The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/172884

Please be advised that this information was generated on 2020-01-17 and may be subject to change.
The Circumstances of Self-Knowledge

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op dinsdag 23 mei 2017
om 14.30 uur precies

door

Fleur Jongepier

geboren op 20 oktober 1986 te Eindhoven
Promotoren:
Prof. dr. M.V.P. Slors
Prof. dr. J.A.M. Bransen
Prof. dr. Q. Cassam

Manuscriptcommissie:
Prof. dr. L.B.W. Geurts
Prof. dr. E. Schwitzgebel
Prof. dr. M. Düwell

Work on this thesis was funded by the Netherlands Organization for Scientific Research (NWO). Research project no. 322-20-003.
## Contents

Acknowledgements

1. Introduction  
   1. The Gist of the Thesis: Atomism versus Holism 13  
   2. Qualifications, Assumptions and Limitations 18  
   3. Plan for the Thesis 21

PART 1: Expressivism

2. Knowing Your Mind by Speaking Your Mind 27  
   1. Introduction 27  
   2. Expressivism: The Basics 28  
   3. Bar-On’s Neo-Expressivism 32  
   4. Expressivism as a Theory of Self-Knowledge 35  
   5. Avowals as Expressing Beliefs 38  
   6. Avowals as Expressing Justified Beliefs 40  
   7. Avowals as Expressing True Beliefs 42  
   8. Some Prima Facie Worries 46  
   9. Conclusion 50

3. The Limits of Expressivism: Failures of Self-Expression 53  
   1. Introduction 53  
   2. Ceteris Paribus I: Assuming I Am Avowing 55  
   3. Ceteris Paribus II: Assuming I Am Not Severely Intoxicated or Asleep 64  
   4. Self-Expression as Self-Interpretation 72  
   5. Diagnosis: Self-Expression and Communication 76  
   6. Avowals as Assertions: Theoretical Advantages 82  
   7. Conclusion 87
PART 2: Rationalism

4. Knowing Your Mind by Making Up Your Mind
   1. Introduction
   2. The Agential Account of Self-Knowledge
   3. The Rationalist Account of Self-Knowledge
      3.1 The Transparency Procedure
      3.2 Alienation
   4. Is the Transparency Procedure Necessary and Sufficient for
      Self-Knowledge?
      4.1 Common Objections to Rationalism
      4.2 Possible Responses
   5. A Transcendental Take on Transparency?
   6. Is the Transparency Procedure an Inferential Procedure?
   7. Conclusion

5. The Limits of Rationalism: Failures of Transparency
   1. Introduction
   2. Radical Rationalism and the Anscombean Constraint
      2.1 The Anscombean Constraint
      2.2 Radical Rationalism
   3. Prima Facie Challenges and Problems for Radical Rationalism
   4. Moderate Rationalism
   5. The 'Filtering Strategy' as a Response to the 'Garbage in,
      Garbage out' Objection
      5.1 Why it Filters Out Too Much
      5.2 Why 'Considered Reasons' are Neither Necessary Nor Sufficient
         for Self-Knowledge
   6. Holist Rationalism: The Self as Part of the World?
   7. Conclusion

6. The Limits of Activism
   1. Introduction
   2. Boyle's Epistemic View: Reflectivism
   3. Boyle's Metaphysical View: Activism
4. Radical Versus Moderate Dispositionalism 154
5. The Activist Argument Against Dispositionalism 160
   5.1 Prima Facie Doubts Against Dispositionalism 160
   5.2 The Principal Objection: Meeting 'Boyle's Constraint' 162
6. Response to Boyle's Constraint 163
7. The Apparent Gap Between Judgement and Belief 168
   7.1 The Implicit Racist, the Unethical Professor and the Family Man 168
   7.2 The 'No Proper Judgement' Response 171
8. The 'What Gap?' Response 173
9. Meta-Theoretical Questions about Metaphysics of Mind 177
10. Conclusion 182

Appendix: A Dispositionalist Account of Self-Knowledge and an Activist Account of First-Person Authority 185

7. The Limits of Rationalism: Failures of Autonomy 189
   1. Introduction 189
   2. Rationalism as a Normative Project 192
   3. Deliberation in Oppressive Circumstances: Three Cases 198
   4. Having a Mind of One's Own 203
   5. Procedural Accounts of Autonomy 206
   6. Relational Accounts of Autonomy 210
   7. Implications for Self-Knowledge 215
      7.1 The Simple View 215
      7.2 A Nuanced View: Types of Self-Knowledge 219
   8. The Metaphysical and Epistemic Sources of Atomism 221
   9. Conclusion 229

8. Taking Stock and Looking Forward 233

Bibliography 239
Samenvatting 253
Curriculum Vitae 263
I have had the luxury of having been supervised by no less than three excellent philosophers: Marc Slors, Jan Bransen, and Quassim Cassam. I am grateful to all three of them for both their enormously helpful intellectual-cum-academic guidance, but also their less-strictly-academic support.

Marc, I am immensely thankful for your patience with my changes of mind and heart, for your inexhaustible supply of optimism and enthusiasm and for all the discussions we had that were directly or indirectly thesis-related (as well as those that were completely thesis-unrelated). Also, the way in which philosophy is practiced in Nijmegen, I came to experience, is constructive and respectful; an atmosphere of thinking-with rather than thinking-against. You play an important role in creating this atmosphere, and I am very grateful for being able to look back at many wonderful years in Nijmegen.

Jan, I want to thank you for, well, being you. You would always ask the questions I was hoping no one would ask, and would press me on the (ir)relevance of certain philosophical issues I was working on. You would warn me against scholastic swamps, strawmen and various cans of worms. I've come to accept, indeed embrace, my own version of Jan popping up in inner speech, asking me "Who cares?" every now and again. I can only hope my inner Jan will remain with me for a long while.

