. Appeal to pragmatic pressures, say, still wouldn't be enough even if they were always present. Indeed, this counterfactual is particularly pertinent. As I will argue in chapter six, all confabulation cases are motivated. But still, this motivational factor isn't enough to overcome the TWO METHODS PROBLEM and fully explain confabulation.

For subjects' desires to explain why they use computation/inference as opposed to RTM, subjects must have some sort of awareness of the fact they wish to avoid. E.g. we would need to say that Alicia is at some level aware that she is racist for her desire to have the required motivational impact. But thinking that subjects have this awareness is unattractive in at least three ways. (i) We would have to say that subjects have this awareness in virtue of an additional inferential mechanism that always operates. After all, subjects wouldn't use the distinctive method to become aware of the troubling factor – e.g., Alicia's racism. But accepting that this awareness is inferential just amounts to accepting the Orthodoxy. We would explain why subjects infer in confabulation cases by saying that subjects always employ inference in attempting to learn why they have their attitudes. Yet this is what Carruthers wants to say in the first place. (ii) In confabulation cases, the subjects are not actually at all aware of the fact they wish to avoid – e.g., that they are racist. As such, the suggestion is *prima facie* implausible. (iii) Accepting that subjects always engage in this sort of inference sits uneasily with agentialism. We would have to say that people take themselves as the objects of inquiry even when using RTM. Yet, according to RTM, individuals engage in deliberation about the external world in which they do not take themselves to be such an object.

As such, I agree that RTM as it stands cannot explain confabulation. This is because confabulations are formed by computation/inference, but *dual method* accounts cannot satisfactorily explain why subjects use computation/inference as opposed the distinctive method in these cases. Therefore, we must accept that all self-ascriptions are either formed by computation or by inference.

4.5 Conclusion

We have, then, reason to accept the Orthodoxy. Specifically, we have reason to either accept that all self-knowledge of why we have our attitudes and perform actions is acquired using computation, or that it is acquired using inference (which may be underpinned by computational processing). This is because *dual method* theories cannot explain confabulation since they cannot overcome the TWO METHOD PROBLEM.

Chapter 5 Problems with the Orthodoxy

Let me recap where we stand. I have introduced the Orthodoxy about self-knowledge of why we hold our attitudes. The Orthodoxy claims that self-knowledge of why we hold our attitudes does not significantly differ from our knowledge of why other people hold their attitudes. There are two forms of the Orthodoxy. Computationalism takes it that self- and other-ascriptions are formed by similar subpersonal processing. Inferentialism takes it that self- and other-ascriptions are both formed using inference, at the personal level. So, proponents of the Orthodoxy, such as Cassam and Carruthers, endorse either a computationalist or inferentialist account and deny that there is anything else significant that would render self-knowledge significantly different from other-knowledge. (To anticipate my position, I will accept computationalism, but reject the Orthodoxy.)

Then in chapter four, I introduced the best case for the Orthodoxy. I concluded that one should accept either computationalism or inferentialism. At least of the options on the table, then, it seems that we should accept the Orthodoxy. I argued that we should accept computationalism or inferentialism because false self-ascriptions are formed using computation/inference. If we claim that at least some knowledgeable self-ascriptions are not formed in this way but using another method (e.g., RTM), then we need to say when it is subjects use computation/inference and when they do not. But we cannot say in a non-ad hoc way when subjects will use the other method as opposed to RTM. As a result, we have reason on explanatory grounds to accept that just one method underpins all self-ascriptions: computation/inference.

This chapter argues that, even though we have reason to accept the Orthodoxy, the Orthodoxy also faces a range of problems; we thus reach an impasse. The chapter comprises two parts. §5.1 returns again to the Orthodoxy's inference to the best explanation (IBE). I still agree that the IBE gives us reason on explanatory grounds to accept either computationalism or inferentialism, because *dual method* accounts cannot explain the data. But the Orthodoxy's own explanation is weaker than it initially seems (i.e., here I challenge CLAIM TWO of the IBE). §5.2 criticises the inferentialist and computationalist accounts more broadly.

5.1 The Orthodoxy's Inference to the Best Explanation

The Orthodoxy's explanation of self-ignorance and confabulation is weaker than it initially seems. This is because neither inferentialism nor computationalism fully explain the patterning of subjects' mistaken self-ascriptions. That is, neither account can fully predict when subjects will make mistakes and the sorts of mistakes subjects will make. Here, I take the 'patterning' of mistaken self-ascriptions to consist in regularities concerning both when it is that subjects make mistakes and the mistakes' contents. Inferentialism/computationalism can make correct predictions to an extent (which Carruthers (2013) discusses at length) but finds it hard to account for important details. §5.1.1 will outline a key pattern concerning confabulation cases, and §5.1.2 discusses the Orthodoxy's difficulties accounting for it. I should note from the start the most obvious prediction the Orthodoxy can make: S will mistakenly explain an attitude/action with reference to *x* when S would mistakenly explain another person's attitude/action with reference to *x*.[1] As I will show, this conflicts with the data.

5.1.1 Explananda: The Confabulation Asymmetry

We can note a key pattern in confabulation cases – what I will call the Confabulation Asymmetry:

> Subjects tend to mistakenly ascribe motivating reasons to themselves more readily than to others.

Subjects make mistakes that they would not make about another person in an identical situation – the subject would instead explain the other person's attitude correctly or make a mistake with a different content. In arguing for the Confabulation Asymmetry, I will first present empirical data that subjects' mistaken self-ascriptions tend to be of the form that the subject has a motivating reason. I then present data that suggests that subjects make mistakes with this content even when they wouldn't do so concerning other people.

First, let me motivate the thought that subjects tend to mistakenly self-ascribe motivating reasons as opposed to purely causal reasons. While not frequently noted, we can see this pattern in the stockings experiment and Haidt's study, and also in a range of other experiments.[2]

[1] Cox (2018) discusses this prediction. Carruthers himself suggests something very similar – that his account 'predicts that confabulation should occur whenever there is sensory evidence of a sort that might mislead a third party' (2013: 3).

[2] For examples, see Hall et al. (2013), Johansson et al. (2005), Johansson et al. (2006) (all choice blindness studies), and Haidt (2001). Also, for a similar pattern in explanations of action, see for example: studies such as those involving split-brain patients (e.g. Gazzaniga (2000)) and hypnotised subjects (as discussed, e.g., in Wegner (2002)). This pattern is noted in Cox (2018), Knobe and Malle (2002), and Sandis (2015).

For instance, subjects make this sort of mistake in choice blindness studies.[3] Choice blindness occurs when individuals select something, say an object or a theoretical position, and fail to notice when it is switched for another. In some experiments, the subjects are then asked why they picked the item (which, unbeknownst to them, they had not in fact selected). This then leads them to confabulate a motivating reason.

We find one example of choice blindness in (Hall et al., 2012). The experimenters told participants to mark their agreement with various ethical statements on a scale of 1 (completely disagree) – 9 (completely agree). Afterwards, some of the statements were reversed and read to the subjects under the guise of what they had agreed or disagreed with. Subjects were frequently unaware of this swap, either at the time or when asked afterwards, and '69% of all the participants accepted at least one of the two altered statement/rating relations' (*Ibid.* 4). More relevant for us is what happened when the experimenters asked the subjects why they held this view (which they did not originally select). The participants provided explanations which were false insofar as they pertained to a view that they did not hold.[4] These explanations took one particular form: mistakenly attributing motivating reasons to themselves. For example, two participants originally agreed that 'even if an action might harm the innocent, it can still be morally permissible to perform it.' The rating they gave was then reversed, e.g., from a 9 to a 1. The subjects tried to explain why they supposedly held the opposite position with the following:

> "No, no one should have to get hurt"
>
> "No, well, I don't think it's ever ok … I'm not exactly sure how to explain this, but innocents should never be hurt, you know, one should always find other ways of doing it" (Hall, Johansson, & Strandberg 2012: supporting information 1).

The participants, then, provided motivating reasons as opposed to a purely causal explanation. They did not, for example, reply with 'I think hurting innocents is wrong because my parents drummed it into me and I've internalised the lesson well.'

Second, subjects tend to cite motivating reasons when making mistaken *self*-ascriptions in particular. The type of confabulation that interests me here specifically concerns explanations of our own attitudes. The propensity to mistakenly ascribe motivating reasons seems to occur regarding self-ascription in particular. This is not to say that we make more mistakes about our

[3] Thanks to Jordi Fernández and Ema Sullivan-Bissett for pointing out the applicability of these cases. See Scaife (2014: §2.4) for a discussion of choice blindness in relation to confabulation.
[4] I should note, though, that Hall et al. construe choice blindness in terms of the subjects' attitudes actually changing (e.g., *Ibid.* 5). McIver Lopes (2014) also favours this interpretation. Yet even if this is the case, the subjects still make a mistake about why they have this new attitude.

minds, but simply that the mistakes we do make exhibit a pattern which contrasts with other-ascriptions (both veridical and false). Unfortunately, this asymmetry has not been directly tested for. Nevertheless, I take the claim to be intuitive, and further, it receives support from the following two sets of studies concerning the *actor-observer asymmetry* and the *bias blind spot*.

5.1.1.1 Actor-observer asymmetry

Subjects tend to provide motivating reasons when explaining their actions while observers give more purely causal explanatory reasons (Malle et al., 2007).[5] We can see one instance of this pattern in study five of Malle, Knobe, and Nelson (2007). Here, individuals were asked to "describe 'the last time [they] had an interesting conflict with a romantic partner, friend, or parent'" (*Ibid.* 502), and to explain a range of their and their opponent's behaviours. Another participant, unrelated to the first, was also requested to explain the same behaviours based on the first subject's account of the conflict. The actors' explanations contained a greater number of motivating reasons compared to the observers', and the observers' explanations included a larger quantity of purely causal explanatory reasons. Indeed, this was the case regardless of whether the observer knew the actor (*Ibid.* 503).

Admittedly, in investigating the explanation of action, the studies do not examine this paper's specific concern – subjects' *erroneous* explanations of their *attitudes*. Yet, on the basis of the experiment below, this pattern would likely occur in subjects' explanations of attitudes as well.[6] And given subjects' tendency to confabulate, we can suppose that at least some of the self-ascriptions in the studies would have been false.

5.1.1.2 The bias blind spot

Subjects tend to mistakenly think that they, but not others, form attitudes in a bias-free way. This stems from the bias blind spot, whereby subjects notice their own biases less than other people's (Pronin et al., 2002).[7] We can see this in the following study by Pronin, Lin, and Ross (2002). In it, pairs of subjects completed a 'social intelligence test' with one being told their mark was above average, the other, below average. When asked to what extent they thought it was a good test and that its results would match up with those of similar ones, those who were told they did relatively well were likelier to appraise it higher on both fronts. Further, the subjects were then

[5] Malle et al. present the contrast in terms of 'reasons' and 'causal history' explanations, but I do not think our terminology differs substantively. On this asymmetry, see also Malle (2011) and Knobe and Malle (2002). This, they persuasively argue, is the best way of understanding the self and other asymmetry that some have tried to account for in terms of the *fundamental attribution error*, e.g., Jones and Nisbett (1972).
[6] Jones's discussion (2002: 227-8) and the cited study in Gilbert and Mulkay (1984) also suggest this.
[7] Shermer (2012) also gives a nice summary.

informed about 'a self-protective tendency' that leads to such results in one's views about an assessment. Yet when asked, the individuals were more likely to take the other participant's views about the quality of the test as having been affected by their results, than to realise that the same could be said about their own. Individuals are on occasion, then, more likely to falsely maintain that their own judgements in particular do not result from purely causal explanatory reasons that indicate bias.[8] Further, following earlier observations, we can suppose that if the subjects were asked why they made the judgement they did, they would confabulate motivating reasons in their own case but provide (correct) purely causal explanations of others' judgements.

Subjects are more likely, then, to mistakenly use motivating reasons when explaining their own attitudes compared to those of others. At the very least, it is a plausible prediction, and making it would be a mark in favour of an explanation of confabulation.[9]

5.1.2 The Orthodoxy and the Confabulation Asymmetry

Inferentialism/computationalism has difficulty accounting for the Confabulation Asymmetry when explaining the patterning of cases. First, the Confabulation Asymmetry is incompatible with the prediction that most obviously falls out of the Orthodoxy: subjects make mistakes, and the same sorts of mistakes, that they make in other-ascription cases. But we have seen that subjects' mistakes exhibit a pattern we don't see with other-ascription: subjects tend to mistakenly ascribe motivating reasons to themselves more readily than to others. In the remainder of this section, I reject another way in which the Orthodoxy might try to explain the Confabulation Asymmetry and then end with some general remarks about the Orthodoxy's prospects.

The Orthodoxy's proponents might make the following suggestion. Subjects mistakenly self-ascribe motivating reasons when they would not ascribe them (correctly or incorrectly) to others because of the additional evidence subjects have concerning themselves, such as mental images and feelings. Wilson, for example, writes that our extra evidence can sometimes serve as red-herrings, so to speak, and result in mistakes (2002: 108-10).[10] So subjects will make mistakes when they have uniquely misleading evidence about themselves.

[8] Pronin does write elsewhere that 'sometimes the 'bias blind spot' is primarily caused by people's unwarranted denials of their own biases, whereas at other times it is more attributable to people's overestimations of others' bias' (2007: 41). The subjects in the above study, though, do underestimate of the influence of bias in themselves (see Pronin et al. (2002: 377)), and therefore do make false self-ascriptions.

[9] See also Nisbett and Wilson (1977: 273) for a case in which subjects recognise the possibility that others' tolerance for electric shocks might have been manipulated by the experimenter, but not their own.

[10] Pronin and Kugler (2007: 566) consider this option as a way of explaining the bias blind spot but reject it.

Here, though, I say two things in reply. First, Pronin and Kugler (2007) performed an experiment which suggests additional evidence does not cause the bias blind spot. It was similar to Pronin et al. (2002) as outlined above. This time, though, the experimenters gave some participants reports of what the other subject was thinking about when appraising the test. Yet, access to this information barely affected the degree to which individuals thought the other's score in the test influenced their judgement about it (2007: 571-2). Second, and more generally, subjects do not just have additional evidence suggesting that they formed their attitudes on the basis of reasons. Some of it favours believing that they lack a reason for their attitudes. For example, say the subject in the stockings experiment holds that people often prefer stockings on the basis of their sheerness. Yet, they have evidence that places doubt on the applicability of that theory in their own case – that they do not remember deliberating about the stockings before picking their favourite, say, or the fact that the stockings currently look the same to them. Carruthers, Wilson and company, then, need to say why only some additional evidence influences subjects' ascriptions.

More generally, the Orthodoxy will find it hard to explain the patterning of cases in a way which accommodates the Confabulation Asymmetry. In doing so, the Orthodoxy must still keep their explanation as parsimonious and unified as possible. They should avoid appealing to a long list of heuristics (processing rules).[11] Such an appeal would preserve their explanation's scope only to sacrifice other explanatory virtues.

***

To conclude, then, the Orthodoxy fails to satisfactorily explain the patterning of self-ignorance and confabulation. This means that the Orthodoxy's explanation of such cases is not as good as it could be. As such, if an account of self-knowledge could overcome the TWO METHODS PROBLEM *and* explain the patterning of confabulation cases, we would have reason to accept it over Cassam's and Carruthers'.

5.2 The Orthodox account in general

I now argue that the Orthodoxy is generally unappealing. I will say why we should reject the Orthodoxy's account of both the method underpinning self-knowledge (§2.1) and the way in which self-ascriptions are warranted (§2.2).

[11] E.g., Pronin (2007) references three in explaining the bias blind spot.

5.2.1 The method underpinning self-knowledge

This subsection criticises the Orthodox account of the method underpinning self-knowledge of motivating reasons. Recall that proponents of the Orthodoxy either accept a computational or an inferentialist account of the method used to acquire self-knowledge of motivating reasons. That is, the Orthodoxy takes it that this type of self-knowledge is importantly akin to other-knowledge, and that we acquire both either as a result of inference or computational processing. I will provide two arguments against this view.

5.2.1.1 The dual role of the question 'why?'

Recall the question 'why?' – you know that a subject has an attitude or is performing an action, and you ask them 'why?'. For example, you know that Sally believes that *it will snow*, or that she is going to the shops, and you ask her 'why?'. This question has a dual role, which I will use in the Dual Role Argument:

> PREMISE ONE. The question 'why?' has a dual role (either when posed by others or ourselves). The question at once requests an explanation and justification for the attitude/action.
>
> PREMISE TWO. We wouldn't be taking the question 'why?' seriously if we used inference to answer it or if the self-ascription was formed by computational processing alone.
>
> PREMISE THREE. We take the question 'why?' seriously.
>
> CONCLUSION: Our answers to the question 'why?' are not formed by inference or computational processing alone.

I will motivate the premises in turn.

According to PREMISE ONE, the question 'why?' does not just request an explanation for the attitude/action: it also asks for what justifies the attitude/action.[12] So, when you ask Sally 'why?' concerning her belief that *it will snow*, you both ask what explains her belief and for evidence in its favour. Sally fails to answer the question 'correctly' if she does not provide this justification. That the question 'why?' has this role shouldn't be too controversial since we need not commit to any view about what grounds it. I take it that the question asks for justification (as opposed to just explanation) because of our rational agency and responsibility regarding our attitudes. But, one might reject this. One might instead think that the question's dual role stems from the nature of

[12] See Anscombe (2000).

interpersonal exchanges which subjects have internalised such that the question plays the role even when they pose it to themselves.[13]

One reason for thinking that the question 'why?' plays this dual role is that we do not just treat others' answers to it as explanations. We also, and indeed primarily, assess them as justifications. For example, say that Sally tells you that she believes that *it will snow* because of the weather forecast. Here you could intelligibly engage with the motivating reason Sally provided *qua* justification. You may well offer defeaters for this justification, e.g., 'there's just been a flood warning' (a countervailing defeater) or 'you looked at a weather forecast for the wrong place' (undercutting).[14] This is not to comment on her answer as an explanation. The fact that, unbeknownst to Sally, her justification for believing that *it will rain* has been defeated does not say anything about the reason for which she actually holds that belief.

PREMISE TWO states that we wouldn't be taking the question 'why?' seriously if we used inference or computation alone to answer it. (By 'computation alone' I wish to preclude the option that computational processing underpins a personal level method other than inference. An account like that would avoid the Dual Role Argument, as I argue in chapter six.) This is because computation/inference is not a good method to use when providing justification for the lower-order attitude. At least, it is not a good method if we use the specific sorts of transitions and inferences that the Orthodoxy appeals to. Merely considering the evidence about what our motivating reasons are does not take into account whether they are good reasons. We might now be aware of stronger reasons. And just because we once took the considerations to be good reasons, it does not mean that, as far as we are concerned, they still are. Also, using inference/computation can give rise to purely causal explanations as well as reason explanations. That is, if Sally answers the question 'why do I believe it will snow?' in this way, she would just as appropriately reply 'because I am generally rational.' Yet this in itself is not to justify the attitude, unlike if Sally used a reason explanation, e.g., 'for the reason that the weather forecast says it will.'

PREMISE THREE states that we do generally take the question 'why?' seriously. It is not that we disregard it completely. This seems intuitive: people do generally seem to explain their attitudes with reason explanations that both explain and justify their attitudes.

[13] E.g. Mercier and Sperber (2011) argue that our capacity for reasoning evolved for interpersonal argumentation. It is because of this that we, say, try to find reasons in support of our claims.
[14] On types of defeat, see e.g. Schroeder (2015: 227).

This then leads to the CONCLUSION. The question 'why?' has a dual role with which subjects generally seem to act in accordance. Yet subjects would not be doing so if they used computation/inference. Therefore, we should reject the claim that subjects answer the question using computation/inference.

5.2.1.2 Rational relations

My second argument concerns the thought that our attitudes bear direct rational relations to our explanations of them. This sort of strategy should be familiar from the discussion of agentialism in chapter two. I will call this the Rational Relations Argument:

> PREMISE ONE. Our explanation of our attitude *A* bears direct rational relations to *A*.
>
> PREMISE TWO. Inferentialist and computationalist accounts of the method underpinning self-knowledge are incompatible with PREMISE ONE.
>
> CONCLUSION. Inferentialist and computationalist accounts of the method underpinning self-knowledge are false.

PREMISE ONE states that our explanations of our attitudes directly bear on the rationality of those attitudes and vice-versa. That is, when I believe that *I believe that p because q*, it is rationally evaluable as a belief about my motivating reason and also as a justification of my belief that *p*.

This can be clarified by looking at Moore paradoxical (MP) statements. Recall that a standard MP statement concerns belief, and can take the form:

> I believe that *q*, but *q* is false. (E.g., I believe that *it will snow*, but it will not snow.)[15]

We can also formulate parallel MP statements concerning motivating reasons. It would be most odd for a subject to assert or believe one of the following, and their doing so would suggest that they are not ideally rational:

> I believe that *q* for the reason that *p*, but *p* is false. (E.g., I believe that *it will snow* for the reason that the weather forecast says that it will snow, but the weather forecast does not say that it will snow.)

[15] Other MP statements for belief take the form '*p*, but I do not believe that *p*.' These, though, lack clear correlates in the case of motivating reasons.

> I believe that *q* for the reason that *p*, but *p* is not a reason for believing that *q*. (E.g., I believe that *it will snow* for the reason that the weather forecast says that it will snow, but the weather forecast is not a reason for believing that *it will snow*.)

Indeed, additional combinations of views are rationally prohibited concerning why we have our attitudes:

> I believe that *q* for no reason. (E.g., I believe that *it will snow* for no reason.)
>
> I do not know why I believe that *q*. (E.g., I do not know why I believe that *it will snow*.)
>
> I believe that *q* because of a purely causal explanatory reason. (E.g. I believe that *it will snow* because I want to go sledging.)

In all these cases, believing that *it will snow* isn't rational by the subject's own lights – it doesn't make sense to her as something to believe. And also, by her lights, these self-ascriptions are irrational. And that, it may seem, is because direct rational relations hold between the self-ascription and the lower-order belief. E.g., it doesn't make sense to the subject to believe that 'I believe that *q* for the reason that *p*' *because* she takes *p* to be a bad reason. And it doesn't make sense to ascribe 'no reason' for the belief *because* doing so makes it irrational to continue holding the belief. To clarify, these are bi-directional rational relations. The rationality of the self-ascription affects the rationality of the lower-order belief, and the rationality of the lower-order belief affects the rationality of the self-ascription.[16]

What about PREMISE TWO – that inferentialism and computationalism are incompatible with these direct rational relations? The rational norms governing self-ascription differ from those governing computation/inference in important ways. Under inferentialism/computationalism, subjects could, at least sometimes, self-ascribe a motivating reason while believing that 'it is a bad reason' and not do anything rationally wrong. This is like how it can be rational from your perspective to believe that Sally has the motivating reason that *p*, even if you also think that *p* is a bad reason. I will say more concerning inferentialism and computationalism in turn.

Regarding inferentialism, I can say the following of PREMISE TWO. When we form a self-ascription through inference, we are subject to certain norms. If we fail to satisfy these, then the self-ascription will not seem rational from our perspective. That is, it will not make sense to have that belief as far as we are concerned. One such (defeasible) norm is to base one's belief on reasons

[16] I take the rational relations to be bi-directional, but the Rational Relations Argument still holds if we take the relation to be one-way.

for holding that belief. As such, we would be perfectly rational in self-ascribing the motivating reason that *p* while believing that *p* is a bad reason provided we have sufficient evidence that we have that motivating reason. Further, if our self-ascriptions were inferential, we would not be subject to certain other norms. Namely, the rationality of the lower-order attitude would be unaffected by our self-ascription. It would make sense in our eyes to hold the lower-order attitude provided it is based on reasons, regardless of any self-ascription concerning it. After all, inference doesn't affect its subject matter in any way. E.g., it will still be sensible from Sally's perspective to believe that *it will snow* because she does so for the reason that *the weather forecast says that it will snow*, even if she also believes that *I believe that it will snow for no reason*. But in the self-ascription case, whether it makes sense for us to hold the lower order belief depends on the higher-order one, and vice versa.

Computationalism finds it even harder to account for the rational relations because computationalism operates at the wrong level of explanation to say anything about the subject's rationality. After all, the computations are performed by the mindreading module. The module does not form beliefs on the basis of *reasons*; beliefs cannot seem *rational* from the module's perspective. As such, nothing in computationalism fully accounts for how the higher- or lower-order attitudes seem rational from the subjects' own perspective, or how these attitudes rationally relate.

PREMISE ONE and TWO together, then, give us the conclusion: that inferentialism and computationalism are false.