Quassim, I want to thank you for many stimulating discussions, for your continuing support, for encouraging me find things out on my own, for your flexibility when I avowed my intentions about a thesis outline one moment and released myself of my resolutions the next (so much the worse for rationalism about self-knowledge), for writing Self-knowledge for Humans and for doing Philosophy for Humans. I am grateful for your overall kindheartedness and hospitably (and cooking skills) in the countryside, of which I have very fond memories.

Marc, Jan and Quassim come from rather different areas of philosophy (roughly; mind, ethics and epistemology, respectively), and the three of them haven't been in a room together up until the defence. Nonetheless, their supervision felt 'unified'. For
Acknowledgements

One thing, I gradually found out that what I most want to do is a kind of philosophy that is located precisely on the intersections of their respective specialisations, and I learned that this was a methodological outlook with respect to which all three were very encouraging. Second, an important thing my supervisors seem to share is a general approach to what philosophy is, why it matters, and what it is philosophers ought to be doing. I’m happy to say that their view and way of doing philosophy has very much shaped my thinking and I’m sure will continue to have its effect on me in the future.

I furthermore want to thank my colleagues and friends in- and outside of the Nijmegen philosophy of mind & language group (Leon, Derek, Daphne, Roy, Katia, Frank, Sammie, Leopold, Bart, Corien, Kees and all the rest) for brown bag lunches, serious and not-so-serious chats, bouldering uijjes (and putting up with my coffeesnobbery). A special word of thanks to my officemate Chris, not only for often kindly informing about ‘my thesis and my cats’ (which, he rightly thought, were roughly at the same level), but also for taking care of – and eventually adopting – the orchid that flowered so explosively over the years only thanks to his care.

Thanks are due also to colleagues in other departments of philosophy in the lowlands. Joel Anderson, Henk van Gils, Caroline Harnacke, Annemarie Kalis, Naomi Kloosterboer, Anthonie Meijers, Jeroen de Ridder, Ingrid Robins, Katrien Schaubroek and Naomi van Steenbergen deserve special mention.

I spent some time at the faculty of philosophy at Warwick during my PhD, and I want to thank Johannes Roessler, Naomi Eilan and, again, Quassim (at the time not yet my supervisor) for stimulating discussions, and Irina, Roberta and other PhDs and postdocs for a wonderful time (who would have thought one could have a good time in Coventry, of all places!).

Special thanks go out to my so-called ‘paranimfen’: Priscilla and Beatrijs. I look forward to being flanked by my two ‘best women’, who I am happy to consider not just as my (ex-)colleagues, but as close friends. (Beatrijs: I’m looking forward to the both of us being on the ‘other side’ next year! Priscilla: I genuinely miss your witty comments and questions – academic philosophy is worse off without you.)
I want to thank my family and my family-in-law for their loving support, putting up with last-minute cancellations, and for taking care of Japi & Pep when having to take much needed breaks from academia.

Let me end by thanking the person who knows me better than I know myself, who was unrelenting in raising critical questions but equally unrelenting in helping me find answers to them and getting me back up on my feet, and, who, most important of all, taught a hard learner that life isn’t all about philosophy. Sem, you make not doing philosophy so much fun.¹

¹ And I don’t just mean I enjoyed watching all eighty-six episodes of Spooks (or indeed, all eighty-six episodes of The Sopranos) with you. Anyway, as you may have feared, this footnote is only here for rhetorical purposes. Given your alter ego of ‘Mr. Footnote’, I could not give more credit to your influence on me had I not ended these Acknowledgements, or indeed began my thesis, with a bit of teasing in a footnote devoted to you.
I

Introduction

1. The gist of the thesis: atomism versus holism

Do you have self-knowledge of your intention to be at work on time tomorrow if you’ve had a few drinks too many? Do you know that you want a divorce if you express your desire during a fit of anger? Do you know your own desires, hopes and beliefs if you’re depressed, insecure or got out of bed on the wrong side? Do you have self-knowledge of your desire to buy a healthy quinoa salad rather than fish and chips for lunch if it’s evident that you’ve been ‘nudged’ into doing so? Do you have self-knowledge of your belief that having a baby boy is better than having a baby girl if you’ve been manipulated by state propaganda?

These are fundamental and difficult questions about self-knowledge and about what it means to know your own beliefs, desires, hopes, intentions and other attitudes. The answers to these questions are by no means obvious. The aim of this thesis is not to provide a concrete answer to such questions, but rather to ask what is required in order to answer them. More specifically, the aim is to explore whether contemporary philosophical theories of self-knowledge have the materials to handle such questions in a satisfactory manner, i.e. in a way that respects their subtleties and intricacies.

The current philosophical debate on self-knowledge is mostly concerned with the question of whether self-knowledge is a matter of looking into our minds (introspectionism), interpreting our minds (interpretivism), speaking our minds (expressivism) or making up our minds (rationalism). Hence, when working on self-knowledge, one of the first questions one is confronted with is whether one is an ‘introspectionist’, ‘interpretationist’, ‘expressivist’, ‘rationalist’ or perhaps a ‘pluralist’ of some sort.1 In other words, much of the current self-knowledge

1 According to introspectionist or ‘hybrid’ introspectionist accounts, self-knowledge is a matter of inner looking or ‘detecting’ one’s own mental states (e.g. Armstrong 1968; Lycan 1996; Goldman 1993; 2006; Nichols and Stich 2003; Rosenthal 2005; Peels forthcoming). Self-ascriptions or second-order beliefs about one’s first-order state
debate focuses on the methods or procedures of self-knowledge – their differences, similarities and (in)compatibilities.2

The focus of this thesis will instead be on what appears to be a widely shared though implicit assumption, which I’ll refer to as the assumption of atomism regarding self-knowledge. This is an assumption about what is required for someone to acquire knowledge of her attitudes. The atomist assumption comes down to this: following the theorist’s preferred procedure or method (such as speaking or making up one’s mind) is sufficient for a subject to acquire knowledge of her attitudes. The alternative ‘holist’ approach can be understood negatively as the reverse of atomism: the mere following of any of these standard methods or procedures by itself does not guarantee that they will yield self-knowledge.

The plan is to articulate and problematize the atomist approach to the question of self-knowledge. More specifically, my aim is to address the preconditions of self-knowledge, which I will address by asking under what circumstances following some particular method is actually knowledge-conducive. In so doing I will concentrate, specifically, on the expressivist and rationalist accounts.

The atomist view as I have just described it is a caricature, which, I take it, hardly anyone will be happy to defend. In recognition of this fact, I distinguish between two types of atomism. According to what I’ll call a radical atomist approach to self-knowledge, the question of whether the ascription of a belief or other attitude to oneself counts as self-knowledge is independent of the circumstances under which one followed method X: there are no ‘bad’ circumstances of self-knowledge. So even someone who is depressed, in a fit of anger, manipulated by state propaganda, under the influence of mind-altering substances, hypnotized or tortured still has knowledge of her attitudes, as long as she arrived at her self-ascription in the ‘right’ way.

are understood as ‘descriptions’ or ‘reports’ of underlying states. Interpretationists and/or inferentialists think that one’s second-order belief is the result of an (unconscious or sub-personal) process of self-interpretation, a process of ‘first-person mindreading’ or a process of inferences made on the basis of evidence (Berm 1972; Gopnik 1993; Carruthers 2000; 2011; Schwitzgebel 2001; 2010; 2012; 2013; 2016; Lawlor 2003; 2009; Cassam 2010; 2011; 2014; 2015). Expressivism, rationalism and interpretivism will be explained in detail in the chapters to come. I will not discuss pluralist accounts in this thesis, but see, for example, Andrews (2015), Samoilova (2015) and Coliva (2016).

2 A brief note on terminology: the notion of a method or procedure can be understood either in a narrow way, as referring to a way of doing something step by step or on the basis of a pre-established manual or recipe or, more broadly, as ‘a way of doing something’ or an ‘act or a manner of proceeding in any action or process’. I shall understand the notions of methods or procedures in the latter, broader, sense, i.e. as a way of doing something, more specifically as a way of making or arriving at a self-ascription or coming to judge or coming to believe that one is that P.
This view is a caricature, I think, because most philosophers working on self-knowledge will allow, implicitly or explicitly, that there are at least certain circumstances in which a self-ascription does not count as self-knowledge even if the appropriate method is followed. If, say, a ‘mentalist’ like Derren Brown were to hypnotize you into thinking you’re a chicken, and you go on to think you need to brood your eggs, then, in some sense at least, you lacked knowledge of your beliefs, desires or intentions, such as not wanting to be made a fool of, the belief that you are not a chicken, the desire to be unhypnotized, and so on. The sort of self-knowledge you have while hypnotized is only knowledge of your occurrent thoughts.\(^3\)

Hence, I shall distinguish radical atomism from what I call moderate atomism, which can be understood as a ‘hedged’ version of atomism. It is moderate atomism that I shall be centrally concerned with. A moderate atomist claims that even though there are certain extreme circumstances in which a self-ascription does not count as self-knowledge, in ‘normal’, ‘standard’ or the ‘right’ conditions, the proposed method is indeed knowledge-conducive, i.e. will allow the subject to know what her attitudes are. The crux of the difference, then, between moderate atomism and holism is that even the moderate atomist agrees with the holist that radical atomism is mistaken; s/he thinks that the success of some or all of the above-mentioned methods can be said to hold \(\textit{ceteris paribus}\), that is, all else being equal, right or normal, or in the absence of countervailing forces or disturbing factors. Whereas a moderate atomist thinks that we can talk meaningfully about ‘exceptions to the rule’, I will try to make a case for the idea that in the end, the exceptions are the rule.

The moderate atomist thus proceeds to concern herself with ‘matters of method’ and questions regarding procedures by appealing to certain \(\textit{ceteris paribus}\) clauses. This, in turn, allows her to deal with, i.e. set aside, the sort of cases mentioned at the beginning of this introduction. We can ignore angry spouses, tipsy colleagues and depressed friends in our philosophical theories of self-knowledge because they’re the exception to the rule, and they do not generalize in such a way that would require abandoning moderate atomism and moving towards a holist approach. The main challenge for the moderate atomist is to appeal to some non-arbitrary cut-off point between normal and abnormal circumstances and to define the right circumstances in a way that does not beg the question, i.e. does not define countervailing forces in terms of self-ignorance or self-deception.

---

3 I say a bit more about the thought/attitude distinction below.
It's important to stress from the outset that I will not argue for the alternative holist account of self-knowledge directly, only indirectly by arguing against moderate atomism. I will argue that the moderate atomist's appeal to normal circumstances is hard to maintain, and that, in the end, the (implicit) ceteris paribus clauses end up doing all of the work to determine the question of whether or not someone has self-knowledge. The argumentative strategy pursued in this thesis, then, is to show that moderate atomism is not a stable position and that it collapses into either radical atomism or holism.