***

I have provided two arguments for denying that self-knowledge of why we have our attitudes is acquired using computation or inference. Before continuing, let me note three things. i. I take the Dual Role and Rational Relations Arguments to importantly relate to each other: the dual role of the question 'why?' in part stems from the rational relations between the subject's explanations of her attitudes and the attitudes themselves. But, we might reject this, and one can accept the first argument while rejecting the second. ii. The phenomena cited in the arguments relate to the obligation for self-knowledge I introduced in chapter two – the *knowledgeable reason explanation* obligation. I return to this in the conclusion. iii. These two arguments would also apply to a quasi-perceptual account of self-knowledge of motivating reasons. Recall the position in logical space that takes it that S learns that she believes that *q* for the reason that *p* because S quasi-perceives that she believes that *q* for the reason that *p*. S would quasi-perceive that she has the motivating reason in a similar way to how S perceives that Ellie cut the brownies. But the above arguments also speak against accepting a quasi-perceptual account. If S uses quasi-perception to learn that

she has a given motivating reason, she would not be taking the question 'why?' seriously. After all, S's detection mechanism would detect whether S has the motivating reason that *p* but not whether *p* is a good reason. Also, the quasi-perceptual account is incompatible with there being direct rational relations that hold between S's belief that *q* and S's belief that *she believes that q for the reason that p*. (This point should be familiar from the initial discussion of agentialism about self-knowledge of belief in chapter two.) The norms governing quasi-perception would be like those governing perception and at odds with the norms that seem to govern self-knowledge. After all, the rational status of S's motivating reason would not affect how well S is quasi-perceiving it or vice versa.

5.2.2 The warrant for self-knowledge

I will now argue that neither computationalism nor inferentialism can provide a satisfactory account of the way in which our reason explanations are warranted. Inferentialism and computationalism fall short in different ways, so I will consider the views separately.

5.2.2.1 Computationalism and warrant

Say we were to understand the method underpinning self-ascription as Carruthers does – a subpersonal mechanism that transitions between contents. We could only say that subjects' self-ascriptions are warranted in virtue of the fact that the subpersonal processing reliably results in true beliefs. After all, because the computationalist account operates at the subpersonal level, it doesn't make reference to personal level notions like the subject herself recognising reasons for the belief, or being able to access evidence herself. But a reliabilist account falls short for two reasons.

i. I reject reliabilism in general. It seems very much part of our everyday understanding of knowledge that the subject's belief must be in some way rationalisable in their eyes.[17] That is, it makes sense to the subject to hold that belief (I discuss this notion more in chapter 6.) Take the clairvoyant case from BonJour (1980). Norman has a reliable belief-forming mechanism – his clairvoyance – which gives rise to the belief that the president is in New York. The belief is indeed true, but as far as Norman himself is concerned, the belief just pops into his head out of nowhere,

[17] In considering the reliabilist's explanation of how we do understand knowledge, as opposed to, say, a reliabilist account of how we *should* understand it, we can target arch-reliabilist Goldman (2000) on the terms of his own project.

and he will be as surprised as anyone if it turns out to be true. It seems that we would not want to call Norman's belief knowledge, despite the reliability of his sixth sense.[18]

ii. Even if we accepted that reliability alone *can* ground warrant, we should still deny that it grounds the warrant underpinning self-knowledge. Supposing for a moment that reliable belief-formation suffices for warrant, it nevertheless does not suffice for *well-grounded* belief. Perhaps beliefs can be knowledgeable without being rationalisable to the subject, such as in Norman's case. Nevertheless, not all beliefs will be warranted in this way.[19] For example, my belief that *Managua is the capital of Nicaragua* will be well-formed – it makes sense to me to hold the belief, in this case, because I hold the belief on the basis of evidence that Managua is the capital of Nicaragua. Self-ascriptions, or at least self-ascriptions of the sort that are candidates for distinctive self-knowledge, seem to be instances of well-grounded belief. Suppose for a moment self-knowledge was not. We would be left with cases like the following. I want to go to the beach and James asks me why. The self-ascriptive belief pops into my head that *I want to go because of the sea breeze*, in much the same way as Norman's belief about the president came to him. My belief is true, and my mindreading mechanism is reliable, but I'm a bit baffled by it all and I couldn't justify the self-ascription in any way. If James asks me why I believe it is my reason, I just shrug and reply 'your guess is as good as mine!' Yet this certainly is not what happens in standard cases of self-knowledge. It makes sense to me that I have that reason, and if asked to justify my self-ascription I have something to say, even if it is just that 'I believe that *it's my reason* because it *is* my reason!' or 'I believe that *it's my reason* because it's the best reason!'. I should note, though, that saying that our self-ascriptions are well-formed is not to say that our self-ascriptions must be based on evidence. After all, recall the RTM account from chapter two. According to this account, our self-ascriptions are warranted by an agent's awareness of the motivating reason; subjects do not treat this agent's awareness as evidence that they have the motivating reason.[20]

We cannot, then, fully explain subjects' warrant for self-knowledge in terms of subpersonal processing.

---

[18] For other criticisms of reliabilism see: e.g., Cohen (2002), Conee and Feldman (1998), Plantinga (1993), Vogel (2000), and Zagzebski (2003).

[19] Here we might have something like Sosa's distinction between animal and reflective knowledge in mind (e.g., (2007)).

[20] Proponents of the RTM account would give the following explanation of the way in which subjects respond when pressed on why they believe that *they believe that q for the reason that p*. The above responses (e.g., 'I believe that *it's my reason* because it's the best reason!') are ways of trying to gesture at the agent's awareness that S has in virtue of justifying her belief that *q.*

5.2.2.2 Inferentialism and warrant

According to inferentialism, our self-ascriptions of motivating reasons are inferentially justified. Inferential justification requires that one base the belief in question, e.g., the self-ascription, on other, knowledgeable, beliefs. (One need not actually engage in inference itself.) So, we might say that Sally is justified in believing that *she believes that it will snow for the reason that the weather forecast says it will* because she bases the self-ascription on knowledgeable supporting beliefs. The evidential base will include facts such as that people normally believe that *it will snow* for the reason that the weather forecast says it will and that Sally checked the weather forecast earlier today. Sally's justification for believing the self-ascription is transmitted from her justification for the supporting beliefs. Unlike Carruthers, then, Cassam does not have to depend on reliabilism, and can therefore give a better account of the warrant underpinning self-knowledge. Inferential justification is, after all, standardly an account of well-formed belief, whereby the belief is rational in the eyes of the subject in virtue of being based on evidence.

But an inferentialist account of warrant is still unappealing regarding self-knowledge. This is because it relies on the inferentialist picture of the relevant basing relations. That is, the self-ascription is based on, and only on, evidence that one has the reason in question. But this faces the problems from §5.2.1 concerning inferentialism about the method underpinning self-ascription. Let me recap these arguments, clarifying how they apply to our present purposes.

First, recall the Dual Role Argument: to take the question 'why?' seriously, one must explain *and* justify the relevant attitude or action. But forming a self-ascription on the basis of evidence that one has a given motivating reason is not to take the question 'why?' seriously. It is not to take into account what justifies the given attitude or action.

Second, recall the Rational Relations Argument: whether it is rational by the subject's lights to self-ascribe a reason for an attitude directly affects whether it is rational to hold that attitude and vice versa. But this would not be the case under the inferentialist picture of the basing relations involved. For Cassam, the self-ascription of a reason is based just on evidence that one has that reason. If Cassam is correct, then a self-ascription would be rational if it is based on sufficient evidence, and the rationality of the lower-order attitude would be unaffected. But this is not what we see. In actuality, forming certain self-ascriptions (e.g., causal history explanations) exhibits irrationality, even if there is good evidence for them. And even if subjects have evidence in favour of an attitude, the attitude is still rendered irrational by their lights if they, say, cannot self-ascribe that evidence as their motivating reason for the attitude. The basing relations involved in self-ascription, then, differ from the sort underpinning inferential justification.

5.3 Conclusion

This chapter criticised the Orthodoxy in various ways. I conceded that we do indeed have reason to accept either computationalism or inferentialism on explanatory grounds. But, the Orthodoxy's explanation of the relevant data isn't as strong as it may seem. And, further, their accounts of self-knowledge are generally unappealing.

This seems to leave us with a troubling impasse. We still have reason to accept computationalism or inferentialism. While the accounts' explanations of confabulation are not as strong as they could be, accepting one of the two accounts is the only way to overcome the TWO METHODS PROBLEM. But we also have reason not to accept computationalism or inferentialism. Where do we go from here?

Chapter 6 The *Two Explanations* Account of Self-Knowledge

Let me recap where we stand. So far, I have introduced the Orthodoxy about self-knowledge of why we have our attitudes. This claims that self-knowledge of motivating reasons importantly resembles other-knowledge. There are two versions of the Orthodoxy. One version (*computationalism*) concerns the subpersonal level. Computationalism claims that subpersonal mechanisms form both self- and other-ascriptions by transitioning from representations concerning the relevant subject. The other version (*inferentialism*) pertains to the personal level. Inferentialism claims that both self- and other ascriptions are formed using inference, which the subject herself engages in. In chapter four, I argued that we have reason to accept the Orthodoxy – inference to the best explanation (IBE) concerning confabulation. We should either accept that all self-ascriptions are formed by inference or that they are all formed by computation. This is because *dual method* theories cannot explain confabulation. To recall, *dual method* theories take it that self-ascriptions are formed using either one of two methods. These methods are wholly distinct from each other: a special first-personal method, and inference/computation. So, the picture looked bleak for RTM at the end of chapter four. But then chapter five outlined various ways in which the Orthodoxy and its main argument are weaker than they seem. The Orthodoxy's explanation of confabulation fails to satisfactorily account for the patterning of self-ascriptions. And the Orthodoxy's account of the method and warrant underpinning self-knowledge faces deep worries. Indeed, these latter problems constitute (*pro tanto*) reason not to accept computationalism or inferentialism. So, chapter five ended with an impasse – we have reason in favour the Orthodoxy and reason against it, and seemingly no way forward.

But, some ideas from earlier offer a solution. I noted in chapter four that we can accept either inferentialism or computationalism to avoid TWO METHODS PROBLEM. That is, we can use independent grounds to choose which of the two accounts we accept. And as I mentioned in chapter two, a given phenomenon can be explained at both the subpersonal and personal level.

This chapter proposes a *two explanations approach* to distinctive self-knowledge. At the subpersonal level, both self- other-ascriptions are formed using computational processing. At the personal level, subjects sometimes use a distinctive method. In particular, I endorse the agentialist picture from chapter two, whereby subjects use the reasons transparency method (RTM). So, self-knowledge in some sense resembles other-knowledge, and yet nevertheless fundamentally differs from it. My account preserves the advantages of the Orthodoxy while

avoiding the disadvantages. Further, I also endorse a *two explanations* account of other sorts of self-knowledge, e.g., self-knowledge of attitudes. I return to this in the conclusion.

This chapter proceeds as follows. §6.1 reminds the reader of the personal/subpersonal distinction and draws a lesson from perception to say that we can explain a phenomenon in two ways. §6.2 draws on this to outline my *two explanations* account of self-knowledge. With that in place, §6.3 clarifies why we should accept my overall account and §6.4 considers two objections.

6.1 The personal/subpersonal distinction and a lesson from perception

Recall the personal/subpersonal distinction from chapter two. We can explain a given feature of a subject in two ways. Personal level explanations reference facts that we can attribute to the subject themselves. For example, we might say that Ben buys a Godzilla film because he finds them entertaining – this appeals to Ben's own states of enjoyment. Alternatively, we could also explain that same action at the subpersonal level. The subpersonal explanation references processes involving Ben but which we cannot attribute to Ben himself. For example, we might also say that Ben buys the Godzilla film because watching Godzilla films releases serotonin, which gives rise to a state of enjoyment. But Ben himself doesn't release serotonin. The subpersonal/personal distinction resembles how we can explain the sound of music in a concert hall at different levels: in terms of the orchestra or the individual players themselves.

I should emphasise that I understand the subpersonal/personal distinction in a broad sense: the personal level pertains to the subject themselves, and the subpersonal level concerns all the goings on that cannot be attributed to you, me, or anyone else. I will be neutral on how to cash out the subpersonal/personal distinction more precisely. The distinction is complex and would require a whole other thesis to fully address.[1]

The subpersonal and personal levels importantly relate. As a minimum, we can fairly uncontentiously say that the two levels relate via grounding, and that the subpersonal level grounds the personal level. For instance, to return to our example, that the individual musicians are playing grounds the fact that the orchestra plays a Bach orchestral suite. The players' individual playing makes it the case that the orchestra itself plays. Similarly, the serotonin in Ben's brain makes it the case that he feels enjoyment.

[1] For good overviews, see Bermúdez (Bermúdez, 2005) and Drayson (2014). Under some understandings, the subpersonal/personal distinction will pull apart from the player/orchestra one more than others.

As a further question about how the levels relate, we might wonder whether they are isomorphic. That is, do the patterns and structures we can pick out at the personal level match the structures we see at the subpersonal level? Let us return to our analogy. We could, for example, talk of the whole orchestra getting louder, i.e., crescendoing, because they also crescendo at the suborchestra level – each individual player gets louder. In this case, the two levels are isomorphic. Yet it is less clear that we could talk of playing an orchestral suite at the suborchestra level, or at least not without referring back to the orchestra level. After all, each person's performance may well be far from the finished piece – the second violin's part will contain little of the main melody, for example. Similarly, we might wonder whether the same can be said about certain aspects of our personal level explanations, namely particular patterns in behaviour and thought (e.g., watching television, daydreaming), and structured mental states (e.g., believing that *p*).[2] Do these directly map onto correlates in our subpersonal explanations? That is, are they more like crescendoing or playing an orchestral suite? I don't want to say anything too bold about the nature of subpersonal/personal explanations in general, but my explanations concerning self-knowledge will be isomorphic to a degree although not completely. This will be discussed and motivated throughout the course of the chapter.

Now for the lesson from perception. We can use the subpersonal/personal distinction when explaining self-knowledge in a similar way to how philosophy of perception uses it. My account takes self-knowledge to be computational at the subpersonal level of explanation but non-inferential at the personal level. This is like how, non-controversially, we can think that perception is computational without taking it to be inferential (which is not to deny that computationalism about perception is not, itself, controversial). The classic computationalist account of perception is that of David Marr (2010), which I will now outline.

Marr addresses perception at a specific level of explanation. He differentiates between three: 'computational,' 'algorithmic,' and 'implementational' explanations of a given process (2010: 19-27). It is worth noting, though, that they are not variants of the personal/subpersonal distinction; rather, Marr's three levels all help carve up the subpersonal level. Of these three levels, the computational is the highest. It concerns '*what* the device does and *why*.' For example, Marr notes that we can explain the operation of a cash register in terms of addition and the basic norms of this activity. The algorithmic level concerns *how* a particular mechanism carries out a process. Here, we determine the specific language of the representations and the intermediate processes. For example, the cash register will give an output in Arabic numerals, and as such, it

[2] For a related sort of analogy and useful discussion, see Bermúdez (2005) on Dennett's views on isomorphism.

uses algorithms specific to that number system rather than, say, binary. And the level of implementation concerns the physical underpinnings of the device carrying out the process, such as metal arranged in a particular way. Or, to try and get this distinction as clear as possible, I think we can also illustrate the three levels with a less mathematics-driven example. There are various ways in which we might explain the process of turning information about a friend into a limerick. First, we could talk of writing a limerick and following certain rules in doing so: using the limerick rhythm and AABBA rhyme scheme and starting with 'There once was a...' (this is analogous to the computational level). Alternatively, we might appeal to how the subject is writing it in English and talk about all the grammatical processes involved (algorithmic level). Or, we might talk of how the subject's hand moves on the page and the physical processes underpinning their thought (implementational level).

Marr focusses on explaining early vision at the computational level. He is concerned with how the input from the world is transformed into a visual representation which aids the subject (*Ibid.* 31). As such, the representation's content is of a three-dimensional perspective-independent thing in the world. The visual system reaches this by producing a series of intermediate pre-representations, each one taking the previous one as an input. So, the system takes primitives in the visual field and forms a primal sketch representing basic shape along two dimensions. This sketch is then processed to form a 2 ½-D sketch which represents information about the three dimensions of given surfaces (but not the objects behind those surfaces), and how they are arranged relative to the perceiver. This is processed again to produce the final 3-D representation which grounds our visual experience (what *you*, the subject, see when reading this page).[3]

Each process (e.g., forming a 2 ½-D sketch from the primal sketch) is computational in that the mechanism relies on various 'assumptions' in producing a new sketch from the input. For example, to form the 2 ½-D sketch, the processing system must combine distinct primal sketches from both retinas into one image. In doing so, the system uses the following assumption:

> [I]f a correspondence is established between physically meaningful primitives extracted from the left and right images of a scene that contains a sufficient amount of detail, and if the correspondence satisfies [several other constraints], then that correspondence is physically correct (*Ibid.* 114-5).

But, although vision is computational, vision is not inferential at the personal-level of explanation. No one thinks that the *subject* infers that if the images from their retinas resemble each other, then they concern the same object in the world. Indeed, the subject is unaware that this process

[3] Summaries can be found on Marr (2010: 36-8, 41-2, and Ch. 6).

of combining sketches – stereopsis – happens at all. Instead, at the personal level, the subject simply sees the object – they don't have to engage in inference in order to undergo the perceptual experience.

I want to paint a similar picture regarding self-knowledge – we can, and should, understand self-knowledge differently at the subpersonal and personal levels. Self-knowledge, though, still differs from perception. My subpersonal explanation of self-knowledge is not as 'low level' as Marr's computational account of vision. And more components underpin self-knowledge at both the personal and subpersonal levels – acquiring self-knowledge is not as straightforward as seeing something in the world. Nevertheless, the foregoing discussion of perception contains an instructive moral.

6.2 The *two explanations* account

This section provides my full account of self-knowledge. It starts with a rough example and some words about the relation between the personal and subpersonal explanations in my account. §6.2.1 and §6.2.2 then detail the personal and subpersonal level explanations respectively.

The following exemplifies self-knowledge of motivating reasons under my account:

> I believe that *it will rain*, and you ask me why. I consider the normative reasons for believing that *it will rain*, such as the fact that there are grey clouds, and that the grey clouds are a good reason for having the belief. I can then answer that 'I believe that *it will rain* for the reason that there are grey clouds' having acquired non-inferential self-knowledge. At the subpersonal level, computational processes underpin all this – both the deliberation, and the formation of the self-ascription. The self-ascription is formed by the mindreading module – a part of the brain responsible for both self- and other-ascriptions. It transitions from contents such as 'this system takes the grey clouds to be a good reason for believing that *it will rain*,' and 'this system is not biased' etc. The module then issues the representation: 'this system believes that *it will rain* for the reason that there are grey clouds.' This representation underpins my personal-level self-ascription that 'I believe that *it will rain* for the reason that there are grey clouds.'

My account takes the process underpinning any one self-ascription to be explained in two ways. We should not, though, confuse my account with views that take self-knowledge to be acquired by wholly distinct methods on different occasions, like the *dual method* accounts that incur Carruthers' ire. For example, traditional agentialist accounts take it that on some occasions

subjects learn what their attitudes and reasons are using a transparency method. On other occasions, subjects use inference (in the case of alienated attitudes and confabulation). Inference and the transparency method are distinct methods and are employed one at a time. Indeed, the methods are *wholly* distinct; they do not fundamentally resemble each other in any way. Instead, my account is:

> The *two explanations* account: There are two different explanations of self-knowledge – one at the subpersonal level, and the other at the personal level. At the subpersonal level, both self- and other-ascriptions are formed using similar computational processing. At the personal level, the subject on occasion acquires non-inferential self-knowledge of her motivating reasons. On these occasions, she learns that she has a motivating reason using the reasons transparency method (RTM).[4]

There is *one* sense in which my account could be termed a '*dual methods* account.' I take it that subjects can use either inference or RTM to acquire self-knowledge. In the above scenario, I might have instead inferred that 'I believe that *it will rain* for the reason that there are grey clouds' on the basis of evidence such as the fact that I commented on the clouds earlier. But still, both the self-ascriptions formed using inference and those formed using RTM will be underpinned by relevantly similar computational processing. In the thesis, I therefore reserve the term '*dual methods* account' for the view that self-knowledge can be acquired using either one of two methods that are *wholly* distinct from each other. This I reject.

Before discussing each of my two explanations in more detail, I will say three things about the way in which they relate.

First, the personal-level explanation is primary in an important sense because it sets the explanatory 'agenda' for the subpersonal one (Bermúdez: 2005). At the personal level we demarcate the method responsible for distinctive self-knowledge. Once we have done this, we can then pick out the particular mechanisms we are interested in at the subpersonal level, which will include mechanisms involved in deliberation and the generation of the self-ascription itself.

Second, regarding the sorts of features picked out by the two explanations, those appealed to by the subpersonal explanation ground those appealed to at the personal level. That is, the low-level processing makes it the case that the subject herself uses RTM.

Third, the facts appealed to by the two explanations of self-knowledge will be isomorphic to some extent but not entirely. The explanations feature two things that may or may not be isomorphic to

[4] Cassam discusses (without endorsing) a related picture in (Cassam: 2010b).

each other – the contentful mental states and how they relate to each other. For the most part, the content of thought at the personal level will be grounded by type identical content at the subpersonal level. For example, take my conclusion to the world-directed question 'why believe that *it will rain*?': 'a good reason for believing that *it will rain* is that there are grey clouds.' This conclusion will be grounded in the deliberation module by a representation that 'a good reason is that there are grey clouds.' Yet the contents involved will not be fully isomorphic. Some of the personal level contents will be at least partly grounded by a subpersonal indexical content that refers to the person who happens to be me, but it is not self-conscious, and instead refers to 'this system.' The subpersonal indexical content refers to the person who happens to be me as a whole (as opposed to the part that is currently producing the ascription). So, to clarify, 'this system' refers to the system of modules as a whole that underpins the person, as opposed to any one module. For example, while I conclude that 'I believe that *it will rain* for the reason that there are grey clouds,' this will be grounded by the subpersonal representation that 'this system believes that *it will rain* for the reason that there are grey clouds.' Also, although all the contents entertained by the subject will feature somewhere in the subpersonal processing, the inverse will not be the case. That is, there will be contents in the subpersonal explanation that do not feature in the personal level one (e.g., premises such as 'this system takes the grey clouds to be a good reason,' 'this system is free from bias'). The relations between the representations featuring in the personal and subpersonal explanations will not be at all isomorphic, though. Whereas at the personal level the subject deliberates, the module underpinning this does not. And while we might say that the act of deliberating results in warrant, we cannot talk of warrant at the subpersonal level – the output of the mindreading module will not itself be warranted.

6.2.1 Details: The personal level

Regarding the personal level, I endorse the agentialist picture from chapter two whereby subjects learn that they have a motivating reason using the reasons transparency method (RTM). Rational agency grounds the method and other distinctive features of such self-knowledge. Because it has been a while, let me recap the relevant discussion from chapter two.

According to the RTM account, subjects learn the reason for which they have an attitude by answering the question 'why have that attitude?' where this amounts to considering what normative reasons favour having that attitude. Subjects transition from a conclusion about normative reasons to one about their motivating reason. E.g., I learn why I believe that *it will rain* by answering the question 'why believe that *it will rain*?' My answer to the world-directed question – 'a good reason is that there are grey clouds'– provides me with the answer for the inward-directed question. In that way, I can then form the belief, 'I believe that *it will rain* for the

reason that there are grey clouds.' Further, subjects' self-ascriptions are warranted due to the rational agency that subjects exercise in taking a consideration to be a normative reason. More precisely, I take it that a subject's self-ascription is warranted by her agent's awareness of her motivating reason. The subject has that awareness in virtue of an agent's awareness of taking the consideration to be a normative reason.

Further, our rational agency grounds other features of self-knowledge as well. This includes first-person authority, and a particular obligation. We bear the:

> *Knowledgeable reason explanation* (*KRE*) obligation: The obligation to knowledgeably self-ascribe motivating reasons when explaining our own attitude.

Roughly, this means that subjects ought to use RTM to knowledgeably explain their attitudes. I say more in the conclusion about the obligation's grounds.

6.2.2 Details: The subpersonal level

My explanation at the subpersonal level provides an account of the mechanism underpinning self-knowledge (but not how it is warranted). This section sets it out.

At the subpersonal level, I mostly follow Carruthers' account. Doing so brings explanatory benefits, and at any rate, it shows that we can even accept most of Carruthers' picture and *still* reject his ultimate conclusion – that self- and other-knowledge are ultimately the same. Also, following Carruthers' account is the most pragmatic at this point in the thesis: I need to give a concrete picture of the *two explanations* account, and we have already looked at his view in some depth. As such I will take the following to be the case. At the subpersonal level, the process leading to self-ascription is computational and involves transitions between representations. These computations are carried out by a mindreading module – a distinct section of the mind that forms both self- and other-ascriptions. This mechanism is reliable, and we can specify this reliability in a particular way. The output of the mindreading processes that underpin self-knowledge (and indeed the modules' processes in general) reliably correlate with the relevant mental fact. The output will not *always* match up with the relevant fact, though. In these cases, it might be that the module does not process the input representations in a truth-conducive way or gives too much or not enough weight to certain representations.