The notions of 'atomism' and 'holism' will probably be familiar to most readers, given that these labels are often used in other areas of philosophy. The distinction between the two has, for instance, been applied to issues surrounding theories of meaning (Fodor and Lepore 1992; Hutto 2008), personal identity (Schechtman 1990; Slors 2001), personhood and individual agency (Hobbes 1651; Pettit 1996; 1996; Stoljar 2015), moral responsibility (Vargas 2013) and (moral) reasons for action (Dancy 2004; 2007). As I will understand these terms, atomism and holism are labels used to specify the role that circumstances play.\[4\]

Applied to self-knowledge, atomism can have either of two sources. One is to be a 'constructivist' and think that some procedure or method M (e.g. making up one's mind) constitutes one's attitudes and that by constituting one's attitudes in this way, one automatically comes to know about them. A radical atomist constructivist thinks that by following M one always constitutes one's attitudes one thereby knows

---

4 Dancy gives the following description of the distinction, as applied to (moral) reasons for action, which is in line with the present project: "Atomism holds that any feature that is a reason in favour of action in one case will always be a reason in favour of action wherever it occurs. The same feature always makes the same reason; or, a reason is a general reason. This theory is false; something that is a reason in favour of action in one case may in another case be no reason at all, or even a reason against action. It all depends on the circumstances; reasons are sensitive to context" (Dancy 2007, 80). Manuel Vargas draws the distinction, as applied to moral responsibility, along similar lines: "Let atomism refer to the view that free will is a non-relational property of agents, that is, it is characterizeable in isolation from broader social and physical contexts. ... (atomistic) accounts specify some property—say, a real self, an uncaused event, the presence of reasoning capacities, or what have you—that, at least in principle, one could identify simply by looking inside the agent: If the relevant feature is there, then the agent has free will, independent of context" (Vargas 2013, 204–5). He describes the alternative view as follows: “[The alternative] picture suggests is that we cannot answer the question of whether an agent has free will simply by looking at the agent. What we need to know are facts about both the agent and the circumstances. On this picture, free will turns out to be a relational property, partly constituted by both agent and circumstance, and not the kind of thing that is settled entirely by the presence of, say, a mechanism of practical reasoning or a general cross-situationally stable capacity to recognize and respond to moral considerations.” (Vargas 2013, 206).

5 NB 1 certainly do not mean to suggest that if one accepts atomism or holism in one domain, one would be committed to accept atomism in some other domain. One can certainly have a holist perspective on self-knowledge without being a holist (or, in Dancy's vocabulary, a so-called 'particularist') about (moral) reasons for action.
about; a moderate atomist constructivist, on the other hand, thinks that following M 'under normal circumstances' constitutes one's attitudes one thereby knows about. But not all self-knowledge theorists are constructivists. Alternatively, one might be a 'reliabilist' about self-knowledge and think that the relevant procedure is a highly reliable way of finding out or detecting certain facts of the matter, and that one is entitled to assume that if one follows M, then one has 'latched on to' the relevant (mental) facts. A radical atomist reliabilist thinks M is a reliable method in all circumstances; a moderate atomist reliabilist thinks following M only has the desired result in the 'right' circumstances.

Despite these differences, what those with an atomist view of self-knowledge in general share is the (implicit) idea that what makes self-ascription count as self-knowledge does not require that the subject considers what went on before her self-ascription or how she will be inclined to act in the future (the subject's selfascriptive 'biography') nor what's going on, as it were, around or outside her self-ascription—the selfascriptive 'context', broadly conceived. Instead, what makes someone's self-ascription count as self-knowledge is a question the answer to which can be given by focusing on what went on during, or immediately preceding, the moment of self-ascription. This is so, in particular, for expressivist and rationalist views, since both 'speaking one's mind' and 'making up one's mind' are activities one does in the present, here and now. The result of these activities likewise delivers (knowledge of) what one thinks here and now. The question, therefore, is whether and why these methods, which provide the subject with knowledge of her occurrent states, would ipso facto tell her what her attitudes are: what she really believes, wants, hopes for, expects, prefers, fears, and so on.

To use a metaphor, we might say that whereas an atomist 'zooms in' and focuses on the methods and the moment of self-ascription, someone with a holist perspective on self-knowledge instead 'zooms out' and considers the place that a specific self-ascription occupies within the subject's larger biography and self-conception, as well as the concrete context in which her self-ascription was made, including the overall psychological state or 'mood' of the self-ascriber. She also takes into account the moral–political circumstances in which the subject finds herself, such as whether or not she is subject to the will of another. 6 Someone with

---

6 Another way of looking at it is to say that it's the atomist who zooms out and looks at different people's self-ascriptions from a distance and concludes that some epistemic feature is stable across these cases, whereas a holist zooms in and considers what is going on and what the good-making epistemic features are in each case.
a holist view of self-knowledge thinks that the question of whether some subject S has knowledge of her attitude A by following some procedure or method M can only be answered by addressing 'questions of circumstance'. Questions of circumstance include questions like these: Who is S? Where is S? What will S do? What has S done? How does S feel? What does S know? Who is S talking to? How is S related to A? How is S treated by others? How does S think about or treat herself?

On a holist view, it is not only the brainwashed, drugged and hypnotized among us who might potentially lack knowledge of our attitudes, in spite of having spoken or made up our minds but also people who are simply tired, hungry or insecure, who lost their temper or got out of bed on the wrong side, or people who are distracted, confused, heartbroken or in love. In other words, people like us. On a holist view, then, the question of what is required in order for someone's self-ascription to express self-knowledge can only be answered by tailoring this question to the specific individual who ascribes an attitude to herself, the life that she leads and the context in which she finds herself.

2. Qualifications, assumptions and limitations

Before proceeding, two important qualifications are in order. First, I do not mean to deny in what follows that tipsy colleagues, angry spouses or people under oppressive circumstances don’t have any self-knowledge, even though the process of my clarifying this qualification will require us to have some patience. This will be needed because in Chapter 7 I will suggest that we should not (just) distinguish between different procedures or routes to self-knowledge, but that it will be helpful to talk about different types of self-knowledge and allow for a sort of pluralism that at present is absent in the self-knowledge debate.

I also want to highlight an important assumption of this thesis, which is that I will be exploring theories of self-knowledge, more specifically expressivism and rationalism as theories of attitudinal self-knowledge, that is, as theories of how a subject acquires knowledge of intentional mental states such as beliefs, desires, hopes, intentions, and perhaps (some) preferences and emotions. The contrast is the knowledge we have of 'occurrent states', such as one's bodily sensations (the experience of pain or nausea), one's conscious thoughts or words “running willy-nilly through her head”, as Harry Frankfurt (1988) would put it. Most theorists

---

7 As we’ll see, ‘judgements’ form a special case, at least on the rationalist view.
of self-knowledge are explicit about the fact that their ambitions are to explain attitudinal self-knowledge, not (merely) the epistemic relation we have with regard to our own conscious thoughts, even though many of these theorists often go on to equivocate between the two. An underlying assumption of this thesis is that addressing the question of how someone knows her own attitudes is not necessarily the same as addressing the question of how someone comes to know her occurrent states, or indeed is not necessarily the same as addressing the question of how someone knows her beliefs. It may well be that equivocating between propositional attitudes and conscious thoughts is legitimate and that an account of how a subject comes to know the latter ipso facto gives us an account of how she knows her own desires, hopes and intentions, and (some of) her preferences and emotions. My plan, though, is not to assume that this is so from the outset. The hope is that the thesis as a whole will go some way towards explaining why these two ‘domains’ of the mental are importantly distinct and require different epistemic explanations.

I should also make explicit the limitations of this thesis. As said, I will for the most part concentrate on expressivism and rationalism. This means I will not be discussing a number of other interesting theories of and philosophers interested in self-knowledge in the chapters to come. I will not, for instance, discuss any historic or contemporary introspectionist accounts (e.g. Armstrong 1968; Nichols and Stich 2003), nor will I have much to say about constitutivist theories which have been fairly dominant in the self-knowledge debate. One reason for this is that there’s only so much I can do in the chapters to come, and something has got to give. A better reason, perhaps, for leaving these theories aside is that they are relatively minimalist theories of self-knowledge; self-knowledge only requires that the subject is conceptually competent (constitutivism) and/or that she is equipped with the requisite faculty or mechanism (introspectionism). It is not unlikely therefore that these theories are, in current terminology, versions of radical atomism – the latter of which I will not, principally, be concerned with. Furthermore, traditional constitutivist views are centrally concerned with so-called ‘cogito states’, which concern thoughts of the form ‘I am thinking that I am thinking that P’. In other words, I will for the most part ignore

---

8 Those who mention being interested in propositional attitudes generally, not beliefs specifically or indeed occurrent thoughts, and who go on to equivocate between what attitudes and thoughts include are, for example, Donald Davidson (1987, 2001), Tyler Burge (2013), Annelisa Coliva (2012), Brie Gertler (2012) and, to some extent, Peter Carruthers (2011). Exceptions include, for example, Quassim Cassam (2011, 2014), Eric Schwitzgebel (e.g. 2010), Johannes Roessler (2013, 2015b) and Christina Borgoni (2015).
constitutivism because it is concerned mostly with occurrent thoughts rather than propositional attitudes (by which I mean attitudes other than so-called conscious or present beliefs). Whether or not atomist/holist approaches are (im)plausible when it comes to these occurrent states is not an issue I will be addressing. Another assumption I should mention is that I take it to be natural to think that being in a certain mental state is one thing, but knowledge thereof is another. I can want, feel or perhaps even believe something without knowing that I do; and I can think that I want, feel or believe something without having the relevant attitudes. I will be interested, therefore, in (versions of) theories of self-knowledge that do not claim that self-knowledge somehow ‘comes for free’ or is ‘self-intimating’, and which at least allow for the possibility that there could be circumstances that are epistemically undermining with respect to attitudinal self-knowledge. However, a constitutivist or, according to my definition, a ‘radical’ version of rationalism will be discussed in Chapter 6.

Finally, I should be clear that my claim in the chapters to come is not that all existing theories of self-knowledge are obviously and/or necessarily committed to (moderate) atomism. Rather, my aim is to examine the extent to which certain contemporary views implicitly subscribe to a version of atomism. Or, in cases where it’s not clear whether or not some view is ‘atomist’ or ‘holist’, the strategy will be to read them along atomist lines for the sake of the argument, and to examine what the implications and problems are and whether a holist rendering of the account is possible and is capable of avoiding these problems.

Before moving to the summaries of the chapters to come, let me end this introduction with a passage written by Wilfrid Sellars, who famously began his essay ‘Philosophy and the Scientific Image of Man’ as follows:

The aim of philosophy, abstractly formulated, is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term. Under ‘things in the broadest possible sense’ I include such radically different items as not only ‘cabbages and kings’, but numbers and duties, possibilities and finger snaps, aesthetic experience and death. (Sellars 1963)
The fundamental conclusion of my thesis is perhaps best described as a meta-thesis about the self-knowledge debate as a whole, urging that before we dive into the details of this or that procedure, we first of all need to question the preconditions of self-knowledge: which questions are assumed to have been answered without receiving explicit treatment? We must recognize that metaphysical questions such as what we take beliefs, desires and intentions to be, and epistemological questions about what we take 'knowledge' in 'self-knowledge' to be, are questions that dovetail with the question of how someone acquires self-knowledge.

Echoing Sellars, we might say that the more specific aim of the philosophy of self-knowledge, abstractly formulated, is, or at least should be, to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term. Under ‘things in the broadest possible sense’ I include such radically different things as not only selves and thoughts but friendships and feelings, cultural practices and character traits, morality and propositional attitudes. I have not even begun to chart, let alone give an account of, all of these and other (inter)connections. My hope is that the present thesis will at least succeed in showing that there are such interconnections and that at least some of them are crucial to a proper understanding of the topic and question of self-knowledge.

3. Plan for the thesis

The thesis is divided into two parts. The first part deals with expressivism and the second deals with (different versions of) rationalism. I will spend significantly more time on rationalism than on expressivism, because rationalism can and has been understood in different ways, whereas the difference between various accounts of expressivism appears to be less fundamental. I think it is therefore worthwhile to consider these different versions of rationalism independently.