That said, while Carruthers' picture forms my base, I need to add or clarify four particularly important features of my account.

i. In accepting that the mind is to some extent modular, I nevertheless reject any strong claims about encapsulation. (At its starkest, the 'encapsulation' thesis claims that cognitive modules operate without being influenced by mental states and processes external the given module.)[5] I take it that other cognitive states can penetrate the mindreading module. Specifically, we should think that subjects' desires can shape the mindreading module's computations. After all, as I argued in chapter four, self-ignorance and confabulation are at least sometimes motivated. (Indeed, §6.3 argues that contra Carruthers, a lot of it is.)

To say a bit more, let me recall ALICIA from chapter four:

> ALICIA: Alicia believes that a new colleague Bernice is unpleasant. When asked why, Alicia replies that Bernice did not smile at her in the corridor. Say also that Bernice is black, and that Alicia habitually overlooks black candidates for jobs, and so on. This has been pointed out to her time and time again, but she just shrugs and tries to explain it away. We would say in this case that Alicia was mistaken about why she believes that Bernice is unpleasant. Alicia is ignorant of her racism and instead mistakenly ascribes a motivating reason. Alicia does not actually believe that Bernice is unpleasant for the (supposed) reason that Bernice didn't smile at her. And further, an obvious explanation of this mistake is that Alicia wants to not be racist.

It seems fairly non-controversial that Alicia's desires penetrate the mindreading module to issue in the resulting self-ascription. Her desires would influence the representations that the module uses and the importance the module places on those representations. And indeed, Carruthers himself allows elsewhere that desires can penetrate the mindreading system (2010: 84).

Second, Carruthers does not say anything about the indexical nature of the evidence and resulting ascription, but we should. I want to emphasise that some sort of indexical component is involved, but that, in occurring at the subpersonal level, it will not be first-personal. Rather, the indexical content 'this system' will feature in the resulting ascription and at least some of the evidence. For instance, in my paradigm case, the mechanism transitions from representations such as that '*this system* takes the grey clouds to be a good reason for believing that *it will rain*.' The mechanism will then form the output that '*this system* believes that it will rain for the reason that there are grey clouds.'

Third, my characterisation of the input representations differs from Carruthers'. I allow that facts such as 'this system takes *p* to be a good reason' can constitute the relevant representations. But

[5] On this, see e.g. Carruthers (2006), Fodor (1983), and Prinz (2006).

Carruthers thinks that the mindreading module performs transitions from just sensory evidence to compute the self-ascription, e.g. visual information and inner speech.[6] But, I think various facts will be represented by the mindreading module, and these representations may or may not have been computed using sensory information. Something like the following 'chain' takes place in the mindreading module. The system generates the representation that 'this system takes *p* to be a good reason' by transitioning from the representation that 'this system has inner speech saying that *p* is a good reason.' This representation of inner speech may well be generated by transitioning from sensory information – the inner speech itself. In this case it probably would be formed non-computationally, but I need not commit to this. Also, it is worth noting that I have used doxastic terms in specifying the relevant representations – I have suggested that the system would represent notions such as 'normative reason' and 'belief.' I do this for simplicity but can reject it. Perhaps, like the first-person, these terms cannot be represented by a cognitive module. If this is the case, I would say that the representations pick out these same things – reasons and beliefs – but under other descriptions.

Fourth, my subpersonal explanation appeals to more processes than Carruthers'. He seems to think that the process directly leading to the self-ascription will just consist of computational mindreading. But RTM will also be underpinned by processes in other parts of the mind. Namely, the deliberation involved in RTM in considering the world-directed question will be underpinned by the subpersonal processes grounding deliberation.

Before continuing, I should note that while my characterisation of the subpersonal explanation has assumed computationalism about the mind, I can be flexible. My picture is also compatible with accepting a connectionist picture at some level of explanation. According to connectionism, we can best model at least some parts of our mental life in terms of patterns of activation connecting inputs to outputs, rather than as transitions mediated by structured contents.[7] But even if we accept this, we can still say that at a different level, cognition can be understood computationally and that self-knowledge results from computational processing.[8] Further, the basic 'two explanations' approach to self-knowledge is also compatible with denying computationalism across the board. I would only need to insist that relevantly similar processes underpin both self- and other-knowledge at the subpersonal level, and that desires can penetrate these processes.

---

[6] On the importance of inner speech, see Carruthers (2010).
[7] See Bermúdez (2005: ch.5) for an overview.
[8] One might think that the activation patterns are implemented by transitions between representations, and that, at one level, we can also usefully talk of computations, even if the connectionist model is somehow deeper. See Rescorla (2017: §4.1).

***

I have provided a lot of details here, so I will now take stock. We should explain self-knowledge of why we have our attitudes in two ways. According to my personal-level explanation, subjects use RTM (the reasons transparency method) to learn of their motivating reason for having attitude *A*: subjects consider what normative reasons there are for having *A*. In concluding that, say, *p* is a (normative) reason, subjects can come to know that *p* is their (motivating) reason. Both the way subjects form the self-ascription, and the way in which it is warranted, is non-inferential. And we can also offer a subpersonal level explanation of the self-ascription. The subject's use of RTM is underpinned by processing in parts of the brain responsible for deliberation and forming both self- and other-ascriptions (i.e. mindreading). The mindreading module transitions from, say, the sensory information involved in inner speech to form the representation that 'this system has *A* for the reason that *p*.' This representation grounds the self-ascription – 'I have *A* for the reason that *p*.'

The subpersonal and personal level explanations are compatible but different, like those we can give of perception. In virtue of the personal level explanation, self-knowledge under this picture differs significantly from other-knowledge. I will say more about how this is the case in the next chapter and the conclusion. But, already we can note that there is a distinct method and warrant by which subjects acquire self-knowledge, grounded by rational agency. Further, the subject possesses first-person authority and is subject to the KRE obligation. Although self- and other-knowledge resemble each other in some ways, they nevertheless fundamentally differ in others.

To take stock, let's use a table again to spell out how self-knowledge is distinctive under my account. We can start filling in some of the squares. I will further discuss some of the others later.

	Two explanations approach: RTM underpinned by computationalism
Extra-reliable	No
Self-intimation	Yes (See next chapter)
Distinctive method and warrant	Yes (at the personal level)
First-person authority	Yes (Grounded in various features. See conclusion)
Role for rational agency	Yes

6.3 Why accept my account?

Now that I have set out my view, I will clarify why we should accept it; hopefully it is starting to become clear that this approach allows us to have the best of both worlds. The *two explanations* approach overcomes all the negatives we encountered regarding the Orthodoxy while keeping the positives. This section returns to the criticisms of the Orthodoxy from chapter four. I will first consider the general plausibility of my *two explanations* approach in contrast to the Orthodoxy (§6.3.1), before then discussing my account's explanatory power. I will argue that my account inherits most of the explanatory advantages of Carruthers' account (§6.3.2) and also adds a new one (§6.3.3).

6.3.1 How I avoid the general problems with the Orthodoxy

Recall the general worries with the Orthodoxy that I presented in chapter four. Inferentialism or computationalism (or at least computationalism on its own) provide an unappealing account of the method involved. They conflict with the dual role of the question 'why?', and the direct rational relations between explaining one's attitude and the attitude in question. And inferentialism/computationalism also fails to satisfactorily explain the warrant for self-knowledge. I'll now consider these issues in turn and how my account avoids these problems.

First, recall that the question 'why?' has a dual role – asking one to both explain and justify the attitude in question. Using inference or computation alone to answer the question is not to take the question seriously. After all, the mindreading system need not take into account whether *p* is a good reason. Indeed, the deliberation system might even represent that '*p* is a good reason' while the subject herself is indifferent. But RTM accounts for the dual role. We explain our attitudes by considering what justifies them and taking something as a good reason. Evidential processing underpins this process, but at the subpersonal level. As such, the subject herself doesn't rely on evidence that she has the reason. In this way, my two explanations account featuring RTM is better than just appealing to computationalism on its own.

Second, I argued that a subject's explanation of her attitude bears direct rational relations to the attitude itself. The irrationality of Moore paradoxical statements illustrates this, e.g., 'I believe that *q* for the reason that *p*, but *p* is a bad reason for believing that *q.*' This combination of beliefs renders both the self-ascription and the belief that *p* in some way irrational from the subject's perspective. The RTM account can explain this direct relation. When a subject employs RTM, the rationality of both her belief that *q* and her belief *that I believe that q for the reason that p* is at stake to her. This is because when the subject employs RTM, she considers what is a normative reason for believing that *q.* So, if the subject is seriously considering her motivating reason as a

normative reason for belief, then she will not conclude that 'I believe that *q* for the reason that *p,* but *p* is a bad reason.' To do otherwise breaches the norms of the process she is engaged in. Again, this is all compatible with thinking that self-knowledge at the subpersonal level is computational.

It should also be clear that we can now explain subjects' warrant for their self-ascriptions. Since we are now positing a personal-level method, we can appeal to something other than reliabilism. As mentioned earlier, I think that knowledge acquired using RTM will be warranted by an agent's awareness of one's motivating reason. Obviously, there is more to do in spelling out such an account of the warrant, but we now have the resources to do so at the personal level.

6.3.2 Explanatory advantages: My account does everything Carruthers' does

We should also accept my *two explanations* account on the grounds that it explains self-ignorance and error. In helping itself to Carruthers' picture, my approach inherits most of its explanatory advantages. (I discuss my slight hedge in a moment.) We can overcome the explanatory problem we encountered in chapter three: avoiding the TWO METHODS PROBLEM and explaining confabulation. Recall that subjects use computation/inference in confabulation cases. To fully explain such cases, the *dual method* theorist must say when subjects do and do not use this other method rather than the distinctive method.

The *two explanations* account sidesteps this problem by appealing to one mechanism that always operates. Subjects' self-ascriptions are underpinned by computation in both confabulation and knowledgeable cases. But subjects also use RTM in some knowledgeable cases, and indeed, some confabulatory ones as well. Any motivational factors will operate by penetrating the mindreading module and influencing its processing.

I should acknowledge that my account does not quite inherit one virtue of the Orthodoxy's explanation. Carruthers prided himself on his explanation's simplicity. In incorporating the agentialist picture, I have added complexity since my account appeals to an additional relevant factor – a method the subject uses. But this is not problematic since my addition's advantages outweigh the extra complexity. Carruthers' account on its own is highly implausible: it conflicts with our intuitions about the method and warrant for self-knowledge. In comparison, my overall picture of self-knowledge is generally more plausible and, indeed, as I will discuss, brings an additional explanatory advantage.

We have, then, various ways in which my account preserves the benefits of computationalism. As such, we have reason why the agentialist should accept computationalism as well as their own

picture – the explanatory benefits this move brings. Indeed, this reason comes in addition to another, obvious one. Agentialism plainly pertains just to the personal level of explanation but the processes appealed to by agentialism can't float free. Plausibly, processes appealed to at the subpersonal level of explanation ground those at the personal level in some way. At least some relation holds, for example, between our pains and c-fibres firing. It would be highly implausible to say that the sorts of processes the agentialist appeals to – deliberation and the transparency method – are not underpinned by low-level processing. As such, agentialists have yet more reason to accept something like the account I propose.

6.3.3 Explanatory advantage: My account does everything Carruthers' does, *and more*

As well as sharing the explanatory pay-off of Carruthers' account, my *two explanations* approach has a further explanatory advantage. It provides a good explanation of confabulation that accounts for the Confabulation Asymmetry while requiring only one independently plausible assumption. Recall from chapter three that Carruthers does not give a good account of the Confabulation Asymmetry. In §6.3.3.1 I propose an explanation of confabulation (and the Confabulation Asymmetry) that falls out of my account. Then, in §6.4.3.2, I argue that it is not just *an* explanation, but specifically a *good* explanation. As such, we have more reason to accept my dual explanation approach and RTM.

6.3.3.1 The Confabulation Asymmetry and my proposal

Chapter four introduced an explanandum concerning confabulation cases:

> *Confabulation Asymmetry*: We tend to mistakenly ascribe motivating reasons to ourselves more readily than to others.

My account of self-knowledge gives rise to an explanation that accounts for this. I will set out the explanation roughly at first and then in more detail, before clarifying at the end of this subsection how it meets the explanandum.

Recall that part of my agentialist picture is that we bear the *knowledgeable reason explanation* obligation:

> *Knowledgeable reason explanation* (*KRE*) obligation: The obligation to knowledgeably self-ascribe motivating reasons when explaining one's own attitude.

I will use the KRE obligation in the following proposal:

> *We confabulate, and indeed confabulate with the content we do, because we desire to have fulfilled the KRE obligation (i.e., the obligation to knowledgeably explain our attitudes by reference to motivating reasons).*[9] *These personal-level desires influence the subpersonal processes underpinning the self-ascription.*

According to my proposal, subjects confabulate when they lack an accessible explanation that would enable them to fulfil the obligation, i.e., when the subject lacks a motivating reason. Carruthers could also accept the KRE obligation and provide this explanation too, but I take it to be an advantage of my account of self-knowledge that the explanation falls out of it so easily. After all, the KRE obligation forms a key component of my account of self-knowledge.

Let us consider the proposal in the context of an example and return to the stockings experiment. First things first, this seems to be the sort of situation in which subjects bear the undefeated KRE obligation, at least from their perspective.[10] We can further take it that the individuals desire to have met this obligation (I discuss this assumption in §6.3.3.2). The desire to have fulfilled the obligation leads the subjects to confabulate an answer in the absence of a true one they can provide – they did not form the preference on the basis of reasons. And further, the subjects specifically self-ascribe the reason that the stockings were sheerer, say, because it is a plausible motivating reason. The subject may well use RTM in this instance. At the personal level, they might consider 'why prefer that pair of stockings?' and conclude 'because they are sheerer.' But this isn't the full picture. The subject's use of RTM and all the computational processes underpinning the self-ascription will be influenced by the subject's desire.

With the rough picture in hand, we can now flesh it out with a mechanism by which the motivational factor operates. As discussed in chapter three, a good option is to see motivated confabulation as an instance of self-deception, and specifically self-deception construed along Alfred Mele's (2001) lines. Recall Mele argues that, in self-deception, motivational factors influence belief-formation. The desires lead subjects (or their belief-formation mechanisms) to place too much weight on certain pieces of evidence, ignore other pieces of evidence, and so on. As a result, the subject acquires a self-deceived belief.

---

[9] See Sullivan-Bissett (2015: 552) on these two ways in which motivational factors can influence confabulation.

The desire to have fulfilled the KRE obligation is perhaps a more specific version of Velleman's (1985) 'desire for self-understanding.'

[10] Sometimes the obligation will be undefeated in confabulation cases, but sometimes it might not. My explanation only requires, though, that the circumstances in these instances are sufficiently like those in normal cases that subjects would plausibly believe they bear the obligation.

Further, the fact that the studies concern cases where subjects would plausibly think that they bear the KRE obligation means that Carruthers cannot make a criticism like in (2013: 342).

If we see confabulation as an instance of self-deception construed along Mele's lines, we could understand the mechanism underpinning it more precisely, in the following way. Unbeknownst to the subject, her mindreading module processes the available representations in line with the subject's desire to have fulfilled the *KRE* obligation. These desires lead the mechanism to give rise to a self-ascription which the subjects can provide to the questioner as opposed to admitting their ignorance. These desires also mean the ascription features the specific content that it does. E.g., the subject's desire to have met the obligation leads the mindreading module to place too much weight on the evidence, if there is any, that the subject formed the attitude in question on the basis of a reason. This may involve placing weight on:

- What the subject takes to be a plausible normative reason. In using this piece of evidence, the module would rely on the theory that: if S takes *p* to be a normative reason for preferring *x*, *p* is S' motivating reason for preferring *x*.

- The fact that a subject's attitude can be based on a given reason without resulting from explicit deliberation.

And the subject's desire also causes the module to under-value or even ignore the evidence that the subject did not form their preference on the basis of a reason. This might include the following facts:

- That the subject experienced uncertainty when considering potential normative reasons.

- That the subject cannot remember forming the preference on the basis of the reason in question, or even considering said reason at all.

- That other subjects with the same attitude lack a motivating reason. When coupled with the claim that our minds work in similar ways to other people's, this might suggest that we lack one as well. For example, recall the subjects in the Pronin et al. (2002) study. The subjects' mindreading mechanisms may well have ignored this fact since the subjects' self- and other-attributions clearly differ.

Because the mindreading module processes evidence in this way, the subject's desire to have fulfilled the obligation would lead them to adopt the relevant self-deceptive belief which they then express to the questioner. It may well be that the subject uses RTM, but my account explains why RTM issues in a false self-ascription. RTM is underpinned by computational processing in the mindreading module. The subject's desire to have fulfilled the KRE obligation penetrates the mindreading module and affects its processing. As a result, the module disregards facts that indicate that the subject lacks a motivating reason. It also places too much weight on what the

subject takes to be a normative reason. The module therefore issues in a false ascription that 'this system' has a given motivating reason. This self-ascription underpins the subject's own false self-ascription that she has a given motivating reason.

It should now be clear that my proposal explains the Confabulation Asymmetry. My proposal states that we desire to have fulfilled the obligation to knowledgeably explain *our own* attitudes by reference to motivating reasons, not other people's attitudes.

6.3.3.2 The ways in which the proposal provides a good explanation

I now give three reasons why the proposal is a good explanation.

i. The auxiliary principle is independently plausible

My *two explanations* account gives rise to the proposal provided an additional principle. This is that we don't just bear the obligations; we additionally desire to have fulfilled them. But this principle is independently plausible and not *ad hoc*.

I can firstly note that I need not commit to anything very demanding regarding the desire in question. The desire could be as minimal as a tendency. Additionally, it is plausible that we would bear such a state regarding the *KRE* obligation and that it would play the role I am arguing it does. After all, doing as we ought (generally) reflects well on us and it is reasonably uncontroversial to think that we generally tend to see ourselves in a favourable light. For example, Wilson puts the point as follows:

> People's judgements and interpretations are often guided by […] the desire to view the world in the way that gives them the most pleasure – what can be called the 'feel-good' criterion. […] Just as we possess a potent physical immune system that protects us from threats to our physical well-being, so do we possess a potent psychological immune system that protects us from threats to our psychological wellbeing. When it comes to maintaining a sense of well-being, each of us is the ultimate spin doctor (2002: 38).[11]

I can draw on a range of data when saying that this wish to feel good manifests in positive self-appraisals. This includes some of the empirical support for self-deception to the extent in which it concerns self-deception about ourselves (e.g., Mele 2001: 3, 11). And I can also refer to various cognitive biases that have a similar effect. For example, the study concerning the bias blind spot that I outlined in chapter four showed our blindness to the consequences of the 'self-serving bias,'

[11] See also Gilbert and Wilson (2000).

in which people chalk their achievements down to themselves, but failures to other influences (Pronin et al., 2002: 370, 377).[12] Indeed, it is a sign that something has gone wrong with the subject if they fail to view themselves and their circumstances through slightly rose-tinted spectacles, as we see with so-called 'depressive realism.'[13]

Let me end this subsection by assuaging several worries we might have about attributing this sort of desire to the population at large. We might object that I over-intellectualise matters by saying that subjects desire to have fulfilled the *KRE* obligation, or indeed that they believe that they bear it. After all, the average person probably has not thought about these issues. But, I just want to suggest we have some sort of standing state that can be most simply captured in terms of the obligation. Indeed, introducing talk of the '*knowledgeable reason explanation* obligation' need not be ad hoc – take how we discuss moral reasoning. If a subject faced with the trolley problem says they would kill one person to save five, we might say they believe that one ought to bring about the greatest happiness for the greatest number. Indeed, based on more responses, we may even want to attribute a very fine-grained utilitarian principle to them, e.g., concerning interests or preferences. But this is not to say that the subject thinks in those terms or are in a position to explicitly provide such a principle. Second, there may also be individuals who do not even believe they are subject to the KRE obligation in some undemanding *de re* sense. For example, certain philosophers will deny that we bear the obligation. As such, these individuals would not desire to have fulfilled the KRE obligation. Yet, I need not say that everyone has this belief and desire, just that individuals will confabulate to the extent that they do. There is space here for empirical research.

ii. Avoids worries with other motivational accounts

My proposal avoids Carruthers' problem with the pragmatic pressures option. Recall that one popular account of confabulation says that it is motivated by a desire to fulfil the demands of interpersonal communication. Yet Carruthers observed that subjects confabulate even when no one is paying attention. But according to the proposal, it is not that we want to provide reasons so others think well of us. Rather, we want to have fulfilled the obligation to explain our attitudes by reference to motivating reasons. This is the case even if we are just explaining the attitudes to ourselves.

iii Better than related alternatives

[12] See also Coleman (2015: "self-serving bias") and Turner and Hewstone (2009). Other biases include the *positivity bias*, *unrealistic optimism*, and the *Lake Wobegon effect*; see Coleman (2015).
[13] E.g. Brown (2007), although see Moore and Fresco (2012) for caution in the precise details of the theory.

My explanation – we confabulate because we desire to have fulfilled the *KRE* obligation – is better than (at least) two related alternatives.

First, my explanation uses the *KRE* obligation as opposed to the related but less controversial obligation to form our attitudes on the basis of what we take to be normative reasons. This reference to responsible attitude-formation at the lower-order level is the sort of thing that Pronin et al., for example, seem to have in mind when discussing the possibility that the 'biased [cognitive] searches' we engage in due to the bias blind spot 'may blind us to our shortcomings and enhance our sense of rationality in a way that is undeniably ego enhancing' (2004: 788). We might simply say that subjects are motivated by the desire to have fulfilled the obligation to form their attitudes on the basis of motivating reasons. The KRE obligation does presuppose this more minimal one. Yet, appealing to just the obligation to form attitudes on the basis of reasons only explains the content of the subject's confabulations. It does not explain why they confabulate in the first place – it is unclear why, under such a model, the subjects do not simply admit their ignorance. After all, to say that you do not know why you have an attitude it is not in itself to say anything about why you do actually have it. It may be possible to have a motivating reason you are not aware of. Appealing to the *KRE* obligation, though, helps us explain both aspects of confabulation.

Second, the proposal states that we are motivated by the desire to have fulfilled an *obligation* to knowledgeably explain one's attitude with motivating reasons. An alternative explanation would simply be that we desire to have to have knowledgeably ascribed motivating reasons without there being any normative demands to have done so. But appealing to the KRE obligation provides a good, full and simple explanation, and is independently plausible (recall my arguments for the KRE obligation in chapter two). On the other hand, if one's explanation of confabulation simply references a desire to have knowledgeably ascribed motivating reasons, one still needs to say why subjects have this desire. Since it cannot be to impress others, perhaps we might say that self-knowledge is some sort of non-normative 'epistemic desideratum.' But this is not to say why we would desire a particular type of self-knowledge – why self-knowledge of motivating reasons as opposed to purely causal explanatory ones? It is not obvious how we would cash out a desire for self-knowledge in non-normative terms. Perhaps we might say that self-knowledge is valuable to us because of pragmatic considerations. It helps us assess our attitudes and come to better decisions. And yet knowing the truth – that we lack motivating reasons – would also be useful. Why, then, would subjects overlook the signs that they lack the relevant reasons in confabulation

cases? It starts to look, then, that even if we could make the relevant manoeuvres, they might be on the baroque side, and involve sacrificing simplicity.[14]

***

My *two explanations* account combining computationalism and RTM engenders a good explanation of confabulation that accounts for the Confabulation Asymmetry. As such, we have more reason to accept my account.

6.4 Objections

Having set out my full account and why we should accept it, I now address two sets of worries. We might object that RTM and computationalism are actually mutually exclusive (§6.4.1). We might also insist, my previous observations from chapter two notwithstanding, that confabulation cases render RTM too unreliable for knowledge (§6.4.2).

6.4.1 The two explanations are not compatible

I have claimed that under my *two explanations* account, self-knowledge fundamentally differs from other-knowledge even though both are computational at the subpersonal level. This account captures the best parts of agentialism and the Orthodoxy, thus allowing us to have the best of both worlds.

Yet one might deny that we can do this. Carruthers argues that the distinctive access theorist cannot make this sort of move: accepting his account of self-knowledge while maintaining their own on the grounds that the accounts occupy 'different explanatory spaces' (2011: 21). Carruthers' opponents might, he considers, offer something like my approach here and draw parallels with perception:

> Philosophers who maintain that we have direct perceptual access to the world don't mean to be denying what the cognitive scientists assert [that 'visual processing is heavily inferential in nature'].[15] Rather, they mean only to be emphasising that,

[14] Nevertheless, while I think we should not, I would be happy if one accepts this explanation. Still, it is something that my agentialist account would predict, and thus the explanation still lends agentialism support. Also, it would still be significant for the confabulation literature. I would still have argued for a motivational account of confabulation, and that the relevant motivation concerns self-knowledge.
[15] Here I would refer to 'computation' in this context, not 'inference'.