In terms of the arguments provided against expressivism and rationalism generally, it will become clear, after a while, that a certain pattern begins to emerge. The structure of this thesis is not linear. What I will try to show is that despite the differences between expressivism and (different versions of) rationalism, these views are confronted with similar problems and face similar challenges. This pattern, in its turn, is an indication of the fact that moderate atomism in general is questionable.
In Chapter 2, I describe Dorit Bar-On’s (2004) expressivist view. Expressivism is usually read as merely offering a non-epistemic theory of first-person authority. The main goal of this chapter is to describe expressivism taken as a theory of self-knowledge; more specifically, it is to explain how we might understand that self-ascriptions, or what expressivists refer to as ‘avowals’, can express self-knowledge. In so doing, I explain how we might think of avowals as expressing (second-order) beliefs about one’s mental condition, what would make such beliefs justified and how an expressivist may approach the question regarding the truth conditions of self-ascriptions.

In Chapter 3, after having explained how expressivism can be understood as a theory of self-knowledge, I turn to the question of how it must be evaluated in this regard. I aim to show that on what seems to be a natural reading of the account, expressivism betrays a form of moderate atomism. I argue that the atomist elements in expressivism, and what stands in the way of a holist approach, can be traced back to the expressivist’s reluctance to see avowals as assertions. I suggest that a holist construal requires understanding self-ascriptions as assertions that serve a communicative point but that understanding self-ascriptions in this way is incompatible with one of expressivism’s central claims.

In the next block of chapters, 4–7, I turn to rationalism. In Chapter 4, I first of all outline the rationalist view as set out by Richard Moran (2001), and then discuss the central thesis that we can know our own attitudes by following the so-called ‘transparency procedure’. I discuss a number of common objections to rationalism and consider ways of dealing with them. I end by discussing the relation between judgements on the one hand and attitudes on the other, and point out, following a number of recent authors, that for the transparency procedure to deliver self-knowledge, it must be construed along inferentialist lines.

In Chapter 5, I distinguish between a ‘radical’ and a ‘moderate’ atomist version of rationalism. I take issue with the latter view, according to which the transparency procedure only has the right ‘output’ (self-knowledge) provided that the subject made up her mind and/or followed the transparency procedure in the right or normal circumstances of deliberation. I argue that the moderate rationalist is ultimately unable to provide a satisfying account of what counts as the right circumstances, at least without ruling out self-deception from the outset. This, then, leaves the rationalist the choice of construing rationalism along holist lines or ‘going radical’ and ruling out the appeal to normal circumstances from the outset.
In Chapter 6, I consider an alternative, so-called 'Activist' take on rationalism, as developed by Matthew Boyle (2009b, 2011a, 2011b). I suggest that Boyle’s view appears to be committed to the radical claim that ‘making up our mind’ is a way of acquiring self-knowledge whatever the circumstances or reasons involved in one’s deliberation and so would qualify as a version of radical atomism. My main aim is to show that the question of whether Boyle’s epistemic account of how one knows one’s attitudes is plausible – and thus whether radical atomism is problematic – depends on how one answers the metaphysical question of what attitudes are. To this end, I contrast Boyle’s metaphysical view with an alternative ‘Dispositionalist’ view, taking my cue from Eric Schwitzgebel (e.g. 2001, 2010). My conclusion is that radical atomism is implausible or absurd only given one’s (implicit) answer to the metaphysical question, which in its turn depends on more fundamental issues such as one’s views regarding the role (and relevance) of ‘intuitions’ about certain cases, and one’s meta-philosophical views.

In Chapter 7, I present yet another reading of rationalism by taking the rationalist account as a normative account of why the capacity to make up one's mind matters to autonomy. The connection between autonomy and self-knowledge has received little attention. I suggest that a thoroughly normative reading of the rationalist account may have a lot going for it, but that it has not been systematically compared with other theories of autonomy, such as ‘proceduralist’ theories on the one hand and ‘relational’ theories on the other. This, then, is what I set out to do. I then abstract away from the details of rationalism and conclude that we need to distinguish not just between different objects of self-knowledge or different routes to them (as Moran proposes, ‘alienated’ versus ‘non-alienated’) but also between different types of self-knowledge, such as introspective, non-alienated and autonomous self-knowledge. I finish the chapter by taking a step back to reflect on the deeper metaphysical and epistemic sources of atomism.

In the concluding chapter, I take stock and summarize some of the key points in the thesis, reflect on the contribution of this thesis to our understanding of self-knowledge and spell out possible directions for future research.
PART I

Expressivism
1. Introduction

Expressivism is often understood as, indeed typically presented as, a theory of 'first-person authority' rather than self-knowledge (Gertler 2011a, 2011b; Bar-On and Long 2001; Bar-On 2004). A theory of self-knowledge tells us what makes people's self-ascriptions epistemically privileged, whereas a theory of first-person authority deals with the question of why we generally presume this. The distinction is subtle but important. Being epistemically privileged means that our second-order beliefs about our first-order states are more likely to result in knowledge, compared, for instance, with the way in which we gain knowledge of others' mental states or the external world. First-person authority, on the other hand, means, in very general terms, that the speaker herself is the 'authority' on what state she thinks or says she is in. More specifically, first-person authority refers to the special authority subjects have with respect to expressing or reporting their own mental states. This is shown in the fact that when someone avows her mental states, we are usually not inclined to override, challenge or correct her self-ascription, or to ask her whether she is sure that she is in the self-ascribed state.

A so-called 'epistemic' reading of expressivism, i.e. a reading that takes expressivism to be about self-knowledge, has not received much attention in the literature. The main goal of this chapter is to discuss expressivism taken as a theory of self-knowledge as an account of knowing our minds by speaking our minds. This will pave the way for the subsequent chapter, where I take issue with the expressivist account of self-knowledge in so far as it is an atomist theory of self-knowledge.

The structure of this chapter is as follows. First of all I describe the general expressivist view (Section 2), and Dorit Bar-On's expressivist view more specifically,
given that her theory is the most comprehensive expressivist account to date (Section 3). I then point out, in Section 4, that even though expressivism is typically taken to be a non-epistemic account of the role and peculiarities of avowals, it makes prima facie sense to understand expressivism as an epistemic account of how we acquire self-knowledge, or indeed as offering a theory of something (roughly) like justified true beliefs about our first-order states. I explain how expressivism can be understood epistemically in Sections 5-7. Assessing expressivism as a theory of self-knowledge will involve addressing the question of how (well) it handles failures of self-knowledge, or what Bar-On calls ‘expressive failures’. I end by discussing some prima facie worries about the expressivist account, thereby paving the way for the next chapter, where I take up the question of the (ir)relevance of the circumstances of self-expression.

2. Expressivism: the basics

The ‘expressivist’ theory of self-knowledge includes many different views (Ryle 1949; Fleming 1955; C. Taylor 1985; Finkelstein 2008; Bar-On 2004; Bar-On and Long 2001; J. Roessler 2013, 2015a, 2015b; Jacobsen 1996). The central expressivist thesis, as I understand it, is not primarily metaphysical, nor epistemic, but rather has its main roots in the philosophy of language and action. The central thesis is one that concerns the question of the function of self-ascriptions. The idea is that we should recognize that self-ascriptions, or what expressivists prefer to call ‘avowals’, are linguistic acts that have their proper place in a linguistic, social community – an idea which, in its turn, of course, will have metaphysical and epistemic implications.

Expressivism can best be understood as a critique of and an alternative to detectivism. Detectivism can be described as a view which assumes that our self-knowledge is “grounded on our being aware, or somehow presented with, mental items that we must identify as being (or representing) states of this or that mental kind on the basis of their presented properties” (Shoemaker 1993, 78).² Dorit Bar-On similarly describes what she refers to as the “recognition” approach to self-knowledge, according to which self-ascriptions “must be grounded in our ability to tell through our recognition of distinctive features of a present mental state what content it has” (Bar-On 2004, 170). The self-ascriptions “I want a cup of tea” and

² This is not Shoemaker’s own view, but rather his description of Goldman’s (1993) view.
"I want to pursue a career in philosophy" are thus based or grounded on a judgement of some sort. Self-knowledge, then, is a matter of "recognitional success, which then invites an epistemic explanation". (Bar-On 2004, 198). On such a detectivist view, something like 'brute error' is in principle possible. Just as one can be mistaken in one's perception of a tree outside (it is really just a picture of a tree projected onto a wall), one could likewise 'mistake' one state for another through some kind of epistemic failure (misperception, misinterpretation, etc.). Brute errors are errors that do not involve any rational failure or a failure that is due to the subject's concepts or capacities; rather, a brute error "is one that is simply due to the world failing to cooperate" (Bar-On 2004, 9; see also Burge 1996, 153). David Finkelstein gives a very useful definition of a 'detectivist' epistemology as

a subject's ability to say what she is thinking or feeling is the result of her somehow finding out—whether by observation alone or in conjunction with memory and inference. A detectivist thinks that our ordinary consciousness of at least some significant range of mental states or events is explained by the fact that we are able to detect their presence. (Finkelstein 2007, 216, 2008)

One of the basic claims of expressivism is that detectivism is false. This is, for instance, evident in Ryle's work:

If a person says "I feel bored", or "I feel depressed", we do not ask him for his evidence, or request him to make sure. We may accuse him of shamming to us or to himself, but we do not accuse him of having been careless in his observations or rash in his inferences, since we do not think that his avowal was a report of observations or conclusions. He has not been a good or a bad detective; he has not been a detective at all. (Ryle 1949, 87)³

On a detectivist view, self-ascriptions have a reportive function. On an expressivist view, by contrast, self-ascriptions have an expressive function, that is to say, avowals are not reports of underlying states but directly 'speak from' or 'show' these states, provided that the subject is sincere.

³ Though Ryle is often described as an inferentialist or interpretationalist (and not without reason), his view has importantly expressivist elements, as this passage shows. For an expressivist reading of Ryle, see also Johannes Roessler (2015b).
Expressivists claim that our self-ascriptions or avowals are continuous with other
ways in which we express ourselves that aren’t reportive either. When someone
smiles, it’s implausible to think that her smiling somehow represents or is grounded
in her belief or judgement that she is pleased, and that her smiling behaviour, as it
were, results from that belief. Rather, we think that the smile expresses, i.e. directly
shows, her pleasure. According to expressivists, we should think about avowals in
a similar way. When someone offers the self-ascriptive avowal “I am so pleased to
see you” instead of, for example, smiling, this self-ascriptive expression does not
represent or ‘rest’ on the subject’s belief that she is pleased to see X; she is simply
speaking her mind.

Contemporary expressivists typically start by describing and distancing
themselves from the ‘simple’ or ‘classic’ expressivist account that is often associated
with some of Wittgenstein’s remarks. Consider, for instance, the following well­
known passages:

To say, “I have a pain” is no more a statement about a particular person than
moaning. (Wittgenstein 1953, 67)

The verbal expression of pain replaces crying and does not describe it.
(Wittgenstein 1953, §244).