> phenomenologically, it is the world that is presented to us in perception, not some intermediate entity like a sense datum. And similarly, the worldly contents of our perceptions are thought to justify our corresponding beliefs immediately, without us needing to engage in an inference or rely upon any major premise about the general reliability of experience. Likewise, it might be said, for the claim that we have direct access to our own propositional attitudes. Perhaps this is only supposed to rule out *conscious* forms of self-interpretation, and is hence consistent with [computationalism]' (2013: 22).

Carruthers offers several arguments for rejecting a *two explanations* approach like this; I will consider the two most applicable to RTM:

i. While self-knowledge is indeed non-inferential at the personal level of explanation, other-knowledge can be as well. Therefore, if we still want to say that self- and other-knowledge differ, we must draw this disparity at the subpersonal level. The distinctive access theorist, then, cannot accept that both self- and other-knowledge resemble each other at this level, i.e., that both are computational (2013: 22).

REPLY: Even if we think both self- and other-knowledge are non-inferential at the personal level, they still significantly differ. Subjects only use RTM to learn of *their own* motivating reasons, not other people's. If other-knowledge is non-inferential at the personal level, it will be because subjects, say, directly perceive others' reasons in some way.

ii. RTM claims that subjects use different methods to learn of their and others' minds at the personal level. Self-knowledge must also differ at the subpersonal level of explanation to ground this difference. Proponents of RTM therefore cannot accept Carruthers' account whereby self-knowledge is underpinned by the same type of subpersonal processes as other-knowledge (2013: 23-4).

REPLY: I agree that the distinctive access theorist cannot explain self-knowledge at the subpersonal level *just* by reference to computation in the mindreading module. But we can appeal to other subpersonal processes to ground the distinctive method. For example, RTM is grounded by computations involved in deliberation and attitude-formation as well as those in the mindreading module.[16]

***

[16] Carruthers also denies that we can draw the relevant lessons from philosophy of perception in (2010: 93). Here, his response seems to involve taking his account to be personal-level in an important way. But I endorse a computational account of self-knowledge firmly at the subpersonal level, even if Carruthers himself doesn't.

We can therefore adopt a *two explanations* approach to self-knowledge and maintain distinctive access under such a model.

6.4.2 Confabulation renders RTM insufficiently reliable for knowledge

Another objection uses confabulation cases to argue that RTM is insufficiently reliable to deliver knowledge. As mentioned before, we might plausibly say that subjects use RTM in confabulation cases. After all, the subject in the stockings experiment may well consider what are the good reasons for preferring the chosen stockings. RTM issues in an incorrect explanation because the subject's desire leads the underlying mindreading module to overlook facts that indicate that the subject doesn't prefer the stockings for the reason that they're sheer. It looks, then, that using RTM will result in a lot of 'false positives' whereby the subject self-ascribes a motivating reason she lacks. As such, we might doubt that RTM could *ever* provide the subject with knowledge, even when their self-ascription is true. Since RTM is intended as an account of self-knowledge, we would have reason to reject it.

We should not confuse this criticism with a related one. Recall from chapter three that proponents of the Orthodoxy use confabulation cases to deny that self-ascriptions are more reliable than other-ascriptions. I accepted that self-knowledge is overall no more reliable than other-knowledge but maintained that self-knowledge can be distinctive in virtue of other features. But the issue at stake here is not whether RTM would be especially reliable, but whether it would even be reliable enough to issue knowledge. After all, while I don't think reliable formation suffices for knowledge, it is still plausibly necessary.

I maintain that RTM will be sufficiently reliable for the true self-ascriptions it issues to be knowledgeable. This is for three reasons.

i. Subjects will generally only confabulate when they lack a motivating reason for their attitude, but as a general rule subjects do base their attitudes on reasons. Here I want to say two things.

First, I should emphasise that motivating reasons can be very general. For example, a subject might base their desire to see a film simply on the fact that it would be fun, and not any particular features that make the film entertaining. Even if the subject doesn't base her desire on a very specific consideration, she will still have a motivating reason in this instance.

Second, subjects will often have a motivating reason even if other factors also explain why they have that reason. For example, say I believe that *Bowie dislikes me* for the reason that he left my lap (in this scenario, 'Bowie' is a small ginger cat). Perhaps I only take that to be a reason because I'm generally suspicious and prone to think that everyone dislikes me. But still, Bowie's departure

is my motivating reason for the belief. If he hadn't left my lap, I wouldn't have formed that belief (at least not then).[17] So, when I believe that *I believe that Bowie dislikes me for the reason that he left my lap*, I still correctly ascribe a motivating reason.

ii. While the mindreading module sometimes 'overlooks' evidence that we lack a motivating reason, it does not always do so. There are occasions when subjects instead provide a purely causal explanation or admit that they do not know why they hold an attitude. That is, the subjects may well switch personal-level methods, and use inference instead of RTM.

After all, there are various cases in which subjects do not confabulate, but instead provide a correct purely causal explanation of their attitudes. For example,

> DIANE: Diane tells you she dislikes a new colleague and you ask her why. She replies sheepishly that 'I'm probably just tired – I'm very judgemental when I haven't slept much.'

Indeed, individuals sometimes get it right in experimental settings, too. For example, we see this in Haidt's (2001) study which I mentioned earlier. In the study, experimenters asked subjects why they thought incestuous siblings in a given scenario did something wrong. While the subjects initially confabulated, they eventually admitted that they didn't know why they made the judgement that they did. Also, recall the bias blind spot experiments. Pronin and Kugler (2007: 574-5) performed a related study in which they explicitly informed subjects that individuals are often mistaken about what causes their attitudes. Upon testing, these subjects did not exhibit the bias blind spot. So, it seems that there will be occasions in which relevant evidence will be too salient for the mindreading module to ignore.[18] As such, the subject will not confabulate in all the cases in which they lack a motivating reason but will use personal-level inference instead.

iii. If we confabulated to such an extent that RTM was insufficiently reliable to issue knowledge, then subjects' self-ascriptions would be too unreliable to constitute self-knowledge at all. This is because inferentialism/computationalism would also be too unreliable for knowledge. The Orthodoxy claims that one method underpins mistaken and confabulated ascriptions alike. Given that the Orthodoxy and agentialism are the only views in town, the objector would have to think that subjects *never* knowledgeably self-ascribe motivating reasons. And yet, this does not seem to be the case. We have, then, more reason to suppose that subjects do not confabulate enough to render RTM too unreliable for knowledge.

---

[17] I further discuss the nature of motivating reasons in the next chapter.

[18] Gawronski et al. (2006) also cite a number of striking cases in which they got subjects to correctly ascribe implicit attitudes they might otherwise not have.

6.5 Conclusion

We should accept that self-knowledge is computational and make use of the explanatory benefits this move brings, but nevertheless think that self-knowledge is distinctive. I have proposed a model that gives us the best of both worlds: the *two explanations* account. Both self- and other-knowledge are computational at the subpersonal level. But despite this commonality, self- and other-knowledge still fundamentally differ. At the personal level, self-knowledge is grounded in rational agency. Subjects can use RTM to learn that they have a motivating reason. In such cases, the subject's self-ascription is warranted in virtue of an agent's awareness of the motivating reason. Further, subjects have first-person authority regarding their motivating reasons, and bear the KRE obligation. The next chapter further considers the way in which self-knowledge of motivating reasons is distinctive. Building on the foregoing, I argue that our motivating reasons self-intimate.

Chapter 7 Motivating Reasons as Strongly Self-Intimating

So far, I have argued for a *two explanations* account of self-knowledge of motivating reasons. At the subpersonal level, self-knowledge resembles other-knowledge; the mindreading module forms both self- and other-ascriptions. At the personal level, though, we should endorse an agentialist picture. According to this, subjects use the reasons transparency method (RTM) to learn what their motivating reasons are. Subjects learn the reason for which they hold an attitude or perform an action by considering the world-directed question 'why hold that attitude/perform that action?' The subject's answer to the world-directed question then provides them with the answer to the question concerning what their motivating reason is. I.e., a subject can (non-inferentially) transition from her judgement about normative reasons to one self-ascribing her motivating reason.

I introduced my *two explanations* account as a way of arguing that we have distinctive self-knowledge of why we hold our attitudes and perform actions. But we should go even further when arguing that self-knowledge of motivating reasons is distinctive. This chapter argues that motivating reasons strongly self-intimate. Necessarily, if a subject has a motivating reason, provided they possess the relevant concepts, the subject will be in a position to learn that they have that motivating reason. (In setting out my position, I will for brevity's sake sometimes omit the requirement that subjects possess the relevant concepts.) Indeed, this claim has a surprising upshot. Self-knowledge of reasons possesses an epistemic privilege that self-knowledge of attitudes lack. As I will discuss, our attitudes do not strongly self-intimate.

The chapter proceeds as follows. §7.1 recaps self-intimation (we last encountered the notion in chapter one). §7.2 considers Cassam's argument that our attitudes do not strongly self-intimate. But, I will go on to argue that motivating reasons avoid his worry. And having this other mental feature in mind will help us grasp what it is about our reasons that makes it plausible they are self-intimating. The following two sections make the positive case for thinking that motivating reasons strongly self-intimate. §7.3 offers a quick argument with a bold conclusion: we should accept *simpliciter* that motivating reasons self-intimate. §7.4 provides a slower argument for a more modest claim: if RTM holds, then our motivating reasons self-intimate. Hopefully the thesis so far has already convinced one of RTM. Therefore, I present the main argument also as a way of arguing that our reasons self-intimate. §7.5 considers two objections to the claim that motivating reasons self-intimate. The first objection raises possible counter examples. The second revisits Cassam's argument from §7.2, this time in relation to motivating reasons. §7.6 emphasises an

upshot of this chapter. Our motivating reasons strongly self-intimate even if we doubt for principled reasons that our attitudes do. So, self-knowledge of reasons does not just differ from other-knowledge; it also bears an epistemic advantage that self-knowledge of attitudes lacks.

7.1 Self-intimation

Recall the phenomenon of self-intimation – that a constitutive relation holds between certain conditions and knowing that one is in them. Self-intimation comes in different strengths (these distinctions are delineated in Cassam 2014: Ch. 11).[1] In its *extreme* iteration, a self-intimation thesis claims that necessarily, if a subject bears a particular feature, she will know that bears it. In contrast, what I call *strong self-intimation* claims that there are features that a subject cannot fail to *be in a position* to know that she bears. And *weak self-intimation* claims that there are features that a subject cannot *rationally* fail to be in a position to know that she bears. I will specifically argue that our motivating reasons *strongly* self-intimate.

Let me clarify two things.

i. I take 'being in a position to know' as being placed such that all the subject needs to do to acquire knowledge is employ the relevant method of belief-formation in a responsible way. That is, the subject has all the epistemological assets she would need for knowledge: access to the relevant grounds (even if she does not currently have them), an absence of misleading defeaters, the availability of a reliable method, etc.[2]

I should note that my opponent, Cassam, seems to use a different construal of 'being in a position to know,' but mine is preferable. He seems to see it simply in terms of an absence of obstacles such as one's 'embarrassment or despair' (*Ibid.* 192). Williamson nicely sets out this sort of account of being in a position to know:

> To be in a position to know *p*, it is neither necessary to know *p* nor sufficient to be physically and psychologically capable of knowing *p*. No obstacle must block one's path to knowing *p*. If one is in a position to know *p*, and one has done what one is in a position to do to decide whether *p* is true, then one does know *p* (2002: 59).

[1] Helpfully, see also Williamson (2002).
[2] As such, I'm understanding being in a position to know more broadly than Boyle. He writes just that being in such a position means that one 'needs no further grounds in order knowledgeably to judge' that *p* Boyle (2011: 8).

But requiring the absence of obstacles in a general sense is too restrictive. Some obstacles only block the subject from knowing a fact in that the obstacle blocks belief formation. The obstacle doesn't seem to preclude one from being in a position to know *per se*, just form the belief.

To illustrate, I take it that the subject is in a position to know his belief in the following example. As I understand things, one is not in a position to know a belief that *q* using the transparency procedure when a psychological obstacle prevents one from judging that *q*. But one would be in a position to know the belief if one judged that *q* but an obstacle stopped one from self-ascribing the belief. Say Charlie judges that *horoscopes are nonsense* but does not self-ascribe the belief that *horoscopes are nonsense*. Charlie grew up in a superstitious family and desires to fit in with them. He therefore engages in self-deception and resists self-ascribing the belief that *horoscopes are nonsense*. Perhaps his desire would count as an obstacle to self-knowledge in some sense. But still, Charlie has all the grounds and other epistemic assets he needs to knowledgably self-ascribe the belief if he wished. And indeed, he just seems not to have 'done what [he] is in a position to do to decide whether *p* is true.' After all, if he had used TM responsibly, then he would reach the correct answer. It looks like the desire is only really an obstacle to forming the belief, not to knowledge *per se*. And in that case, we should think that Charlie is in a position to know his belief; the obstacle construal is too restrictive.

I think this seems intuitive. But even if one insists on the alternative understanding of being 'in a position to know,' the epistemic position I have in mind still bears significance. I nevertheless will have shown something interesting about self-knowledge of motivating reasons in arguing for it.

ii. I am interested in being in a position to know that one has a feature when one in fact has it. I.e., I argue that necessarily, when a subject has a motivating reason, she is in a position to know that she has it. I do not think that lacking a motivating reason self-intimates. Subjects can fail to have a motivating reason without being in a position to know this and may falsely self-ascribe a reason instead.

7.2 Against self-intimation of attitudes

In §7.3 and §7.4 I will argue that motivating reasons self-intimate, but before doing so, I will raise a problem for thinking the same about our attitudes. I will return to this objection in §7.5, where I

will show that it does not threaten my picture concerning motivating reasons. As such, my argument that motivating reasons in particular self-intimate will be strengthened.[3]

Cassam denies that our attitudes self-intimate and tells us that:

> Throughout this book [*Self-Knowledge for Humans*] I've operated with a dispositionalist account of belief and other attitudes: to believe that P is to be disposed to think that P, to act as if P is true, to use P as a premise in reasoning, and so on. Merely *having* the dispositions associated with believing that P is no guarantee that you know or believe that you have them, just as believing that you have the relevant dispositions is no guarantee that you have them. Neither ignorance nor error is ruled out, and self-ignorance is possible even if the dispositions you need in order to count as believing that P include the disposition to self-ascribe the belief that P. If you believe that P, and the question arises whether you believe that P, then other things being equal you will judge that you believe that P but it doesn't follow that you believe that you believe that P prior to the question arising. Suppose you believe that the government will be re-elected. The thought that this is what you believe might never have crossed your mind, and if it did cross your mind you might find it hard to admit to yourself. Yet your other dispositions might leave no room for doubt that this is what you believe (2014: 197-8).

This specifically attacks *extreme* self-intimation: necessarily, if a subject believes that *q*, she will know that she believes that *q*. The argument also applies to *strong* self-intimation: necessarily, if a subject believes that *q,* she will be in a position to know that she believes that *q*. We can put the argument as follows: Beliefs are broadly dispositional. There is only one way in which a dispositional state could be self-intimating – if its self-ascription was one of the manifestations. But one's dispositions can be blocked from manifesting. So, a subject can have a belief which consists of the disposition to self-ascribe the belief, but where the subject cannot do so.[4]

This seems right. Certainly, the argument applies to agentialist accounts of how we know our attitudes. This is because the disposition to self-ascribe the belief that *q* using the transparency method (TM) rests on being disposed to judge that *q*. Insofar as we take beliefs to have a dispositional component, which is indeed plausible, one can believe that *q* where the disposition to judge that *q* is blocked. E.g., take the academic in Peacocke (1998: 90). Let's call her Christine.

[3] Also, Williamson (2002) notably denies that sensations self-intimate. I will not discuss Williamson's argument here since I am not concerned with self-knowledge of sensations. But, briefly, I can note that his argument doesn't extend to motivating reasons since it relies on the fact that sensations are gradable in a way that believing for a reason is not.

[4] See also Parrott (2017).Carruthers says something related concerning judgement (2013: 101-2, 104).

Christine judges that *degrees from different universities are equally valuable*. Yet she tends to prefer job candidates who studied in her own country, and so on. We can plausibly say that Christine in fact *believes that some degrees are more valuable than others*, even though she is not prepared to judge that *some degrees are more valuable than others*. Perhaps she has the disposition to make that judgement, but it is systematically prevented from manifesting – she judges that the degrees are equal. In this case, the subject will not be in a position to learn of her belief that *q* by considering whether *q* is true.[5] Christine might employ the transparency method and carefully consider the outward-directed question 'are all degrees equal?' with full conceptual grasp, and still fail to self-ascribe her belief. So, it isn't the case that necessarily, if S believes that *q*, she will be in a position to know that she believes that *q* using TM.

As a result, then, we should reject strong self-intimation concerning belief. It still allows for weak self-intimation – that subjects are necessarily in a position to learn of their beliefs *if they are rational* – but this is outside my scope here. I will return to Cassam's argument in §5 where we will see that it does not apply to motivating reasons. The following two sections now provide positive arguments for thinking that our motivating reasons self-intimate – one quick and one slow.

7.3 Self-intimation and motivating reasons: The Quick Argument

My contention is this: necessarily, if S has a motivating reason that *p*, provided she possesses the relevant concepts, S will be in a position to learn that she has that motivating reason. (It is not simply that S will be in a position to learn the proposition that *p*.) To clarify, recall the different types of reason. We can talk of motivating reasons: the reason for which S has an attitude or performs an action. And there are also purely causal explanatory reasons, like biases and the like. So, there may well be facts (or beliefs about facts) that causally influence a subject's belief. But if S isn't in a position to learn of the relevant consideration as a reason, then it would only be a purely causal explanatory reason and not her motivating reason. It's not even that she would be irrational in having the consideration as her motivating reason – it just isn't a motivating reason at all. I should note that while it is necessary for having a motivating reason that the subject is in a position to learn that she has it, this isn't to say whether ascribing a consideration as her reason is *sufficient* for it being her motivating reason (more on this later. I disagree here with Setiya 2013).

[5] Cassam's argument wouldn't hold if we adopted Gendler's conception of belief as contrasted with alief (2008a, 2008b). But I am persuaded by Cassam (2014: 108-9) against this tactic – we cannot make the same move with other attitudes, so it would separate belief from them too much.

In this section, I provide my Quick Argument for thinking that motivating reasons self-intimate. I set it out in terms of reasons for belief, but it can be generalised (I say something about reasons for other attitudes at the end of §7.4). I start with an intuitive premise before then pinpointing an implication which leads to our conclusion:

> PREMISE. Necessarily, if S has a motivating reason that *p* for having attitude *A*, *p* makes *A* intelligible to S.
>
> CONSEQUENCE. For a motivating reason *p* to make *A* intelligible to S, S must be in a position to learn that her motivating reason is *p*.
>
> CONCLUSION. Necessarily, if S has a motivating reason that *p* for having *A*, then S is in a position to learn that her motivating reason is *p*.

(For simplicity's sake, I assume reasons are propositional, but I would be happy to say that experiences can constitute both normative and motivating reasons.)

First, why accept our PREMISE to the Quick Argument? The PREMISE forms one way of capturing S's point of view when she, say, believes that *q* for a reason. Often philosophers observe when S believes that *q* for the reason that *p*, *p* justifies the proposition that *q* in S's eyes. As such, *p* renders believing that *q* a sensible thing to do in S's eyes. I will discuss this standard thought later. But, I contend, believing something for a reason also makes it intelligible to S that she has this belief about the proposition. It is also *explicable* to S why she has that belief. I don't want to say yet that S possesses an explanation, i.e., that S *believes* that her belief is explained in such-and-such way. That would be essentially to say that S knows why she has the belief, which would be to beg the question. But, provided she possesses the relevant concepts, S will be in a position to form a rationalising explanation of her action/attitude.

Take, for example, my belief that *it will rain*, which is based on the reason that there are grey clouds in the sky. it comes as no surprise to me that I have this belief. I do not find myself inexplicably believing it as if by chance.[6] Rather, it is rationally explicable to me that I have that belief – I believe it on the basis of the grey clouds. On the other hand, consider a belief that isn't based on reasons – that of Norman, BonJour's (1980) clairvoyant. Recall that Norman finds himself with the belief that *the president in town* as a result of a newly-gained reliable sixth sense. And indeed, the belief is true. Yet as far as Norman is concerned, it just happens upon him.

[6] Ward considers a similar sort of scenario and what he terms the 'Separation Thesis,' according to which 'it is possible for me to think that my justification for my present belief that p can lie in completely different considerations from my explanation for my belief that p' (2002: 241). Ward denies this, claiming that one would not believe that *p* in that case.

Regardless of whether or not it is justified, the belief isn't based on reasons. Importantly, Norman's position is not just unusual because nothing in his eyes justifies the proposition that the president in town. The belief just pops into his mind without him having any sort of idea about the causal history of it.

We can clarify what is unusual about Norman by considering a different clairvoyant:

> CLARA: Clara reads the newspaper in the morning. It features an article saying that the president is in town. Clara believes that this article provides a normative reason to believe that the president is in town, but the president's immanent presence didn't 'sink in.' She wouldn't avow the claim that the president is in town, and neither does she act accordingly. That is, although Clara takes there to be good reason to believe the president is in town, she does not actually believe the president is present. Say, though, that in the afternoon, she develops a reliable sixth sense which leads her to believe that *the president is in town*. The belief pops up out of nowhere as far as she is concerned.

Like Norman, Clara also believes that *the president is in town* as a result of a sixth sense. The belief happened upon Clara out of nowhere in the same way as Norman's. But she happens to take there to be normative reasons for believing that *the president is in town*. And perhaps, unbeknownst to her, Clara's belief about the president is even sensitive to her evaluation of the evidence. I.e., if Clara decides that the newspaper is unreliable, she would no longer believe that *the president is in town.* But still, it is just a lucky accident from her perspective that the belief fits with what she takes to be a normative reason. The belief would come as a surprise to Clara – she just finds herself with it out of the blue. She may well be puzzled as to why she suddenly has this belief. In that way, her belief is like Norman's, and unlike my belief that it will rain. So, I contend, Clara does not believe that *the president is in town* for the reason that the newspaper says so.

We might also put this sort of point in a different way. One claim is that in believing that *q* for a reason, I am aware of what, in my mind, propositionally justifies believing that *q*, i.e., what would justify my belief that *q* if I were to form it. This is partly what distinguishes my case from that of Norman. But it seems that I am also aware of what, in my mind, is my doxastic justification, i.e., what justifies my belief which I do in fact hold. (This is not to say that the belief actually *will* be justified.) This distinguishes my case from Norman's, and also from Clara's. Clara is clearly aware of the propositional justification for having the belief: the newspaper article. But she has no awareness at all about the doxastic justification, i.e., what actually justifies the belief itself. But when I believe it will rain, I have some sort of awareness of both. I am aware that the grey clouds justify believing that *it will rain*, and that the grey clouds justify my belief itself, i.e., that the fact there are grey clouds forms part of my belief's evidential base.

We have, then our PREMISE: necessarily, if S has a motivating reason that *p* for having attitude *A*, *p* makes *A* intelligible to S. This PREMISE has an interesting CONSEQUENCE. For S's motivating reason to render her belief intelligible in this way, she does not just need to be prepared to take the consideration to be a normative reason. S also needs to be in a position to know that it is her reason, and that it is why she has the belief. Otherwise, as far as S is concerned, she might as well be Clara. This CONSEQUENCE leads to the conclusion that motivating reasons self-intimate: necessarily, if S has a motivating reason that *p* for having *A*, then S is in a position to learn that her motivating reason is *p*.[7]

It is worth, though, considering a criticism. We might think that understanding what it is to have a reason in this way over-intellectualises matters. After all, we might want to say that non-human animals and children can believe on the basis of reasons.[8] There is a sense in which Fellini the cat believes that *Lizzy is home* for the reason that he hears the door opening. And seven-year-old (human) Harrison believes that *Real Madrid is the best football team* for the reason that Ronaldo is in it. Those considerations do not just play a causal role – Fellini and Harrison's beliefs do not seem alien to them. Yet both Harrison and Fellini lack the concept of a motivating reason and as such are not in a position to self-ascribe their motivating reasons. And my picture might also look overdemanding even in the case of adult humans. After all, it is implausible that for every belief we form on the basis of reasons, we know what those reasons are.

But I am only claiming that subjects will necessarily be *in a position* to know their motivating reasons in virtue of having them, and even that requirement depends on one possessing the requisite concepts. So, I allow for cases in which the subject forms a belief in the quick un-self-reflective way that is commonplace. Such a subject wouldn't have thought about why she holds that belief until she is asked. But still, she is able to form the warranted true belief that she holds the lower-order belief for a given reason. And children and non-human animals can still have beliefs for reasons, even though they cannot self-ascribe them – such individuals simply lack the relevant concepts. Waiving this requirement for those who lack the concept of a reason is not *ad hoc*. The motivating reason plays the sort of role that means that, if Fellini and Harrison did have the relevant concepts, they would be in a position to self-ascribe their motivating reasons. As this is only meant to be a quick argument, I cannot explore what this role amounts to. But briefly, it would probably concern the primitive doxastic control that even Fellini and Harrison would exercise insofar as they can be said to believe on the basis of reasons.

---

[7] Relatedly, see Neta (n.d.) who argues that basing is constituted by ostending a given consideration as one's 'commitment justifier.'

[8] The literature on epistemic basing and related issues often discusses 'children and animals' objections. See e.g., Leite (2008: 422) and McHugh and Way (2016: 187).

That all said, I recognise this is argument is contentious. Usually, discussion of reasons for belief are cashed out in other terms (see below), and I have made some large jumps. So, I will now provide a slower argument. My Main Argument is in some ways less controversial, but still related to the first – premises one (i) and two roughly map onto the first premise of this quick argument.

7.4 Self-intimation and motivating reasons: The Main Argument

My Main Argument is less contentious in two ways. It proceeds from a more common-place position about the nature of motivating reasons that doesn't in itself require awareness of that motivating reason. If this view is the case (constituting PREMISE ONE below), then it follows that subjects would be in a position to learn of their motivating reasons using RTM. The Main Argument is also less contentious than the Quick Argument because the Main Argument's conclusion amounts to a conditional: if RTM is true, then reasons self-intimate. I hope, though, that one accepts the antecedent.

The Main Argument proceeds as follows:

> PREMISE ONE. Necessarily, if S has a motivating reason that *p* for her attitude *A*, then S is prepared to take that *p* to be a normative reason for having *A.*
>
> PREMISE TWO. Necessarily, if S is prepared to take that *p* to be a normative reason for having *A*, then S is in a position to learn that her motivating reason is *p* using RTM.
>
> CONCLUSION. Necessarily, if S has a motivating reason that *p* for having *A*, then S is in a position to learn that her motivating reason is *p* using RTM.

We can also express the argument in the following way. Let MR =(def.) *S has a motivating reason that* p *for her attitude* A; TNR =(def.) *S is prepared to take that* p *to be a normative reason for having* A; and SK =(def.) *S is in a position to learn that her motivating reason is* p *using RTM*. Then we see that the argument is at the very least valid since it amounts to the following argument.

PREMISE ONE: Necessarily, if MR, then TNR.

PREMISE TWO: Necessarily, if TNR, then SK.

Therefore,

CONCLUSION: Necessarily, if MR, then SK.[9]

PREMISE ONE concerns the crucial aspect of motivating reasons that allows us to avoid Cassam's objection. Subjects may be credited with a belief that they do not judge to be true, but they cannot be credited with a motivating reason they are not prepared to take to be a normative reason. It is also worth noting that I take my conclusion to hold for our motivating reasons for all attitudes. That is why PREMISE ONE of the Main Argument references normative reasons in general. I focus on belief since a lot of the relevant literature concerns this. I take normative reasons for belief to only be epistemic (as opposed to practical). Indeed, the following discussion proceeds if that is the case. But the main argument itself pertains to motivating reasons in general and taking the consideration to be a normative reason *simpliciter*, be that epistemic or practical. I revisit this at the end of the section.

I will now argue for PREMISE ONE and TWO in §7.4.1 and §7.4.2 respectively. I will conclude that the Main Argument is sound.

7.4.1 PREMISE ONE of the Main Argument

PREMISE ONE. Necessarily, if S has a motivating reason that *p* for her attitude *A*, then S is prepared to take that *p* to be a normative reason for having *A.*

It is worth noting that I just take it to be *necessary* for having a motivating reason that S is prepared to take *p* to be a normative reason. This is not to say whether it is sufficient. I take it that PREMISE ONE is not sufficient, and that S might be prepared to take a consideration to be a normative reason for believing that *q* without the consideration being her reason for believing that *q*. But my overall argument would also work if one accepted this stronger claim.

PREMISE ONE of the Main Argument makes a claim about the epistemic basing relation – to believe that *q* for the reason that *p* is for your belief that *q* to be based on your belief that *p*. In

[9] We can also put it as:

□ (MR → TNR)
□ (TNR → SK)
⊨
□ (MR → SK)

this subsection, I will introduce some prominent options in the basing debate and where my view fits in (§7.4.1.1). This discussion should prove timely. So far, I have made claims about basing and rationality without discussing it explicitly. I will then argue for this position over the other possibilities (§7.4.1.2) and consider an objection (§7.4.1.3).

7.4.1.1 Basing relations

We can roughly carve up the debate into two camps – those who place a doxastic requirement on basing and those who do not.

Doxastic accounts require that for S's belief that *q* to be based on her reason that *p*, she must believe that *p is a normative reason for believing that q*. At the account's strongest, we might think that it is both necessary and sufficient for believing that *q* on the basis of *p* that one believes that *p is a normative reason for believing that q* (see Setiya 2013, and Leite 2004, 2008). A weaker version is just that it is necessary, although not sufficient, that one believes that *p is a reason for believing that q*. Longino (1977) and Audi (1993) both have views of this sort, and they supplement it with a causal requirement – that the belief that *q* also causes the belief that *p* in the right way.[10]

Non-doxastic accounts deny that basing requires a relevant metabelief – that *p is a normative reason for believing that q*. They then vary as to what *is* necessary. First, causal accounts argue that 'for a belief to be based on a reason, the reason must cause the belief in an appropriate way' (Korcz 2015). For example, under at least some versions of the account, we would say that a subject's belief that *q* is based on the reason that *p* if it originally caused her to have the belief, even if she has now forgotten that reason. Or we might also more plausibly appeal to what *sustains* the belief now.[11] An alternative to the causal account is the dispositional account. This is worth bearing in mind as a contrast to my own, for reasons that should become clear later. The view sees basing a belief on a reason as a matter of possessing the disposition to revise the belief in line with that reason (although that disposition might be blocked). For Ian Evans, 'S's belief that *p* is based on *m* iff S is disposed to revise her belief that p when she loses *m*' (2013: 2952). So, my belief that *it will rain* is based on the grey clouds in the sense that I am disposed to stop believing that *it will rain* if I find out that the grey clouds have dissipated. Believing that [*m* is a reason for believing that *p*] is neither necessary nor sufficient for believing that *p* on the basis of *m* (2013: §4).

[10] Relatedly see also 'treating' accounts – Lord and Sylvan (n.d.) and Neta (n.d.).
[11] Causal accounts are offered by McCain (2012), Moser (1985), and Turri (2011). See Korcz (2015) for helpful discussion. Relatedly, for a counterfactual account of basing, see Swain (1985, 1981).

My claim is that for S to base the belief that *q* on the reason *p*, she must be prepared to take *p* to be a normative reason for believing that *q*. As such, my claim opposes non-doxastic accounts in positing an important role for metabeliefs. In places I will refer back to Evan's approach since it fits better with Cassam's objection, but I trust that my criticism of his account of basing also applies to all non-doxastic accounts, including causal ones.

My basing claim is akin to a doxastic account, but with an important difference. I do not require that subjects *always* take the relevant consideration to be a reason, just that they are *prepared* to do so. So, they might believe that *it will rain* for the reason that there are grey clouds without believing that *a normative reason for believing that it will rain is that there are grey clouds*. But the subject will be prepared to form the belief when faced with the relevant questions, such as 'why believe that *it will rain*?'. It is also worth emphasising that, for the MAIN ARGUMENT to work, we need only accept that this doxastic component is necessary for basing. We need not also think that the doxastic component suffices, and indeed, I think it would not. I take it that a further element is also necessary for basing, such as the appropriate causal links or the disposition to revise one's belief. My claim, then, is far less demanding than it could be.[12]

I hope it is fairly intuitive what being 'prepared' to take *p* to be a reason for believing that *q* involves. It would amount to something like being disposed to generally form the belief that *p is a reason for believing that q*. This disposition would be manifested in various situations, such as if one were considering whether *p* is a reason for the belief, or what reasons speak in favour of believing that *q*. This disposition to generally provide reasons could not be blocked from manifesting. Perhaps the subject might happen not to provide a reason in one or two instances, but this cannot be the general course of affairs.

So, to clarify, I take it that subjects lack motivating reasons in both of these cases:

> ARTHUR: Arthur believes that *Belle is the weakest candidate for a technical job*. Arthur also believes that *women are less capable than men at technical matters*, even though he does not sincerely assert this belief. It may well be that his belief about technical jobs disposes him to revise his belief about Belle in various ways. But he is not prepared to take the fact that *women are less capable* to be a reason for believing that Belle is the weakest candidate – he judges that they are equally capable. As such, Arthur cannot be said to base his belief *that Belle is the weakest candidate* on the supposed fact that *women are less capable*.

[12] Certainly my view is less demanding than Setiya (2013). He seems to think that our motivating reasons would be self-intimating, but as part of a picture where the appropriate metabelief is *sufficient* for basing.

CHO: Cho believes that *externalism is true*. She has just been in a lecture where the lecturer presented an argument for externalism, although she is confused about how all the claims relate to it. Maybe she unconsciously revises her beliefs in line with the argument because it is salient to her. If she discovers the argument is unsound she would no longer have the belief. But she isn't prepared to take it to be a reason for externalism – she is too confused about it and the role it plays. As such, she cannot be said to base her belief that *externalism is true* on the lecturer's argument.

7.4.1.2 Arguments for PREMISE ONE

I will now present three arguments for PREMISE ONE of the Main Argument against alternative accounts of the basing relation. The first two argue for PREMISE ONE against non-doxastic accounts of the basing relation. They also speak in favour of the traditional doxastic account (that having a motivating reason requires taking it to be a normative reason). My last argument argues for my claim in PREMISE ONE against this traditional version.

i. Motivating reasons as rationalisers

Recall PREMISE ONE of the Main Argument: Necessarily, if S has a motivating reason that *p* for attitude *A*, then S is prepared to take that *p* to be a normative reason for having *A*. My argument for PREMISE ONE in this subsection proceeds from an intuitive observation about the role motivating reasons play. My argument then considers what motivating reasons would need to be like to play that role. Believing on the basis of reasons is the archetypal way of producing beliefs that are rational in one's own lights. For a motivating reason to have this rationalising role, S would have to be prepared to take it as a normative reason for the belief. I call this the Rationalising Argument (RA) for PREMISE ONE of the Main Argument. We can spell this out more precisely as follows:

PREMISE ONE. When S bases a belief that *q* on a reason that *p*, *p* makes the belief rational in S's lights.

PREMISE TWO: For *p* to make S's belief that *q* rational in her lights, S must be prepared to take *p* to be a reason for believing that *q.*

CONCLUSION: Necessarily, if S has a motivating reason that *p*, then S is prepared to take that *p* to be a normative reason.

We can find this sort of argument in Leite (2008) and relatedly in Jarvis Thomson (1965) concerning responsible belief formation. Let me say something about these two premises in turn.

First, I should clarify what PREMISE ONE of the Rationalising Argument amounts to. I mentioned when setting out the Quick Argument that motivating reasons play two roles from the subject's point of view. There I discussed how believing for a reason makes it explicable to the subject that she has this belief – it has not just popped up out of thin air. But standardly, philosophers tend to focus on the contention that believing for a reason makes it understandable to S why she *should* have this belief (or why it is permissible). That is, when S believes that *q* for a reason that *p*, the reason means that it makes sense in S's eyes to believe that *q*. This is what I mean by the first premise of the Rationalising Argument.

In arguing that reasons play this role with belief, it is first worth considering how performing an action for a reason rationalises that action. That motivating reasons play a rationalising role is less contentious in the case of action, and we can draw relevant parallels to belief. To start with an example: I go to the shop for the reason that I've run out of hummus. That (at least as far as I know) I have run out of hummus makes going to the shop a reasonable thing to do in my eyes. Or, to use Quinn's (1993) language, the motivating reason means that going to the shop is 'intelligible,' 'sensible,' and it 'makes sense' to me. It would not seem like such a good idea if I believed I had a well-stocked fridge. We can contrast my going to the shop for a reason with cases where subjects fairly uncontentiously lack a reason. It is not that I go to the shop, say, as a result of a compulsion even though I recognise that I do not need anything. I would be alienated from this action – it is a foolish thing to do in my lights.

It looks like motivating reasons also rationalise belief in a similar way. (This is not to say that motivating reasons for belief and action are akin in all important respects. Indeed, at this stage, I need not hold that they are alike in any fundamental way at all, although I do. Here, just let me observe that the case of action nicely illuminates some features of believing for reasons.) When S believes that *q* on the basis of *p*, believing that *q* seems to S like a sensible thing to do in light of *p*. Take for example my belief that *it will rain*, which is based on the reason that there are grey clouds. It makes sense to believe that *it will rain* in light of the grey clouds. I cannot say much more at this point as to why it would make sense to believe it – that would presuppose my account of reasons. But it should, I hope, seem *prima facie* plausible. After all, compare this with a case in which one does not believe for a reason. Recall again BonJour's clairvoyant Norman. Norman's belief that the president is in town is not based on reasons. As such, there is nothing to make it seem sensible in Norman's eyes. As BonJour writes:

> From [Norman's] standpoint, there is apparently no way in which he *could* know the President's whereabouts. Why then does he continue to maintain the belief that the President is in New York City? Why is not the mere fact that there is no way, as far as he

> knows or believes, for him to have obtained this information a sufficient reason for classifying this belief as an unfounded hunch and ceasing to accept it? And if Norman does not do this, is he not thereby being epistemically irrational and irresponsible? (1980: 62-3).

And, even if there is a reliable connection between Norman's belief and the truth, still, 'from his subjective perspective, it *is* an accident that the belief is true' (1980: 63). This should not be contentious – philosophers may well question whether Norman's belief is justified, but not whether it is based on a reason, is intelligible to Norman, and is responsibly formed.

Further, it seems to be a conceptual truth that motivating reasons rationalise one's attitudes. That is, it is not just that subjects' motivating reasons for belief ought to rationalise their beliefs in this way, or that the motivating reasons normally do. Rather, if a consideration does not make it rational in S's eyes to believe that *q*, then it cannot be their reason for believing that *q*. After all, it is hard to make sense of a 'motivating reason' that does not make believing that *q* a rational option in the subject's own lights. In what way would it still be their *reason* for holding the belief? A belief that *p* may well causally influence a subject's belief that *q* in various ways, but this is not to say that *p* would be the reason for which she believes that *q*.

Before continuing, I should clarify two things. First, one may worry that I over-intellectualise matters. One might think it is over-demanding to say that believing that *q* for a reason is to see believing that *q* as rational. But the relevant sense of 'seeing believing that *q* as rational' can be fairly minimal. I do not think subjects must form the belief that *believing that q is rational*. And neither do I think it requires having the concepts of *belief* and *rationality*. I think 'seeing believing that *q* as rational' would probably fundamentally amount to some form of tacit *de re* awareness. Obviously, this is very sketchy, but it is outside my project to say much here. It is enough to note that there seems to be a difference from the subject's perspective between my belief that *it will rain* and Norman's belief, and that it concerns the rationality of believing the relevant propositions from our perspectives. There plainly is some sort of difference, and so I just have whatever this amounts to in mind. And second, the motivating reason makes the belief rational in the subject's eyes in a way that is proportionate to the weight the subject places on it. The subject may take there to be reasons which are outweighed. Accordingly, the reason would make the relevant belief seem like a somewhat sensible thing to believe, but not over all.

Let us now consider PREMISE TWO of the Rationalising Argument: For a reason *p* to make S's belief that *q* rational to her, S must be prepared to take *p* to be a normative (i.e., an epistemic) reason for believing that *q*. It can't be sufficient that S has a disposition to revise her belief in accordance with *p*, say, or that her belief was simply caused by her belief that *p*.

It seems intuitive that a subject would at least need to be prepared to take *p* to be a normative reason for believing that *q* for *p* to rationalise the belief that *q* in her lights. Let us return to the case in which I believe that *it will rain* for the reason that there are grey clouds. As mentioned above, the grey clouds make forming that belief sensible in my eyes. And it seems that this is because I take it to be true that there are grey clouds, and I recognise that this speaks in favour of believing that *it will rain*.

At any rate, motivating reasons as understood by non-doxastic accounts could not play this rationalising role. Non-doxastic accounts allow cases where the subject believes for a reason without it making the belief at all rational in her eyes. To see this better, let us firstly return to the case of action. When a subject *φ*s for the reason *p*, *p* makes *φ*ing rational in subject's eyes. This places constraints on how we understand reasons for action. As Quinn (1993) discusses, one might construe reasons for action as simply 'functional-dispositional' states to seek out certain states of affairs, but then we face compelling counter examples.[13] Quinn presents a subject who has an urge to turn on radios without seeing it as at all worth doing. The urge wouldn't rationalise turning on radios for the subject and as such, cannot be their motivating reason.[14] Indeed, these cases are not just the preserve of thought experiments. Subjects with mental disorders such as OCD have compulsions to do certain things, and yet these compulsions do not rationalise performing the action. Cleaning the table, say, may well have nothing going for it in their eyes, other than perhaps the fact that doing so will temporarily alleviate their anxiety. As a result, their action does not make sense to them; rather, they are alienated from it.

My discussion regarding action bears similarities with the Guise of The Good debate but should not be confused with it. The 'Guise of the Good' thesis claims that for a subject to act at all is to see that action as good in some way.[15] E.g., The radio-obsessive isn't *acting* at all in that he sees nothing of merit in turning on radios. Yet I am not saying anything as committing. I can accept that the radio-obsessive acts in turning on the radios – I just contest that he acts for a reason.

With the practical case in mind, let us return to the epistemic basing relation and the case of Arthur. Say we understand reasons under Ian Evans' dispositional model. Recall that for Evans, 'S's belief that *p* is based on *m* iff S is disposed to revise her belief that p when she loses *m*' (2013: 2952). And, importantly, S's belief can still be based on *m* if the relevant disposition is blocked

[13] I should note that Quinn discusses all this in terms of the role of desire in motivating action. But Setiya (2010) convincingly writes that we can also understand discussions of this sort in terms of reasons.
[14] It is also important for Quinn's purposes that turning on radiators isn't *actually* valuable either. I, though, just want to claim that motivating reasons make something rational in the subject's eyes even if it isn't objectively so.
[15] This position is endorsed by Anscombe (2000) and Quinn (1993) nicely discussed in Setiya (2010).

from manifesting. Evans may well say, then, that Arthur believes that *Belle is the weakest candidate* for the reason that Belle is a woman. (And let's suppose for simplicity's sake that this would be Arthur's only reason for his belief.) Arthur's belief that *Belle is the weakest candidate* is sensitive to his belief that *Belle is a woman* – he will revise his conclusion in line with it. For example, say Arthur were to discover that the names on the CVs were mixed up (he never met the candidates) and it turns out that Belle is actually a man called Boris. Arthur is disposed such that he would stop believing that Belle/Boris is the worst candidate. Nevertheless, Arthur is not prepared to take the fact that *Belle is a woman* to be a reason for believing that *Belle is the weakest candidate* – he judges that women are just as good at technical jobs as men.

Arthur's disposition would not make believing that *Belle is the weakest candidate* rational in his eyes. He would be like BonJour's clairvoyant or someone with obsessive beliefs. Arthur cannot shake his conviction that *Belle is the weakest candidate* and can think of nothing that supports it. Indeed, it would be a lucky accident from his perspective if the fact that *Belle is a woman* turned out to speak in favour of the believed proposition, and indeed if the belief was true. As Leite writes about inference:

> From [the point of view of the agent], to move from premise to conclusion without taking one's premise to support the conclusion is simply to guess. Even if the truth of the resultant belief isn't – relative to certain fact – merely a lucky coincidence, still from one's own point of view it would look at best like a lucky coincidence, and that's why the transition is irresponsible (2008: 424).

For the proposition that *Belle is the weakest candidate* to seem worth believing from Arthur's perspective, Arthur would actually need to take the fact that *Belle is a woman* to be a normative reason or have the implicit awareness that means he would be prepared to do so. Then the belief would be rational in his lights. (Of course, though, Arthur's belief that *Belle is the weakest candidate* would still not be objectively rational, because the fact that *Belle is a woman* is a bad reason for the belief.)

Let me say more about why we cannot credit Arthur with a motivating reason. There are two ways in which we need a doxastic account of motivating reasons in order to explain how motivating reasons can rationalise holding a belief.

i. First, we should note that for believing that *p* to seem sensible to S, S needs to have some indication that the belief would be true. This is because belief aims at the truth. Or, even if there is no norm to believe as many true things as possible, at the very least, beliefs are such that those

we do have ought to be true.[16] As such, it is only sensible for S to believe propositions that are likely to be true as far as she is concerned.[17] To play the rationalising role, then, motivating reasons would have to somehow suggest that the belief would be true if S formed it. But under a non-doxastic account, reasons would not do this. Take the mere disposition to revise one's conclusion in light of another belief. This disposition doesn't tell S anything about whether the two beliefs are connected or whether the basing belief is even true. But, on the other hand, it's clear under a doxastic account how believing for a reason makes believing that *p* sensible. Subjects are prepared to take the reason to be a normative one which, I think, would involve some sort of tacit awareness that the consideration is a normative reason. Forming beliefs in line with normative reasons is a good way of forming true beliefs. So, it makes sense from S's perspective to form beliefs in light of the considerations she is prepared to take to be normative reasons.

ii. Second, if a motivating reason is to make believing that *p* intelligible for S, then it needs to in some way impact on S's consciousness.[18] Say Cho is disposed to judge that *the lecturer's argument is a reason for accepting externalism*, but her confusion about the issues block the disposition from manifesting. Since she has no form of conscious awareness of the connection between the reason and conclusion, it is not clear how the consideration would make believing that *externalism is true* seem rational *to her*.

Putting both PREMISE ONE and PREMISE TWO together, we get the conclusion: Necessarily, if S has a motivating reason that *p*, then S is prepared to take that *p* to be a normative reason.

ii. Motivating reasons and rational control

In this subsection I take a similar route as in (i) to argue for PREMISE ONE of the Main Argument. Again, I proceed from an intuitive observation about the role motivating reasons play and what motivating reasons must like in order to play that role. As well as rationalising subjects' beliefs in their eyes, believing for reasons gives subjects rational control over their beliefs. Exercising this control is part of what it is to be a rational agent, who *does* something in, say, believing that *p* (McHugh 2013: 132-5). But subjects would not exercise rational control in believing *q* for the reason that *p* if they were not prepared to take that *p* to be a normative reason. Let's call this the Rational Control Argument (RCA) for premise one of the Main Argument. RCA proceeds as follows:

[16] See Whiting (2012).
[17] Relatedly, see Whiting (2014).
[18] For a discussion of the importance of conscious deliberation for doxastic agency, see McHugh (2013).

PREMISE ONE. Believing for reasons gives subjects rational control over their beliefs.

PREMISE TWO: For S to exercise rational control in believing that *q* for the reason that *p*, S must be prepared to take *p* to be a normative reason for believing that *q.*

CONCLUSION: Necessarily, if S has a motivating reason that *p* for believing that *q*, then S is prepared to take that *p* to be a normative reason for believing *q.*

In these arguments, I follow Leite (2004) and Longino (1977) in my own way. I will motivate both premises in turn.

There are two aspects to PREMISE ONE of the Rational Control Argument. (i) We do indeed have rational control, and we can acquire beliefs by forming them on the basis of reasons. It isn't that we can only affect our beliefs through cognitive behavioural therapy (CBT), hypnosis, and the like. (ii) Believing for reasons *always* gives us this control. This is not to say whether believing for reasons is the only way in which we can exercise rational control, but that it is at least one way. I will discuss both (i) and (ii) in turn.

i. We can control our beliefs by believing for reasons. One indication is the fact that we often epistemically criticise other people, give others epistemic recommendations, and hold others epistemically responsible.[19] It is commonplace to say things like: 'you shouldn't believe that *q* for the reason *p* because…. *p* is false/*p* isn't a reason/*p* is a reason but is defeated/*p* is a reason but is outweighed.' We should note two things. (a) Intelligibly criticising someone requires that the subject had control over what they did. In this case, it requires that the subject could have ceased to believe that *q* simply for the reason that *p* without, say, having to engage in CBT or hypnosis. Compare this with cases where subjects obviously can't exercise rational control. There we wouldn't hold people responsible. We don't criticise someone with OCD for believing on the basis of a bad reason, and this is because they can't change their belief in line with what they take to be reasons. We might criticise them for not going to the doctors, but we would not simply say 'but that belief is bad because of the following pieces of evidence…'[20] (b) If subjects didn't generally have direct rational control, these sorts of epistemic criticisms wouldn't be entirely relevant. The criticisms would be missing some sort of recommendation about the causal influences one should employ. E.g., we would regularly say things like 'you shouldn't believe that *p*; you should visit a hypnotist.' Yet these criticisms do seem to get to the heart of the matter.

[19] E.g., Boyle (2009, 2011a, 2011b), Leite (2004), and McHugh (2013).
[20] See McHugh (2013).

ii. Further, believing for a reason *always* results in such control. That is, there are no cases in which subjects believe for a reason without exercising rational control over that belief. We don't pepper our epistemic criticisms with provisos. We do not say 'you shouldn't believe that *q* on the basis of *p* because *p* is a bad reason. Unless of course your doxastic control failed, in which case it's fine.' Furthermore, it would be odd if believing for reasons sometimes provided us with doxastic control, other times not. It would seem too contingent and unreliable from the perspective of the subject.[21] If it could still be counted as control, it wouldn't be *rational* control.

Moving onto PREMISE TWO of the Rational Control Argument: for believing on the basis of a reason to give S rational control, S must be prepared to take this reason to be a normative reason. Not being prepared to take *p* to be a normative reason must entail that *p* is not S's reason for her belief. Here I want to say two things.

i. If a non-doxastic account was true, then our epistemic criticisms would miss the mark, and yet they seem appropriate. If S's believing for a reason simply consisted in, say, having a disposition to revise her belief in line with certain evidence, then we would make different criticisms. Instead of telling S that her reason is a bad reason (e.g. '*p* is false'), we would say that she ought to induce different dispositions in herself. After all, whether or not the reason is a good reason wouldn't be relevant per-se; it is not that, for Evans, the subject must be disposed to take the consideration to be a good reason.

ii. If a non-doxastic account was true, then Cho's and Arthur's beliefs would count as rationally controlled by them. Indeed, so would obsessive beliefs (e.g., the obsessive compulsive's belief that the table is dirty). But this isn't the case. Maybe there might be some sense in which the subject, or a certain part of them, controls the belief, but it wouldn't be *rational* control. We should think this for two reasons.

First, if there is a sense in which Cho, Arthur, etc. exercise control over their attitudes, it wouldn't be *rational* control. It wouldn't be the sort of control that subjects exercise when believing for reasons. For S's control to be rational, it must be that S does something that seems sensible in her eyes. But, as we have seen, under a non-doxastic account, S's motivating reasons need not make believing that *p* seem like a rational thing to do from her perspective.

Second, if there is a sense in which Cho, Arthur, etc. exercise control over their attitudes, it wouldn't be rational *control*. Indeed, it doesn't make sense to talk of the subject themselves exercising control at all – instead, it seems to be a subpersonal mechanism. When Cho and Arthur

[21] Relatedly, see also McHugh (2013: §3.2).

believe as they do, their beliefs have happened upon them as a result of some sort of disposition. It is not something *they* do. Indeed, Cho and Arthur still wouldn't exercise rational control if they were disposed to take *p* to be a reason but the disposition was blocked from manifesting. After all, for the subject to exercise control, at least some of the relevant processes need to be conscious so that they're visible to the subject *herself*.[22]

Say, then, that we accept both premises of the Rational Control Argument. These give us our conclusion: necessarily, if S has a motivating reason that *p* for believing that *q*, then S is prepared to take that *p* to be a normative reason for believing *q*.

***

On the basis of the Rationalising and Rational Control Arguments, then, we should accept some form of doxastic account. Now I want to say why we should accept my particular version, whereby subjects only need to be *prepared* to believe that *p* is a reason for believing that *q*.

iii. Psychological plausibility

So far, in arguing for premise one of the Main Argument I haven't ruled out the traditional doxastic account of basing. To recall, traditional doxastic accounts claim that necessarily, if S believes that *q* for the reason that *p*, S believes *that p is a normative reason for believing that q*. Indeed, my Main Argument would also work if one accepted that principle instead. But I prefer my version because the traditional doxastic account is implausible on psychological grounds. It seems implausible that for every belief we hold on the basis of reasons, even trivial beliefs, we actually have two beliefs – the belief that *q*, and the belief that *p is a reason for believing that q.* My view avoids this worry. It concerns just being *prepared* to take the consideration to be a reason. That is, subjects need only be disposed to generally form the belief that *p is a reason* in the appropriate circumstances. Subjects do not actually have to have formed the belief – the relevant circumstances may not have arisen.

***

We have, then, good reason to accept PREMISE ONE of the Main Argument. The claim about basing in PREMISE ONE best captures the role played by motivating reasons for belief. Motivating reasons rationalise the belief (as we saw in the Rationalising Argument) and furnish subjects with rational control (as we saw in the Rational Control Argument). In order to play those roles, motivating reasons have to be the sorts of things that doxastic accounts say they are. Further, we

[22] Leite (2008, 2004).

should accept the precise version of a doxastic account given in PREMISE ONE of the Main Argument because it is generally more plausible.

7.4.1.3 An objection to PREMISE ONE of the Main Argument

We might, though, object in the following way to PREMISE ONE of the Main Argument: it seems to threaten a regress of propositional justification.[23] According to my view, believing that *q* for the reason that *p* requires being prepared to believe that *p is a reason for believing that q*. Sometimes *p* will be a bad reason and the subject won't be justified in believing that *q*. But often, the basing relation will result in justification – S's belief that *q* will be justified because it is based on a reason *p* which is itself justified. I take it that basing requires that the subject is prepared to form the relevant metabelief. But, my opponent might argue that for the basing relation to transmit justification, the metabelief would have to be justified if S formed it. That is, S would have to have propositional justification for believing that *p is a normative reason for believing that q*. And, most obviously, we might think that the metabelief would be propositionally justified in that S has access to a normative reason *r* on the basis of which she could believe that *p is a reason for believing that q*. So, S would have to be prepared to form the belief that *r is a reason for believing that [p is a reason for believing that q]*. Which itself would have to be justified, and so on.

To cash this out with an example: my belief that *it will rain* is justified because justification is transmitted from my belief that *there are grey clouds*. For the basing relation to hold, I also need to be prepared to believe that *the grey clouds are a reason for believing that it will rain*, which we might think must also be propositionally justified. And, most obviously, we might think that it would be propositionally justified in that I have access to a normative reason for which I could believe that it is a reason. So, I would have to potentially form the belief that *the fact that grey clouds reliably correlate with rain is a reason for believing that [the grey clouds are a reason for believing that it will rain]*. And this new belief would itself have to be propositionally justified, and so on.

In replying to the regress objection, I have two options open to me, and I am willing to accept either. First, as Leite (2008) observes, the proposition that *p is a reason for believing that q* isn't part of one's propositional or doxastic justification for believing that *q*. That *p is a reason for believing that q* is required for forming a belief responsibly and plays a 'regulatory role' in the inference. This leads me to suggest, then, that being prepared to believe the proposition is simply an enabling prerequisite for justification to be transmitted from one's belief that *p*. The subject

[23] On this sort of criticism, see, for example, Setiya (2013: 197-8), Leite (2008: 426-7) and McHugh and Way (2016: 318-9).

wouldn't need to have propositional justification for the metabelief. Second, we could say that the regress will stop somewhere with basic justification. Perhaps the proposition that *r is a reason for believing that [p is a reason for believing that q]* will be justified in and of itself. I can leave cashing out the nature of this justification for further research.[24]

7.4.2 PREMISE TWO of Main Argument

> PREMISE TWO. Necessarily, if S is prepared to take that *p* to be a normative reason for having attitude *A*, then S is in a position to learn that her motivating reason is *p* using RTM.

I have already motivated PREMISE TWO throughout the course of the thesis. It stems from the fact that RTM is a route to knowledge, which at this stage I will take as a given. Recall that using RTM is to learn what one's reason for believing that q is by answering the question 'why believe that *q*?'. This amounts to considering the question 'what are the normative reasons for believing that *q*?'. For example:

> I believe that *it will rain* having earlier looked out of the window and seen the looming grey clouds. You learn that I believe that *it will rain* and ask me 'why?'. I look out into the world, so to speak, and consider the reasons in favour of believing that *it will rain*. I conclude that the grey clouds are a normative reason and can then report that I believe that *it will rain* for the reason that there are grey clouds.

S will, then, be in a position to use RTM to learn that her motivating reason for believing that *q* is *p* on those occasions when S is prepared to take *p* to be a normative reason for believing that *q*. This is like how S is in a position to learn that she believes that *q* using TM when she prepared to judge that *q*. As with TM for attitudes, RTM will give S knowledge when her answer to the world-directed question corresponds with the mental feature she is trying to learn of (and only then!).

Further, recall that being in a position to acquire knowledge requires having the relevant epistemological assets. S will have the relevant assets for knowledgeably self-ascribing her motivating reason when she is prepared to judge that *p* is a reason for believing that *q*. I will mention two such assets.

i. Being prepared to judge that *p* is a normative reason means that S has accessible warrant for self-ascribing *p* as her motivating reason. This is because judging that *p* is a normative reason provides sufficient grounds for self-ascribing *p* as her motivating reason. S doesn't require a

[24] The account of basing expressed in PREMISE ONE also overcomes the children and animals objection, for the reasons set out in §7.3.

premise that she is rational and that her reason matches what she takes to be a normative reason. As we saw when I introduced RTM, the warrant for self-ascribing one's motivating reason doesn't come from any additional evidence, but from taking the consideration to be a reason in this context and having an agent's awareness that the consideration is one's motivating reason. Because of this, S will be in a position to use RTM to gain knowledge when she has the unblocked disposition to take *p* to be a reason for believing that *q*. The relevant grounds are accessible in having that unblocked disposition.

ii. Beliefs must be formed in a reliable way in order to constitute knowledge. As I argued in chapter six, self-ascriptions formed using RTM would be reliably true (although not *especially* reliable). For example, I noted that subjects do at least sometimes infer instead of employing RTM on occasions when using RTM would not be knowledge-conducive.

As a result, we should accept PREMISE TWO of the MAIN ARGUMENT.

***

Taking the two premises of the MAIN ARGUMENT together we get our conclusion. To recall, PREMISE ONE of the argument stated that necessarily, if S has a motivating reason that *p* for attitude *S*, then S is prepared to take that *p* to be a normative reason for having *A*. We should accept PREMISE ONE on the grounds of the Rationalising and Rational Control Arguments, and because it is more plausible than traditional doxastic accounts. PREMISE TWO stated that necessarily, if S is prepared to take *p* to be a normative reason for having *A*, then S is in a position to learn that her motivating reason is *p* using RTM. As discussed, PREMISE TWO follows from my account that I proposed in chapter six. Together, PREMISE ONE and PREMISE TWO result in the conclusion: necessarily, if S has a motivating reason that *p* for having *A*, then S is in a position to learn that her motivating reason is *p* using RTM. Or, to put the thrust of this section differently: I have argued for a doxastic account of motivating reasons. Motivating reasons understood in this way will self-intimate under RTM.

As a result, I have given another way in which self-knowledge of motivating reasons is distinctive, in addition to those discussed in chapter six. At the subpersonal level, a subject's self-knowledge of her motivating reasons resembles her knowledge of other people's reasons. But, I have argued that at the personal level, the subject can use a distinctive method (RTM) and warrant to learn of her own motivating reasons but not other people's. And I can now add that RTM plays an important metaphysical role. Having a motivating reason entails being in a position to learn that one has the motivating reason using RTM.

Having given both the Quick and Slow Arguments in terms of reasons for belief, let me briefly consider reasons for other attitudes. I take it that necessarily, if S has a motivating reason *simpliciter*, S will be in a position to learn that she has it using RTM. E.g., necessarily, when Tom desires a book for the reason that it looks interesting, Tom will be in a position to know that he desires the book for the reason that it looks interesting.

All my preceding arguments extend to motivating reasons for attitudes other than belief. To give one example, take the Rationalising Argument for PREMISE ONE of the Main Argument. I contend that necessarily, if S desires *x* for the reason that *p*, S is prepared to take *p* to be a reason for desiring *x*. When S desire something for a reason, holding that desire is made intelligible to her. In that way, we can contrast desiring something for a reason with brute cravings. For example, when Tom desire a book for the reason that it looks interesting, desiring the book seems sensible to him. It is not like how one craves food in being hungry. But motivating reasons for desire wouldn't play this role if the subject wasn't prepared to take the relevant consideration to be a normative reason. Wanting the book would seem like an odd thing to do from Tom's perspective if he couldn't avow his belief that *the book is interesting*, or if Tom believed that *interestingness isn't a reason for desiring a book*. As such, necessarily, Tom will be prepared to take the relevant consideration to be a normative reason and will therefore be in a position to learn that he has the motivating reason using RTM.

Indeed, we should expect my self-intimation claim to have this broad scope since it is fairly general. It just makes a commitment about what having a motivating reason *necessarily* requires; it says nothing about sufficiency. Different kinds of motivating reasons may also have different additional necessary conditions. While I do claim that some relevant similarities hold between motivating reasons for all attitudes (and indeed actions), that shouldn't be too contentious. If we can talk of motivating reasons in these diverse cases at all, it will surely be because the reasons all play certain basic roles, such as making the attitude or action rational to the subject.[25]

[25] Regarding my arguments over the past two sections, see also Bilgrami (2006). Bilgrami claims that 'given agency, if S desires (believes) that p, then S believes that she desires (believes) that p' (*Ibid.* 138). Roughly, Bilgrami argues in the following way: S's beliefs and desires rationalise the actions she performs as an agent, and S must believe that she has these beliefs and desires in order for them to play this rationalising role. See also Bar-On (2007) on Bilgrami's account.

7.5 Objections

In §7.5.1, I consider some possible counterexamples to the claim that motivating reasons strongly self-intimate. Then, in §7.5.2 I revisit Cassam's objection to self-intimation from §7.2. Now that the shape of my proposal is more apparent, we should be able to see why self-intimation regarding motivating reasons avoids the worries facing self-intimation theses concerning attitudes.

7.5.1 Possible counterexamples

One might object that subjects sometimes have motivating reasons without being in a position to self-ascribe the reasons using RTM. One might raise the following counterexamples to my self-intimation claim:

> FORGOTTEN REASONS: Ellie believes that *epistemic externalism is false*. She originally formed the belief on the basis of an *a priori* argument. At the time, Ellie took the argument to be a reason for believing that *epistemic externalism is false*. She has, though, now forgotten that argument. She can't even remember that she once encountered an argument. As a result, when I ask Ellie why she believes that *epistemic externalism is false*, she replies that she doesn't know.
>
> SLIPPED REASONS: Will reads an article listing four facts concerning a politician, Jones. Will then forms the belief that *Jones is awful* on the basis of all four of these facts. Later, I learn of Will's belief and ask him why he believes that *Jones is awful*. Will considers the reasons that justify his belief, i.e., the good reasons for believing that *Jones is awful*. Will can only recall three, and can only reply with: 'Jones is racist, has bad foreign policy, is sexist, and ... what was that last one again?!' Later in the day it dawns on Will 'ah yes, Jones is generally incompetent!'. While Will hadn't forgotten the reason, it had slipped his mind when I asked him why he believes that *Jones is awful*.

I will address each case in turn; I deny that either constitute a counter example to my self-intimation claim.

Ellie doesn't have a motivating reason she isn't in a position to know, since, I contest, the *a priori* argument is no longer her reason for believing that *externalism is false*. It once was, but not anymore. Instead she now bases her belief on a consideration such as the memory that the belief is true. I take it that it is necessary for having a motivating reason that *p* that one is prepared to take *p* to be a reason for the belief. Yet Ellie has lost the disposition to generally form the belief that *the a priori argument is a reason for believing that externalism is false*. It doesn't even seem

like she has the disposition to do so but that it is systematically blocked. Ellie might come to regain the disposition if she went over her lecture notes, but as it is, nothing in her psychology links her to the *a priori* argument.[26] I do not think, then, that FORGOTTEN REASONS forms a counterexample to my account.

SLIPPED REASONS requires a different answer. It's implausible to deny that Jones' general incompetence is one of the reasons for which Will believes that *Jones is awful*. After all, basing relations would hardly disappear and reappear in such a spontaneous way. But in this case, I want to say that Will is still in a position to learn of all four motivating reasons using RTM. Will just isn't in a position to form the belief that *he believes that Jones is awful for the reason that Jones is incompetent.* It is like how one can be in a position to know that there is a pomegranate on the table by looking, even if the term has slipped one's mind ('there's fruit on the table … you know, the one with the big seeds, what's it called again? … this is so annoying – I know this!!'). Recall that I take being in a position to know as amounting to possessing the relevant epistemological assets – having accessible grounds, that the belief be reliably formed if one were to form it, and so on. And Will does have these assets. Will still has accessible grounds because he is still disposed to *generally* form the belief that *Jones' incompetence is a normative reason*. Given a little more time he will form the belief that *Jones' incompetence is a normative reason*. When Will forms that belief, he will be warranted in self-ascribing Jones' incompetence as his motivating reason. And when Will does use RTM to form the belief that Jones' incompetence is his motivating reason, the self-ascription will have been formed by a reliable method.

7.5.2 Cassam's objection

Cassam's argument for denying that attitudes strongly self-intimate does not apply to motivating reasons. Recall that Cassam rightly thinks that a subject can believe that *q* even if she isn't prepared to judge that *q*. The right overall disposition suffices for having a belief, even if the disposition is blocked from ever manifesting. But, we saw in §7.4.1 that motivating reasons differ from attitudes in that regard. In order to be credited with a motivating reason *p* for holding an attitude, S must be prepared to take *p* to be a normative reason for holding the attitude. That is, S must be disposed to believe that *p is a normative reason for holding the attitude*, and this disposition cannot be blocked from manifesting. S must have this unblocked disposition if her motivating reason is to make holding the attitude rational in her lights, and if her motivating

[26] Also, dispositional and some causal accounts would deny that Ellie's belief is based on the *a priori* argument. They would say that a belief about the *a priori* argument initiated her belief that *epistemic externalism is true* but no longer sustains it. Rather, her belief that *epistemic externalism is true* is now sustained by the memory that epistemic externalism is true.

reason is to provide S with rational control over the attitude. If instead the disposition to form the metabelief was blocked, then S would be like Norman as far as she is concerned from her point of view.

7.6 Upshot

I conclude, then, that motivating reasons strongly self-intimate. Consequently, self-knowledge of motivating reasons is special relative to other-knowledge. And interestingly, it is also special relative to self-knowledge of attitudes. Attitudes do not strongly self-intimate. S can believe that *q* without being in a position to know that she believes that *q.* We might want to say that S cannot *rationally* believe that *q* without being in a position to know that she believes that *q.* But even given that, S could still be credited with a *belief* in the relevant cases.

This upshot might seem surprising, but it shouldn't. While S can have a belief that she is not prepared to judge to be true, S cannot have a motivating reason that she is not prepared to take to be a normative reason. And this should be plausible for two reasons.

i. Motivating reasons and attitudes are different kinds of things. After all, they occupy distinct roles. For instance, beliefs guide action. And a state can guide action in accordance with *q* even if the subject lacks the disposition to judge that *q* or the disposition is blocked from being manifested. But motivating reasons fundamentally rationalise subjects' attitudes. It doesn't make sense to call a consideration S's reason for her belief if it doesn't rationalise that belief in her lights or provide her with rational control. And, as we have seen, a consideration can't fulfil the required roles if S is not prepared to take the consideration to be a normative reason for that belief.

ii. We seem to already recognise this fundamental difference between attitudes and motivating reasons in how we talk. It makes sense to speak of S having an 'irrational belief' where this is to say that the belief is bad by S's own lights. But it seems very odd to say that S believes for a reason that is irrational by her own lights. The belief might be irrational from S's perspective in that the reason might be outweighed. But it seems odd to say that the individual reason itself is irrational. The only way it would make sense to talk of an 'irrational reason' is if we wanted to say that it was objectively irrational – that the reason happened to be bad even though S thought otherwise.

7.7 Conclusion

I have argued that self-knowledge of motivating reasons is metaphysically distinctive. Our motivating reasons strongly self-intimate: necessarily, if we have a motivating reason we will be in a position to know that we have it. Indeed, this is the case even though our attitudes do not strongly self-intimate. So, this chapter has the further consequence that self-knowledge of motivating reasons is not just distinctive relative to other-knowledge; it is also distinctive relative to self-knowledge of attitudes.

Chapter 8 Conclusions

Here, I summarise the thesis (§8.1) before elaborating on its main claim: that we have distinctive self-knowledge of our motivating reasons (§8.2). I then identify other important conclusions to have emerged from the thesis (§8.3).

8.1 Recap

This thesis answered the following questions:

> ***Thesis questions:*** *Is self-knowledge of why we have our attitudes and actions a distinctive species of knowledge? In what ways is it/is it not?*

I argued that we have distinctive self-knowledge of why we hold attitudes and perform actions. I argued for this by arguing for a particular account of this type of self-knowledge. My account concerns our motivating reasons in particular – i.e., I argued that subjects have distinctive access to the reasons for which they hold attitudes and perform actions.

In answering the questions, the thesis took the following shape. Chapter two introduced what seem to be the options on the table: the Orthodoxy and agentialism. The Orthodoxy claims that self-knowledge of why we hold our attitudes importantly resembles other-knowledge. Both are acquired in the same way by a process that transitions between evidential contents. We might position this process at the personal level (whereby the subject herself engages in inference) or the subpersonal level (whereby a subsystem performs computation). The Orthodoxy constitutes the mainstream position, and even many who argue that we have distinctive access to other mental features reject distinctive access to why we have our attitudes. Alternatively, agentialism argues that our rational agency grounds distinctive self-knowledge of why we have our attitudes. According to my version of agentialism, subjects learn of their motivating reasons by using the reasons transparency method (RTM) – they answer the question 'why do I have that attitude?' by answering the question 'why have that attitude?'. Chapter three then cleared some ground to show that the thesis questions are open and worth considering further. In chapter four, I presented reason to accept the Orthodoxy – an inference to the best explanation of self-ignorance and confabulation. Chapter five then criticised the Orthodoxy and argued that we have reason not to accept it. This seemed to leave us in an impasse, but chapter six presented my own full account as a solution. I endorse a *two explanations* account. We can and should explain self-

knowledge of motivating reasons at both the subpersonal and personal levels. At the personal level, we should accept the agentialist picture: the subject learns of her motivating reasons for an attitude by employing RTM. At the subpersonal level, this is all underpinned by computational processing. These subpersonal processes importantly resemble those underpinning other-knowledge. Self- and other-knowledge bear some similarities, then, but still fundamentally differ in virtue of the personal-level picture. Chapter seven built on these discussions to argue that self-knowledge of motivating reasons is distinctive in a further important way: our motivating reasons strongly self-intimate. Indeed, this is the case even though our attitudes do not strongly self-intimate. Contra the Orthodoxy, then, our self-knowledge of why we have our attitudes is not just distinctive relative to our knowledge of other people. It is distinctive relative to our self-knowledge of attitudes as well.

8.2 'Self-knowledge of motivating reasons is importantly distinctive'; the main claim discussed

This section elaborates on my answer to the thesis questions. It will: clarify the ways in which self-knowledge of why we have our attitudes is distinctive (§8.2.1); extend the foregoing discussion about reasons for attitudes to motivating reasons generally (§8.2.2).

8.2.1 The four ways in which self-knowledge of motivating reasons is distinctive

Recall from the introduction the list of ways in which we might say that certain types of self-knowledge are distinctive:

1. Self-knowledge is epistemically special in one or both of the following ways:

    1.a. Self-knowledge is especially reliable.

    1.b. Certain features self-intimate.

2. We are in a position to use a distinctive method and warrant to learn of certain features of ourselves.

3. We have first-person authority regarding certain features of ourselves.

4. Some self-knowledge is grounded by our position of agency concerning our attitudes.

I take self-knowledge of our motivating reasons to be distinctive in all four ways. Subsections 8.2.2.1–8.2.1.4 discuss each in turn, although they interconnect.

8.2.1.1 Self-knowledge is epistemically special in one or both of the following ways: a. Self-knowledge is especially reliable; b. Certain features self-intimate

Self-knowledge of motivating reasons bears the second epistemic advantage (b), but not the first (a). As I argued in chapter seven, our motivating reasons strongly self-intimate. I.e., necessarily, if S holds an attitude for the reason that *p*, she will be in a position to know that she holds the attitude for the reason that *p*. That said, I am happy to accept that self-knowledge of motivating reasons is not specially reliable. While subjects cannot have a motivating reason without being in a position to know that they have it, subjects sometimes self-ascribe motivating reasons that they lack.

8.2.1.2 We are in a position to use a distinctive method and warrant to learn of certain features of ourselves

Subjects can use a special method to learn what their motivating reasons are – the reasons transparency method (RTM). For example, I learn why I believe that *it will rain* by considering what normative reasons there are for believing that *it will rain*. I conclude that the grey clouds are a good reason and can thereby know that I believe that *it will rain* for the reason that there are grey clouds. But I can't acquire knowledge of another person's reasons using RTM. RTM would be too unreliable in the other-case to produce knowledge – what comes to my mind as a normative reason will strongly correlate with my motivating reasons, but not yours. As a result, even in cases where using RTM ascribed the correct motivating reason to another person, the ascription wouldn't constitute knowledge.

Self-knowledge acquired using RTM is warranted in a distinctive way as well – in virtue of the subject's agent's awareness of having the motivating reason. The discussion in chapter seven puts us in a better place to say something about how this warrant would look. To state my position baldly, I endorse the following picture. S has an agent's awareness of her motivating reason that *p* in virtue of an agent's awareness of taking *p* to be a normative reason. This is because, in taking *p* to be a normative reason, S partly makes it the case that *p* is her motivating reason and she is aware of this constitutive connection. S's agent's awareness of her motivating reason that *p* then warrants her in self-ascribing the motivating reason that *p*.[1]

[1] This picture is based on O'Brien's account of self-knowledge of judgement in O'Brien (2007, 2003).

Let me give an example to illustrate how self-ascriptions formed using RTM are warranted. Say you ask me why I believe that *it will rain*. To answer your question about myself, I consider the world-directed question 'why believe that *it will rain*?', i.e., 'what are the normative reasons for believing that *it will rain*?'. I conclude that a normative reason for believing that *it will rain* is that there are grey clouds, and then self-ascribe the grey clouds as my motivating reason for believing that *it will rain*. In taking the grey clouds to be a normative reason for believing that *it will rain*, I have an agent's awareness of taking the grey clouds to be a normative reason for believing that *it will rain*. I am *doing something* in taking the consideration to be a normative reason and am aware of what I am doing. I have this awareness because other options are open to me – I could think that the grey clouds are a bad reason instead – and I am aware that these are all options for me as an agent to act on. Furthermore, in being aware of taking the grey clouds to be a normative reason, I am aware of my motivating reason. Part of a subject's understanding of what it is to hold a belief for a reason is that doing so requires being prepared to take the consideration to be a normative reason. In support, we can note the following. We saw in chapter seven various examples suggesting that part of our understanding of a motivating reason is that one must be prepared to take the consideration to be a normative reason. This is reflected in our intuitions about when subjects do and do not have motivating reasons. As a result of all this I have an agent's awareness that I believe that *it will rain* for the reason that there are grey clouds, and this awareness warrants me in self-ascribing the reason. There is, of course, a lot more to say about distinctly agential awareness and how it might apply to this case, but this is all space for further research.[2]

8.2.1.3 We have first-person authority regarding certain features of ourselves

We have first-person authority regarding our explanations of our attitudes and actions. To say that we have 'first-person authority' amounts to the claim that we usually take other peoples' 'word for it' on certain matters concerning themselves. This becomes clearer if we consider cases where we fail to accord individuals first-person authority. Recall the examples from chapter two:

> SEMINAR ONE
>
> Suki: Why do you want to go to the seminar?
>
> Felix: Because it will be interesting.

[2] A lot of the discussions of agential awareness centre around subject's awareness of intentional action (see the Roessler and Eilan (2003) edited collection for an indicative selection of work). We can, and should, also further investigate the experiential awareness involved in holding attitudes on the basis of reasons.

Suki: No, you want to go for the reason that it will help your general philosophical education.

SEMINAR TWO

Suki: Why do you want to go to the seminar?

Felix: No reason.

Suki: No, you want to go for the reason that it will be interesting.

Suki's responses seem to be most peculiar in both these cases. Standardly, we would accept what Felix says without question.

I can now account for first-person authority more fully than before. Earlier I said in a flat-footed way that first-person authority stems from our rational control concerning attitudes and motivating reasons. I.e., my authority concerning my reasons is more like the authority of a parent concerning their child than the authority of a train spotter regarding trains. But material from chapter seven helps flesh this out further (and indeed will show the limitations of the analogy).

Subjects have first-person authority regarding their motivating reasons because motivating reasons strongly self-intimate. Necessarily, if S has a motivating reason that *p*, then S is in a position to know that she has the motivating reason that *p*. And, accordingly, if S is not in a position to learn of some influence on her attitude, then the influence cannot be her motivating reason. The influence may have caused S's attitude, but it will not be the reason for which she holds the attitude. And indeed, it seems plausible that subjects generally use this epistemic position when they have it, and that the relevant considerations don't slip their minds very often. As discussed in chapter two, the fact that *p* will tend to be salient to S when she has the motivating reason that *p.* So, others defer to the S's own word on the matter because that word is effectively a precondition on her having that reason.[3] My account of first-person authority therefore contrasts with that of quasi-perceptual theories of self-knowledge. Quasi-perceptual views say that subjects have first-person authority because their self-ascriptions are especially reliable. But I take it that special reliability doesn't play a role at all (because subjects in fact lack special reliability).

[3] Setiya (2013) also emphasises the constitutive role of answering the question 'why?'.

Let's consider our earlier examples in light of my picture. Felix had first-person authority in both SEMINAR ONE and SEMINAR TWO which Suki should have recognised. In SEMINAR ONE, Felix's epistemic position regarding his motivating reason partly makes it the case that he intends to go to the seminar for the reason that it will be interesting. Of course, Felix will confabulate from time to time, as we all do. But still, his testimony is especially important. That he ascribes the motivating reason is essentially a prerequisite for him having that reason, even though it is not sufficient. And further, in SEMINAR TWO, Felix has first-person authority that he lacks a motivating reason. Unless it had slipped his mind, Felix would have said that he intends to go to the seminar for the reason that it will be interesting. It looks like he isn't in a position to ascribe that consideration using RTM, even if it plays some purely causal role. Accordingly, Felix's 'no reason' answer in effect makes it the case that he has no reason, and Suki should defer to him on this matter. First-person authority stems, then, from the fact that motivating reasons self-intimate, not that self-ascriptions are especially reliable (they aren't).

8.2.1.4 Some self-knowledge is grounded by our position of agency concerning our attitudes.

Our rational agency grounds self-knowledge of why we hold our attitudes. Recall from earlier that we bear an agential relation to our attitudes (e.g., beliefs, desires, hopes). Like playing the recorder and eating a Bakewell tart, believing and desiring are things I can be said to *do*. This position of agency means that I am responsible for my attitudes. The resulting picture is one in which I ought to revise my attitudes in line with normative reasons and have motivating reasons for my attitudes. I will discuss the various ways in which our rational agency grounds self-knowledge in turn. I will say that rational agency places constraints on our account of self-knowledge (i), and then recap the positive picture that our rational agency grounds. This will involve looking at the other ways in which self-knowledge is distinctive: the method (ii), the warrant (iii), self-intimation (iv), and first-person authority (v). I then discuss the place of rational obligation in this account (vi).

i. As a negative point, our position of agency regarding our attitudes precludes the Orthodoxy. Aspects of our rational agency gave rise to two arguments from chapter five against the Orthodoxy. Recall the Dual Role Argument. Here I observed that the question 'why?' concerning our attitudes does not just ask for an explanation; it also requests justification. The questioner takes us to be responsible and accountable for our attitudes and presumes that we ought to be able to justify our attitudes. Recall also the Rational Relations Argument. I argued that our explanation of an attitude bears direct rational relations to that attitude. The explanation we give of an attitude directly affects the rationality of the attitude and vice versa. Again, this is because

we bear responsibility for our attitudes. For the attitude to be rational from our perspective, we must take it to be based on good reasons.

Furthermore, our rational agency doesn't just preclude certain accounts of self-knowledge. It also grounds a positive account. The following discusses the ways in which rational agency does so.

ii. Recall the method underpinning self-knowledge of motivating reasons – RTM. We learn why we have an attitude by answering the question 'why have that attitude?'. We take a given consideration to be a good reason and thereby self-ascribe that consideration as our motivating reason. Importantly, we exercise rational agency in employing RTM. We take a consideration to be a normative reason, partly making it the case that we hold the attitude on the basis of that reason.

And I take self-ascriptions of motivating reasons formed by RTM to be warranted by an agent's awareness of the motivating reason. As discussed before, we exercise rational agency in considering the normative reasons and partly making it the case that we have a motivating reason. This agency then grounds our agents' awareness – an awareness of what we are doing in virtue of being the ones doing it.

iii. Our motivating reasons self-intimate as a result of our rational agency. Our rational agency gives rise to PREMISE ONE of the Main Argument for self-intimation of motivating reasons. To recall the Main Argument from chapter seven:

> PREMISE ONE: Necessarily, if S has a motivating reason that *p* for her attitude *A,* then S is prepared to take that *p* to be a normative reason for having *A.*
>
> PREMISE ONE: Necessarily, if S is prepared to take that *p* to be a normative reason for having *A*, then S is in a position to learn that her motivating reason is *p* using RTM.
>
> CONCLUSION: Necessarily, if S has a motivating reason that *p* for *A*, then S is in a position to learn that her motivating reason is that *p* using RTM.

I argued for PREMISE ONE on the basis of the role that motivating reasons play for us. Holding an attitude for the reason that *p* makes holding that attitude sensible to us, i.e., believing for a reason enables us to believe responsibly. And our motivating reasons enable us to have direct rational control over our attitudes. For our motivating reason that *p* to play those rational roles requires that we are prepared to take *p* to be a normative reason. So, it seems that features that motivating reasons bear in virtue of our agential relation to our attitudes entail, coupled with PREMISE TWO, that our motivating reasons self-intimate.

iv. In turn, rational agency also grounds our first-person authority – recall that our first-person authority importantly stems from the fact that motivating reasons self-intimate.[4]

iv. Recall that according to my agentialist picture, we bear obligations concerning self-knowledge of motivating reasons. It is not just that we can know certain features of ourselves in a special way, but that we *ought* to – we do something wrong if we cannot. Earlier I noted that we bear the:

> *Knowledgeable Reason Explanation* (*KRE*) obligation: The obligation to knowledgeably self-ascribe motivating reasons when explaining one's own attitude.

I am now in a position to say what grounds the KRE obligation. Importantly, the KRE obligation amalgamates two more basic obligations: the obligation to be in a position to know why we hold an attitude, where this is a reason explanation; the obligation to use reason explanations when explaining our attitudes as opposed to purely causal explanations. I will consider these two obligations in turn; both stem from our rational agency and its implications for self-knowledge.

i. We ought to be in a position to self-ascribe motivating reasons for our attitudes. Say I believe that *it will rain*. I ought to be in a position to knowledgeably explain that belief with a reason explanation, such as that I believe that *it will rain* for the reason that there are grey clouds. (This is not to say yet that I actually *should* make the most of this position and form the explanation.) This obligation stems from the (defeasible) obligation to form our attitudes on the basis of reasons. This is because part of being an agent is being responsible for our attitudes, and as such, we can be criticised if we form them willy-nilly. Therefore, if I have a belief that it will rain, I ought to hold it on the basis of a reason. That I ought to have a motivating reason for my attitude means that I also ought to be in a position to know why I have that attitude.[5] This stems from the fact that our motivating reasons self-intimate. Part of satisfying the obligation to have a motivating reason is to be in a position to know that we have it. E.g., if I was not in a position to self-ascribe the grey clouds as a motivating reason for my belief using RTM, then the consideration would not be my motivating reason. The belief that there are grey clouds would at most be a purely causal explanatory reason for my belief that *it will rain*.

ii. We also ought to make use of our special epistemic position and use reason explanations when explaining our attitudes. That is, we ought to use a reason explanation when answering the question 'why?'. It is not just that we ought to be in a position to form knowledgeable reason

[4] On the role of agency in first-person authority, see especially Parrott (2015).
[5] I therefore accept that obligations are at least sometimes closed under entailment, that is, that bearing one obligation can at least sometimes entails also bearing another obligation.

explanations of our attitudes – the KRE obligation requires more than that. Even if we are in a position to form a knowledgeable reason explanation, we still do something rationally wrong if we use purely causal factors when explaining our attitudes. E.g., I exhibit irrationality if I explain my belief that *it will rain* by saying that 'I am a pessimistic person.'

We ought to use reason explanations when explaining our attitudes because of the dual role of the question 'why?'. As I have discussed, the question 'why?' elicits both an explanation and justification of one's attitude. But we can only satisfy the norm of the question 'why?' and our position of responsibility for our attitude by using reason explanations, not purely causal explanations. Reason explanations both explain and justify attitudes, but purely causal explanations only explain.

These two obligations underpin the KRE obligation: the obligation to knowledgeably self-ascribe motivating reasons when explaining one's own attitudes. We ought to be in a position to knowledgeably form reason explanations of our attitudes – we ought to have reasons for our attitudes, and part of having a motivating reason is being in a position to self-ascribe it. And further, we ought to use this reason explanation as opposed to any other so that we can do both things the question 'why?' asks of us – explain a*nd* justify the attitude.

***

I have shown, then, various ways in which our self-knowledge of why we hold our attitudes is distinctive. This is all broadly put, and there is space for research in fully exploring what this distinctiveness amounts to.

8.2.2 The reasons for which we act

The foregoing extends to the reasons for which we act. We have distinctive self-knowledge of our motivating reasons for action which is special in all four ways. We learn of the reasons for which we act using RTM, agent's awareness warrants the self-ascriptions, our reasons for action self-intimate, and we have first-person authority regarding why we act. For example, say I'm going to the shop and bump into Lucas who asks me why I'm doing this. I can learn why I am going to the shop by answering the question 'why go to the shop?'. I conclude that a good reason for going to the shop is that I've run out of hummus and can tell Lucas that I'm going for the reason that I've run out of hummus. This subsection first says why RTM extends to reasons for action (i). Then, since one might doubt this claim in particular, I argue that reasons for action also self-intimate (ii).

i. The RTM account extends to reasons for action because the arguments from chapter five for rejecting the Orthodoxy also apply in the practical case. Recall the Dual Role argument. As with

our attitudes, the question 'why?' plays a dual role regarding our actions. Answering the question involves both explaining and justifying what we are doing. After all, other people often address our self-ascriptions in this context as justifications. It would be perfectly normal for Lucas to offer undercutting or countervailing defeat in response to my answer to the question 'why?': 'but there's hummus in your fridge!' or 'you're eating too much hummus.' And my self-ascription also bears rational relations to the action itself. It would be irrational to sincerely reply that 'I am going to the shop for the reason that I've run out of hummus, but I haven't run out of hummus.' This self-ascription renders my action irrational by my own lights. And, as I argued in chapters five and six, only a method like RTM could account for the dual role of the question 'why?' and the direct rational relations of this sort.

ii. I should also say something about self-intimation since one might in particular deny that reasons for action self-intimate. I take it that we can easily apply the MAIN ARGUMENT to reasons for action. After all, the MAIN ARGUMENT in the case of belief took some of its force by drawing similarities with reasons for action. Our motivating reasons for acting rationalise our actions and provide us with rational control. As such, necessarily, if we *φ* for the reason that *p,* we will be prepared to take *p* to be a normative reason for *φ*ing. Therefore, if we are *φ*ing for the reason that *p,* we will be in a position to use RTM to learn that we are *φ*ing for the reason that *p.*

But one might deny that motivating reasons for action sufficiently resemble those for belief; perhaps only reasons for belief self-intimate. Setiya (2013) argues that it is necessary and sufficient for believing that *q* for the reason that *p* that one believes that *p* is a good reason for believing that *q*. But, Setiya contends, it is not even necessary for *φ*ing for the reason that *p* that one take *p* to be a good reason for *φ*ing (2013, 2010). I.e., the relevant metabelief is necessary and sufficient for believing for a reason, but neither necessary nor sufficient for *φ*ing for a reason. If this is the case, then my MAIN ARGUMENT would apply to reasons for belief but not action. It wouldn't be true that, necessarily, if S *φ*s for the reason that *p*, then S is prepared to take that *p* to be a normative reason for *φ*ing. As such, it wouldn't be the case that, necessarily, if one *φ*s for the reason that *p,* then one is in a position to know that one *φ*s for the reason that *p.*

Setiya provides the following argument for rejecting a doxastic account of *φ*ing for a reason:

> [PREMISE ONE.] It is sufficient to answer the question "Why?" that one has a belief of the form, "I am doing *φ* because *p*," in the sense of "because" that gives an agent's [motivating] reason.
>
> [PREMISE TWO.] That I am doing *φ* because *p*, in this sense, is consistent with the fact that *p* not being a [normative] reason for me to *φ*.

> [PREMISE THREE.] If one proposition is consistent with the negation of another, it is possible to believe the first without believing the second.
>
> So:
>
> [CONCLUSION.] It is possible to believe that I am doing *φ* because *p*, and thus to answer the question "Why?" without believing that the fact that *p* is a [normative] reason for me to *φ* (Setiya 2013: 193).

I dispute PREMISE THREE of Setiya's argument since it begs the question against my conclusion. In certain cases, it *is* impossible for subjects to believe two consistent propositions. *I* cannot believe that *my motivating reason for φing is p* and not believe, when prompted, that *p is a normative reason for φing*. If I was not prepared to believe that *p is a normative reason for φing* in the course of answering the question 'why?', then I cannot be credited with having the motivating reason that *q*. The fact that *p*, or my belief about it, may well have caused me to *φ*. But it is not clear how it would be my motivating reason. There would be something *non-rational* about it and would not play the sorts of *rational* roles that motivating *reasons* to play.[6]

> ***

We can therefore extend my answer to the thesis question to reasons for action.

8.3 Other important ideas from the thesis

So far, I have discussed my main conclusion which concerns the self-knowledge of motivating reasons. But this thesis also provides conclusions that bear on a range of other topics. I discuss four such conclusions in subsections §8.3.1-8.3.4.

8.3.1 The *two explanations* account and self-knowledge of attitudes

I argued for the *two explanations* account in the context of self-knowledge of motivating reasons, but we should extend the account to distinctive self-knowledge generally. That is, we should also explain our self-knowledge of features such as attitudes at both the subpersonal and personal levels. This conclusion is significant: as we have seen, most discussions apart from Carruthers'

[6] Relatedly, recall the combinations of beliefs underpinning Moore paradoxical statements, e.g., when S believes both that *she believes that q* and that *q is false.* S can hold these beliefs. But, she is irrational in doing so, even though the content of the beliefs are consistent with each other.

concern the personal level, and the method the subject herself uses to acquire self-knowledge. To illustrate, I will briefly introduce the *two explanations* account of self-knowledge of belief.

Say, for example, that I know that I believe *it will rain*; we can explain this self-knowledge in two ways. At the personal level, I use a transparency method. I learn whether I believe that *it will rain* by considering the world-directed question 'will it rain?'. I make up my mind and conclude that *it will rain* and can thereby self-ascribe the belief that *it will rain*. In making up my mind and forming my belief that *it will rain*, I have an agent's awareness which warrants my self-ascription. The personal level method will be underpinned by processes in the mindreading and deliberation modules. The processes in the mindreading module will resemble those underpinning other-knowledge and will form the self-ascription by transitioning from contents such as 'this system judges that *it will rain*.' Nevertheless, self-knowledge of belief still fundamentally differs from our knowledge of others' beliefs.

It is not just that we *can* extend the *two explanations* approach to these other types of self-knowledge. We actively *should*, for two reasons. First, recall that we adopted the *two explanations* approach in response to an inference to the best explanation for computationalism/inferentialism about self-knowledge (see chapters four and six). The neo-Ryleans primarily direct that argument at self-knowledge of attitudes. But, as with self-knowledge of reasons, the *two explanations* approach allows us to explain the data while denying neo-Ryleanism proper. The explanation of self-ignorance and error concerning our attitudes would be similar to the explanation I gave in the case of motivating reasons. Subjects self-ascribe attitudes that they in fact lack because computation mechanisms underpinning self-ascription place too much weight on certain pieces of evidence and not enough on others. Second, it is generally plausible that self-knowledge of attitudes would be grounded by the sorts of processes I appeal to. Subjects' use of the transparency method does not float free their minds; it must be grounded by subpersonal processing. Denying this would be implausible.

We have, then, a bold new way of understanding self-knowledge in general. We must bear the *two explanations* approach in mind when investigating self-knowledge, and the approach forms a new research programme in its own right. I had to be brief in setting out the *two explanations* approach and numerous questions remain. How exactly do the subpersonal and personal levels interact? What more can we say about the relation between the subpersonal indexical content and the personal level self-referential content? Are there any exceptions to the *two explanations* approach to self-knowledge? What about self-knowledge of experience and proprioception? Does the need to give an explanation at the subpersonal level place constraints on our personal level picture and vice versa? These are all questions for further research.

8.3.2 Motivating reasons

Chapter six provided an account of what it is to hold attitudes and perform actions on the basis of reasons. Indeed, I should emphasise that I take this to be a unified account of both practical and epistemic motivating reasons. This in itself provides a promising line of further study. While attention is increasingly being paid to how practical and epistemic normativity relate, discussions tend to focus on the nature of normative reasons rather than motivating ones. (See the McHugh et al. 2018 edited collection for an indicative selection of work on this cross over.)

I have argued for two necessary (though not sufficient) conditions on having a motivating reason:

i. *Necessarily, if S has a motivating reason that p, S will be prepared to take p to be a normative reason.* Here I systematised arguments in the literature to argue against non-doxastic accounts of basing: such accounts are incompatible with the rational role that motivating reasons play. Further, I argued that my account improves on classic doxastic views since it is more plausible, psychologically speaking.

ii. *Necessarily, if one has a motivating reason that p, one will be in a position to know that one has the motivating reason that p.* As we saw, the basing literature tends to split into two camps – doxastic and non-doxastic accounts. That is, there is much argument about whether basing requires taking the relevant consideration to be a good reason. But I argued in chapter seven that self-knowledge also plays a role in what it is to have a motivating reason.

8.3.3 Confabulation

I have argued for a novel explanation of confabulation. (By 'confabulation', I mean our mistakes about why we have an attitude or perform an action.) I contend that:

> *We confabulate, and indeed confabulate with the content we do, because we desire to have fulfilled the KRE obligation (i.e. the obligation to knowledgeably explain our attitudes and actions by reference to motivating reasons).*

This explanation, I argued, fares better than other options in the existing literature: that confabulation is motivated by a desire to impress the questioner, or that it results from the fact that subjects form all self-ascriptions using a third-personal method. But, while we should accept that all self-ascriptions are underpinned by some sort of third-personal mechanism, we also need to appeal to motivational factors to explain the data.

8.3.4 A metaphilosophical point

The thesis provides a cautionary tale. Theoretical and empirically-informed philosophy must converse, and each constrains the other. We have seen that agentialists like Moran get something right. They best capture the subject's perspective and our warrant for self-ascriptions. But focusing on abstract issues like transcendental requirements leads thinkers like Moran to pay insufficient attention to empirical constraints. Conversely, Carruthers offers a systematic explanation of self-ignorance and error and considers a range of data. But he fails to take into account the warrant for self-knowledge and the place of epistemic obligations. Fully understanding the human mind requires the input of both philosophy and psychology/cognitive science.

***

The question of how we know why we hold our attitudes and perform actions is live and fruitful. I have hopefully shown that self-knowledge of motivating reasons is importantly distinctive compared to other-knowledge, and indeed, distinctive relative to self-knowledge of attitudes as well. But the conclusions reached in this thesis are just the start; they raise many questions for further research.

List of References

Achinstein, P., 1985. The Nature of Explanation. Oxford University Press, New York.

Alston, W., 1971. Varieties of Privileged Access. American Philosophical Quarterly 8, 223–241.

Alvarez, M., 2017. Reasons for Action: Justification, Motivation, Explanation [WWW Document]. The Stanford Encyclopedia of Philosophy (Winter 2017 Edition), Edward N. Zalta (ed.). URL https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/ (Accessed 2.7.18).

Alvarez, M., 2010. Kinds of Reasons: An Essay in the Philosophy of Action. Oxford University Press, New York.

Anscombe, G.E.M., 2000. Intention, 2nd edition. ed. Harvard University Press, Cambridge, MA; London.

Armstrong, D.M., 2001. A Materialist Theory of the Mind. Routledge, London; New York.

Aydede, M., 2013. Pain [WWW Document]. The Stanford Encyclopedia of Philosophy (Spring 2013 Edition), Edward N. Zalta (ed.). URL https://plato.stanford.edu/archives/spr2013/entries/pain/ (Accessed 3.7.18).

Baier, K., 1965. The Moral Point of View: A Rational Basis of Ethics. Random House, New York.

Bar-On, D., 2007. Review of Akeel Bilgrami, Self-Knowledge and Resentment. Notre Dame Philosophical Reviews 9.

Bar-On, D., 2004. Speaking My Mind: Expression and Self-Knowledge. Oxford University Press, Oxford.

Bermúdez, J.L., 2005. Philosophy of Psychology: A contemporary introduction. Routledge, New York and London.

Bermúdez, J.L., 2000. Personal and Subpersonal; A difference Without a Distinction. Philosophical Explorations 3, 63–82.

Bilgrami, A., 2006. Self-Knowledge and Resentment. Harvard University Press, Cambridge, MA; London.

BonJour, L., 2003. Back to Foundationalism, in: Epistemic Justification: Internalism Vs. Externalism, Foundations Vs. Virtues. Blackwell Publishing, Oxford, pp. 60–76.

BonJour, L., 1985. The structure of empirical knowledge. Harvard University Press, Cambridge, Mass.

BonJour, L., 1980. Externalist Theories of Empirical Knowledge. Midwest Studies in Philosophy 5, 53–73.

Bortolotti, L., Cox, R.E., 2009. 'Faultless' ignorance: Strengths and limitations of epistemic definitions of confabulation. Consciousness and Cognition 18, 952–965.

Boyle, M., 2015. Critical Study: Cassam on Self-Knowledge for Humans. European Journal of Philosophy 23, 337–348.

List of References

Boyle, M., 2011a. Transparent Self-Knowledge. Proceedings of the Aristotelian Society Supplementary Volume 85, 223–241.

Boyle, M., 2011b. “Making up Your Mind” and the Activity of Reason. Philosophers’ Imprint 11, 1–24.

Boyle, M., 2009a. Two Kinds of Self-Knowledge. Philosophy and Phenomenological Research 78, 133–164.

Boyle, M., 2009b. Active Belief. Canadian Journal of Philosophy 39, 119–147.

Boyle, M., n.d. Transparency and Reflection. URL https://tbrunoni.expressions.syr.edu/wp-content/uploads/2013/07/Transparency-and-Reflection-by-Matthew-Boyle.pdf. (Accessed 11.02.2013).

Brown, J., 2007. The Self. Psychology Press.

Burge, T., 1999. A century of deflation and a moment about self-knowledge. Proceedings and Addresses of the American Philosophical Association 73, 25–46.

Burge, T., 1996. Our Entitlement to Self-Knowledge. Proceedings of the Aristotelian Society 96, 91–116. https://doi.org/10.1093/aristotelian/96.1.91

Byrne, A., 2011. Transparency, Belief, Intention. Proceedings of the Aristotelian Society Supplementary Volume 85, 201–221.

Byrne, A., 2005. Introspection. Philosophical Topics 33, 79–104.

Carruthers, P., 2013. The Opacity of Mind: An Integrative Theory of Self-Knowledge. Oxford University Press, Oxford.

Carruthers, P., 2010. Introspection: Divided and Partly Eliminated. Philosophical and Phenomenological Research 80, 76–111.

Carruthers, P., 2006. The Architecture of the Mind. Oxford University Press, Oxford.

Cassam, Q., 2017. What asymmetry? Knowledge of self, knowledge of others, and the inferentialist challenge. Synthese 194, 723–741.

Cassam, Q., 2014. Self-Knowledge for Humans. Oxford University Press, Oxford.

Cassam, Q., 2011. Knowing What You Believe. Proceedings of the Aristotelian Society 111, 1–23.

Cassam, Q., 2010a. Judging, Believing, and Thinking. Philosophical Issues 80–95.

Cassam, Q., 2010b. How We Know What We Think. Revue de Métaphysique et de Morale 4, 553–569.

Chalmers, D.J., 2003. The Content and Epistemology of Phenomenal Belief, in: Smith, Q., Jokic, A. (Eds.), Consciousness: New Philosophical Perspectives. Clarendon Press, Oxford, pp. 220–272.

Chisholm, R., 1982. The Foundations of Knowing. The Harvester Press, Brighton.

Cohen, S., 2002. Basic knowledge and the problem of easy knowledge. Philosophical and Phenomenological Research 65, 309–329.

Coleman, A.M., 2015. A Dictionary of Psychology, 4th edition, lt;http://www.oxfordreference.com/view/10.1093/acref/9780199657681.001.0001/acref-9780199657681> ed. Oxford University Press.

Coliva, A., 2012. Introduction, in: The Self and Self-Knowledge. Oxford University Press, Oxford, pp. 1–14.

Conee, E., Feldman, R., 1998. The Generality Problem for Reliabilism. Philosophical Studies 89, 1–29.

Cox, R., 2018. Knowing Why. Mind & Language 33, 177–197.

Davidson, D., 1963. Actions, Reasons, and Causes. The Journal of Philosophy 60, 685–700.

deRosset, L., 2013. Grounding Explanations. Philosophers' Imprint 13.

Descartes, R., 1644. Principles of Philosophy, in: Cottingham, J., Stoothoff, R., Murdoch, D. (Eds.), The Philosophical Writings of Descartes Volume I. Cambridge University Press, Cambridge, UK.

Douglas, H.E., 2009. Reintroducing Prediction to Explanation. Philosophy of Science 76, 444–463.

Drayson, Z., 2014. The Personal/Subpersonal Distinction. Philosophy Compass 9, 338–346.

Drayson, Z., 2012. The Uses and Abuses of the Personal/Subpersonal Distinction. Philosophical Perspectives 26, 1–18.

Dretske, F., 1969. Seeing and Knowing. Routledge & Kegan Paul, London.

Evans, G., 1982. The Varieties of Reference. Clarendon Press, Oxford; New York.

Evans, I., 2013. The problem of the basing relation. Synthese 190, 2943–2957.

Fernández, J., 2013. Transparent Minds: A Study of Self-Knowledge. Oxford University Press, Oxford.

Fernández, J., 2003. Privileged Access Naturalized. Philosophical Quarterly 53, 352–372.

Finklestein, D., 2003. Expression and the Inner. Harvard University Press, Cambridge, MA.

Fodor, J.A., 1983. The Modularity of Mind. MIT Press, Cambridge, MA.

Fotopoulou, A., 2009. Disentangling the motivational theories of confabulation, in: Hirstein, W. (Ed.), Confabulation: Views from Neuroscience, Psychiatry, Psychology and Philosophy. Oxford University Press, New York, p. Chapter 12.

Gallois, A., 1996. The World Without, the Mind Within: An Essay on First-Person Authority. Cambridge University Press, Cambridge, UK.

Gawronski, B., Hofmann, W., Wilbur, C.J., 2006. Are "implicit" attitudes unconscious? Consciousness and Cognition 15.

Gazzaniga, M.S., 2000. Cerebral specialization and interhemispheric communication Does the corpus callosum enable the human condition? Brain 123, 1293–1326.

Gendler, T.S., 2008a. Alief and Belief. Journal of Philosophy 105(10), 634–663.

Gendler, T.S., 2008b. Alief in Action (and Reaction). Mind and Language 23(5), 552–585.

Gertler, B., 2015. Self-Knowledge [WWW Document]. The Stanford Encylopedia of Philosophy (Summer 2015 Edition). URL http://plato.stanford.edu/archives/sum2015/entries/self-knowledge/

Gertler, B., 2012. Renewed Acquaintance, in: Smithies, D., Stoljar, D. (Eds.), Introspection and Consciousness. Oxford University Press, New York, pp. 89–123.

Gertler, B., 2011. Self-Knowledge. Routledge, London; New York.

Gertler, B., 2001. Introspecting Phenomenal States. Philosophy and Phenomenological Research 63, 305–328.

Gilbert, D.T., Wilson, T.D., 2000. Miswanting: Some problems in the forecasting of future affective states, in: Forgas, J.P. (Ed.), Thinking and Feeling: The Role of Affect in Social Cognition. Cambridge University Press, Cambridge, p. Ch. 8.

Haidt, J., 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review 108, 814–834.

Haidt, J., Bjorklund, F., 2008. Social intuitionists answer six questions about morality, in: Sinnott-Armstrong, W. (Ed.), Moral Psychology, Volume 2: The Cognitive Science of Morality: Intuition and Diversity. The MIT Press, Cambridge, MA, pp. 181–217.

Haidt, J., Björklund, F., Murphy, S., 2000. Moral Dumbfounding: When Intuition Finds No Reason [unpublished manuscript]. URL http://faculty.virginia.edu/haidtlab/articles/manuscripts/haidt.bjorklund.working-paper.when%20intuition%20finds%20no%20reason.pub603.doc. (Accessed 23.03.2015).

Haidt, J., Koller, S.H., Dias, M.G., 1993. Affect, culture, and morality, or is it wrong to eat your dog? Journal of Personality and Social Psychology 65, 613–28.

Hall, L., Johansson, P., Strandberg, T., 2012. Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. PLoS ONE [online] 7, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0045457.

Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., Johansson, P., 2013. How the Polls Can Be Both Spot On and Dead Wrong: Using Choice Blindness to Shift Political Attitudes and Voter Intentions. PLoS ONE [online] 8, http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0060554.

Harman, G.H., 1965. The Inference to the Best Explanation. The Philosophical Review 74, 88–95.

Hieronymi, P., 2011. Reasons for Action. Proceedings of the Aristotelian Society 111, 407–427.

Hieronymi, P., 2009. Two Kinds of Agency, in: O'Brien, L., Soteriou, M. (Eds.), Mental Actions. Oxford University Press, New York, pp. 138–162.

Hieronymi, P., 2008. Responsibility for Believing. Synthese 161, 357–373.

Hieronymi, P., 2006. Controlling Attitudes. Pacific Philosophical Quarterly 87, 45–74.

Hirstein, W., 2009. Confabulation, in: Bayne, T., Cleermans, A., Wilken (Eds.), The Oxford Companion to Consciousness. Oxford University Press, Oxford, pp. 174–77.

Hirstein, W., 2005. Brain Fiction: Self-Deception and the Riddle of Confabulation. MIT Press, Cambridge, MA.

Horgan, T., 2012. Introspection about Phenomenal Consciousness: Running the Gamut from Infallibility to Impotence, in: Smithies, D., Stoljar, D. (Eds.), Introspection and Consciousness. Oxford University Press, New York.

Hornsby, J., 1997. Simple Mindedness: In Defense of Naive Naturalism in the Philosophy of Mind. Harvard University Press.

Huemer, M., 2007. Compassionate Phenomenal Conservatism. Philosophical and Phenomenological Research 74, 30–55.

Hurlburt, R.T., Schwitzgebel, E., 2007. Describing Inner Experience? Proponent Meets Skeptic. MIT Press.

Johansson, P., Hall, L., Sikström, S., Olsson, A., 2005. Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. Science 310, 116–9.

Johansson, P., Hall, L., Sikström, S., Tärning, B., Lind, A., 2006. How something can be said about telling more than we can know: On choice blindness and introspection. Consciousness and Cognition 15, 673–692.

Jones, E.E., Nisbett, R.E., 1972. The actor and the observer: Divergent perceptions of the cause of behavior, in: Jones, E.E., Kanhouse, D.E., Kelley, H.H., Nisbett, R.E., Valins, S., Weiner, B. (Eds.), Attribution: Perceiving the Causes of Behavior. General Learning Press, Morristown, pp. 79–94.

Jones, W.E., 2002. Believing and Doxastic Instability. Philosophical Studies 111, 217–249.

Jongepier, F., Strijbos, D., 2015. Introduction: self-knowledge in perspective. Philosophical Explorations 18, 123–133.

Kahneman, D., 2012. Thinking, Fast and Slow. Penguin Books, London.

Kant, I., 1958. Critique of Pure Reason. MacMillan & Co Ltd, London.

Knobe, J.M., Malle, B.F., 2002. Self and Other in the Explanation of Behavior: 30 Years Later. URL https://campuspress.yale.edu/joshuaknobe/files/2016/02/PsyBel-17odk82.pdf. (Accessed 24.08.2016).

Leite, A., 2008. Believing one's reasons are good. Synthese 161.

Leite, A., 2004. On Justifying and Being Justified. Philosophical Issues 14, 219–253.

Lipton, P., 2004. Inference to the Best Explanation, 2nd edition. ed. Routledge, London; New York.

Locke, J., 1689. An Essay Concerning Human Understanding. Oxford University Press, Oxford.

Lopes, D.M., 2014. Feckless Reason, in: Currie, G., Kieran, M., Meskin, A., Robson, J. (Eds.), Aesthetics and the Sciences of Mind. Oxford University Press.

Lord, E., Sylvan, K., n.d. Believing for Normative Reasons: Prime, Not Composite.

Lycan, W., 1996. Consciousness and Experience. MIT Press/ Bradford Books, Cambridge, MA.

Lycan, W.G., 1995. Consciousness as Internal Monitoring, I: The Third Philosophical Perspectives Lecture. Philosophical Perspectives 9, 1–14.

Macdonald, C., 2014. In My "Mind's Eye": Introspectionism, Detectivism, and the Basis of Authoritative Self-Knowledge. Synthese 1–26.

Macdonald, C., 1998. Externalism and Authoritative Self-Knowledge, in: Wright, C., Smith, B.C., Macdonald, C. (Eds.), Knowing Our Own Minds. Clarendon Press, Oxford, p. Ch. 5.

Mackonis, A., 2013. Inference to the best explanation, coherence and other explanatory virtues. Synthese 190, 975–995.

Malle, B.F., 2011. Time to Give Up the Dogmas of Attribution: An Alternative Theory of Behavior Explanation. Advances in Experimental Social Psychology 44, 297–352.

Malle, B.F., Knobe, J.M., Nelson, S.E., 2007. Actor–Observer Asymmetries in Explanations of Behaviour: New Answers to an Old Question. Journal of Personality and Social Psychology 93, 491–514.

McCain, K., 2012. The Interventionist Account of Causation and the Basing Relation. Philosophical Studies 159, 357–382.

McHugh, C., 2017. Attitudinal control. Synthese 194, 2745–2762.

McHugh, C., 2013. Epistemic Responsibility and Doxastic Agency. Philosophical Issues 23, 132–157.

McHugh, C., Way, J., 2016. Against the taking condition. Philosophical Issues 26, 314–331.

McHugh, C., Way, J., Whiting, D., 2018. Normativity: Epistemic and Practical. Oxford University Press.

McKay, R., Kinsbourne, M., 2010. Confabulation, delusion, and anosognosia: Motivational factors and false claims. Cognitive Neuropsychiatry 15, 288–318.

Mele, A., 2001. Self-deception unmasked. Princeton University Press, Princeton.

Mercier, H., Sperber, D., 2011. Why do humans reason? Arguments for an argumentative theory. Behavioral and Brain Sciences 34, 57–111.

Millar, A., 2010. Knowledge from Indicators, in: Pritchard, A., Millar, A., Haddock, A. (Eds.), The Nature and Value of Knowledge: Three Investigations. Oxford University Press, Oxford, pp. 144–163.

Moore, G.E., 1942. A reply to my critics, in: Schilpp (Ed.), The Philosophy of G.E. Moore. Northwestern University, Evanston, IL.

Moore, M.T., Fresco, D.M., 2012. Depressive realism: A meta-analytic review. Clinical Psychology Review 32, 495–509.

Moran, R., 2012. Self-Knowledge, “Transparency”, and the Forms of Activity, in: Smithies, D., Stoljar, D. (Eds.), Introspection and Consciousness. Oxford University Press, New York, p. Ch. 8.

Moran, R., 2004. Replies to Heal, Reginster, Wilson, and Lear. Philosophy and Phenomenological Research 69, 455–472.

Moran, R., 2003. Responses to O’Brien and Shoemaker. European Journal of Philosophy 11(3), 402–419.

Moran, R., 2001. Authority and Estrangement: An Essay on Self-Knowledge. Princeton University Press, Princeton; Oxford.

Moser, P.K., 1985. Empirical Justification. Springer.

Nanay, B., 2011. Do We Sense Modalities with Our Sense Modalities? Ratio 24, 299–310.

Neta, R., n.d. Basing and Treating. URL https://philosophy.unc.edu/files/2014/06/Basing-and-Treating1.pdf. (Accessed 09.05.2017).

Nisbett, R., Ross, L., 1980. Human Inference: Strategies and Shortcomings of Social Judgement. Englewood Cliffs, N.J.

Nisbett, R.E., Wilson, T.D., 1977. Telling More Than We Can Know: Verbal Reports on Mental Processes. Psychological Review 84, 231–259.

O'Brien, L., 2007. Self-Knowing Agents. Oxford University Press, Oxford; New York.

O'Brien, L., 2005. Self-Knowledge, Agency and Force. Philosophy and Phenomenological Research 71(3), 580–601.

O'Brien, L., 2003. Moran on Agency and Self-Knowledge. European Journal of Philosophy 113, 375–390.

Parrott, M., 2017. Self-Blindness and Self-Knowledge. Philosophers' Imprint 17, 1–22.

Parrott, M., 2015. Expressing first-person authority. Philosophical Studies 172, 2215–2237.

Peacocke, C., 1998. Conscious Attitudes, Attention, and Self-Knowledge, in: Wright, C., Smith, B.C., Macdonald, C. (Eds.), Knowing Our Own Minds. Clarendon Press, Oxford, pp. 64–98.

Pitt, D., 2004. The Phenomenology of Cognition or "What Is It like to Think That P?" Philosophy and Phenomenological Research 69, 1–36.

Plantinga, A., 1993. Warrant and Proper Function. Oxford University Press, New York.

Prinz, J.J., 2006. Is the mind really modular?, in: Stainton, R.J. (Ed.), Contemporary Debates in Cognitive Science. Blackwell, pp. 22–36.

Pronin, E., 2007. Perception and misperception of bias in human judgment. Trends in Cognitive Sciences 11, 37–43.

Pronin, E., Gilovich, T., Ross, L., 2004. Objectivity in the eye of the beholder: divergent perceptions of bias in self versus others. Psychological Review 111, 781–799.

Pronin, E., Kugler, M.B., 2007. Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. Journal of Experimental Social Psychology 43, 565–578.

Pronin, E., Lin, D.Y., Ross, L., 2002. The Bias Blind Spot: Perceptions of Bias in Self Versus Others. Personality and Social Psychology Bulletin 28, 369–381.

Pryor, J., 2000. The Skeptic and the Dogmatist. Noûs 34, 517–49.

Quinn, W., 1993. Putting Rationality in its Place, in: Quinn, W., Foot, P. (Eds.), Morality and Action. Cambridge University Press, Cambridge, UK, pp. 228–255.

Reed, B., 2011. Certainty [WWW Document]. The Stanford Encyclopedia of Philosophy (Winter 2011 Edition), Edward N. Zalta (ed.). URL https://plato.stanford.edu/archives/win2011/entries/certainty/ (Accessed 3.7.18).

Rescorla, M., 2017. The Computational Theory of Mind [WWW Document]. The Stanford Encyclopedia of Philosophy (Spring 2017 Edition), Edward N. Zalta (ed.). URL https://plato.stanford.edu/archives/spr2017/entries/computational-mind/

Reutlinger, A., 2017. Explanation beyond causation? New directions in the philosophy of scientific explanation. Philosophy Compass 12, 1–11.

Roessler, J., Eilan, N. (Eds.), 2003. Agency and Self-Awareness: Issues in Philosophy and Psychology. Oxford University Press, New York.

Russell, B., 1917. Knowledge by Acquaintance and Knowledge by Description, in: Mysticism and Logic. George Allen and Unwin, London.

Ryle, G., 2009. The Concept of Mind. Routledge, London.

Sandis, C., 2015. Verbal Reports and 'Real' Reasons: Confabulation and Conflation. Ethical Theory and Moral Practice 18.

Sartre, J.-P., 1956. Being and Nothingness. Philosophical Library, New York.

Scaife, R., 2014. A Problem for Self-Knowledge: The Implications of Taking Confabulation Seriously. Acta Analytica 29, 469–485.

Scanlon, T.M., 1998. What We Owe to Each Other. The Belknap Press of Harvard University Press, Cambridge, MA; London.

Schroeder, M., 2015. Knowledge Is Belief For Sufficient (Objective and Subjective) Reason, in: Gendler, T.S., Hawthorne, J. (Eds.), Oxford Studies in Epistemology Volume 5. Oxford University Press, pp. 226–250.

Schwitzgebel, E., 2012. Introspection, What?, in: Smithies, D., Stoljar, D. (Eds.), Introspection and Consciousness. Oxford University Press.

Schwitzgebel, E., 2009. Knowing Your Own Beliefs. Canadian Journal of Philosophy 35, 41–62.

Schwitzgebel, E., 2008. The Unreliability of Naive Introspection. Philosophical Review 117, 245–273.

Schwitzgebel, E., n.d. Self-Ignorance. URL www.faculty.ucr.edu/~eschwitz/SchwitzPapers/SelfUcs-101118.pdf. (Accessed 25.02.2015).

Setiya, K., 2013. Epistemic Agency: Some Doubts. Philosophical Issues 23, 179–198.

Setiya, K., 2010. Sympathy for the Devil, in: Tenenbaum, S. (Ed.), Desire, Practical Reason, and the Good. Oxford University Press, pp. 82–110.

Shermer, M., 2012. The Believing Brain: From ghosts and Gods to Politics and Conspiracies - How We Construct Beliefs and Reinforce Them as Truths. St Martin's Press, New York.

Shoemaker, S., 2012. Self-Intimation and Second-Order Belief, in: Smithies, D., Stoljar, D. (Eds.), Introspection and Consciousness. Oxford University Press, Oxford; New York, p. Ch. 9.

Shoemaker, S., 1994. Self-Knowledge and "Inner Sense." Philosophical and Phenomenological Research 54, 249–314.

Siegel, S., 2012. Cognitive Penetration and Perceptual Justification. Noûs 46, 201–222.

Siegel, S., 2009. The Visual Experience of Causation. The Philosophical Quarterly 59, 519–540.

Siegel, S., 2006. Which Properties are Represented in Perception, in: Gendler, T.S., Hawthorne, J. (Eds.), Perceptual Experience. Oxford University Press, New York.

Siewert, C.P., 1998. The Significance of Consciousness. Princeton University Press, Princeton.

Smithies, D., 2012. A Simple Theory of Introspection, in: Smithies, D., Stoljar, D. (Eds.), Introspection and Consciousness. Oxford University Press, New York, pp. 259–293.

Smithies, D., Stoljar, D., 2012. Introspection and Consciousness: An Overview, in: Introspection and Consciousness. Oxford University Press, New York, pp. 3–26.

Sosa, E., 2007. A Virtue Epistemology: Apt Belief and Reflective Knowledge, Volume I. Oxford University Press.

Stern, R., 2015. Transcendental Arguments [WWW Document]. The Stanford Encyclopedia of Philosophy (Summer 2015 Edition). URL <http://plato.stanford.edu/archives/sum2015/entries/transcendental-arguments/>

Swain, M., 1985. Justification, Reasons and Reliability. Synthese 64, 69–92.

Swain, M., 1981. Reasons and Knowledge. Cornell University Press, Ithaca, NY.

Sylvan, K., 2016a. Epistemic Reasons I: Normativity. Philosophy Compass 11, 364–376.

Sylvan, K., 2016b. Epistemic Reasons II: Basing. Philosophy Compass 11, 377–389.

Thagard, P.R., 1978. The best explanation: Criteria for theory choice. Journal of Philosophy 75, 76–92.

Turner, R.N., Hewstone, M., 2009. Attribution Biases. Encyclopedia of Group Processes & Intergroup Relations.

Turri, J., 2011. Believing For a Reason. Erkenntnis 383–397.

Velleman, J.D., 1985. Practical Reflection. The Philosophical Review 94, 33–61.

Vogel, J., 2000. Reliabilism Levelled. The Journal of Philosophy 97, 602–623.

Wegner, D.M., 2002. The Illusion of Conscious Will. MIT Press, Cambridge, MA.

Whiting, D., 2014. Epistemic Norms: New Essays on Action, Belief, and Assertion, in: Littlejohn, C., Turri, J. (Eds.), Epistemic Norms: New Essays on Action, Belief, and Assertion. Oxford University Press, Oxford, pp. 219–237.

Williamson, T., 2002. Knowledge and its Limits. Oxford University Press.

Wilson, T. de C., Nisbett, R.E., 1978. The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. Social Psychology 41, 118–31.

Wilson, T.D., 2002. Strangers to Ourselves. Belknap, Cambridge, MA; London.

Wilson, T.D., Dunn, D.S., Kraft, D., Lisle, D.J., 1989. Introspection, Attitude Change, and Attitude-Behavior Consistency: The Disruptive Effects of Explaining Why We Feel the Way We Do. Advances in Experimental Social Psychology 22, 287–343.

Wilson, T.D., Dunn, E.W., 2004. Self-knowledge: Its Limits, Value, and Potential for Improvement. Annual Review of Psychology 55, 493–518.

Wittgenstein, L., 1953. Philosophical Investigations. Oxford.

Woodward, J., 2017. Scientific Explanation. The Stanford Encyclopedia of Philosophy (Fall 2017 Edition).

Zagzebski, L., 2003. The Search for the Source of Epistemic Good. Metaphilosophy 34, 12–28.