It can’t be said of me at all (except perhaps as a joke) that I know I am in pain.
What is it supposed to mean—except perhaps that I am in pain? Other people
cannot be said to learn of my sensations only from my behaviour,—for I cannot
be said to learn of them. I have them. The truth is: it makes sense to say about
other people that they doubt whether I am in pain; but not to say it about myself.
(Wittgenstein 1953, §246)

I can know what someone else is thinking, not what I am thinking. It is correct
to say “I know what you are thinking”, and wrong to say “I know what I am
thinking.” (A whole cloud of philosophy condensed into a drop of grammar.)
(Wittgenstein 1953, §315)

On the Wittgensteinian expressivist account, my sincere self-ascriptions are not
grounded in my second-order beliefs about my first-order state (are not the result
of scanning or detecting). In fact, the claim is much more radical yet, namely our avowals do not report anything at all. We are misled if we think that our avowals are instances of describing and reporting purely because of the grammar of avowals, in which I seem to be ascribing some state to myself (e.g. “I am sad”, “I want a slice of cheesecake”, etc.).

The problem with the Wittgensteinian version of expressivism is that avowals are not truth-evaluable, and there is thus a deep discontinuity between my saying “I feel bored” and your saying “You feel bored”. You are making a truth-evaluable claim about the state I am in, whereas I am not. This is strange, because on the face of it, my utterance seems to be one in which I do in fact ascribe a certain condition to myself. Also, it seems that one’s self-ascriptive expressions, such as “I feel bored”, can be used in a variety of inferences, e.g. “I am bored, but you are not, so only one person is bored”. But if we follow Wittgenstein and think that saying “I am bored” is rather like ‘moaning’ and hence is not truth apt, then we cannot explain the fact that my utterance is similar to and semantically continuous with other ascriptions in this way.

Hence, contemporary expressivists move from a simple or traditional expressivist view to versions that are able to accommodate the fact that sincere, self-ascriptive expressions are truth-evaluable. According to Finkelstein, avowals have an “assertoric function” (Finkelstein 2008, 101), and according to Bar-On, any account of avowals must respect the claim that avowals are interchangeable salva veritate in context with certain unproblematic statements and can figure in certain logical inferences (Bar-On 2004, 10). The latter is what Bar-On refers to as the desideratum of ‘Semantic Continuity’.

Contemporary expressivism thus tries to pull off a delicate balancing act between, on the one hand, seeing linguistic self-ascriptions as very similar to other expressive acts (smiling, etc.) and, on the other hand, seeing avowals at the same time as very different from expressive acts like smiling, in order to secure semantic continuity between avowals and other linguistic acts, and the idea that avowals (though not smiles or moans) can be false.4

I turn to the way in which expressivists accommodate the phenomenon of ‘false avowals’ or ‘expressive failures’ in a moment. First, I describe the idea of self-expression in more detail. I do so by focusing on Bar-On’s neo-expressivist account

---

4 I address the relevance of securing semantic continuity in the subsequent chapter, where I argue that the expressivist is forced to choose between seeing avowals as expressions (similar to smiling) or seeing them as speech acts that can be true or false.
Chapter 2

(2004, 2015; Bar-On and Long 2001). I focus on her account specifically because her theory is the most comprehensive expressivist account to date and also because, apart from Bar-On, not much has been written on what constitutes ‘expressive failures’, which I will argue is an important aspect of any plausible expressivist theory of self-knowledge.

3. Bar-On’s neo-expressivism

Bar-On describes the specific type of act that avowing is by drawing on the work of Wilfrid Sellars (1969). The point Bar-On takes from Sellars is distinguishing between two senses of the notion of ‘expression’: expression in the action sense and expression in the semantic sense. When someone gives expression to a state of mind in the action sense (when she ‘a-expresses’ her state), she expresses a state of hers by intentionally doing something. Expression in the action sense is what avowals share with other natural expressions:

You could clap your hands enthusiastically, emit a “Yeal!”, say “This is great!” or avow “I’m so excited”. Pre-theoretically, there seems to be very little to distinguish the different acts in terms of the process they involve. (...) [They all] seem to issue directly from your excitement, without the intervention of an inward reflection on, or even recognition of, your condition. (...) These considerations may speak in favor of a prima facie similarity in the process involved in all these performances. (2004, 252–53)

A-expression is a relation that holds between an agent, her mental state and an expressive vehicle (gestures, facial expressions or verbal behaviour). This should be distinguished from expression in the semantic sense (s-expression), which is a relation that holds between “contentful tokens, such as sentences, and their semantic contents” (Bar-On 2015, 140). So, saying “I’d like some tea” is both an a-expressive and an s-expressive act. But pointing at ‘Earl Grey’ on the drinks menu

5 In her book, Bar-On also mentions that there’s a causal sense of expression: “an utterance or piece of behavior expresses an underlying state by being the culmination of a causal process beginning with that state” (2004, 216). I leave this third category aside for now, as does Bar-On (2015).

6 I return to the relevance of expression being an intentional action and its relation to how expressing oneself delivers self-knowledge below.
Knowing Your Mind by Speaking Your Mind

does not s-express anything, though it still expresses my desire for Earl Grey tea in the action sense. The fact that avowals are expressions in both senses is what makes them of special interest.

Avowals are, according to Bar-On, similar to other natural expressions like smiling. What is distinctive about naturally expressive behaviour, she writes "is not only that it shows the conditions or states it expresses, but that it is behavior that is sufficient to show the relevant conditions or states" (Bar-On 2004, 275). The contrast here is between 'showing' something and 'communicating' something, the latter involving telling someone or conveying information about yourself (such as what mental state you’re in). To express one’s mental states, on her view, is to “show the conditions they express, rather than merely communicate information” (Bar-On 2004, 273). According to Bar-On,

avowals wear the conditions they are supposed to express on their linguistic sleeve, as it were. An avowal such as “I wish we’d get some rain today” explicitly names a kind of condition (a hope) and articulates its content (that it rain today), as well as ascribing it to a certain individual; it reveals the kind of state the avower expresses (as well as its intentional content, when it has one) through what the sentence expresses in the semantic sense. (2015, 142 emphases removed)

In his article ‘Language as Thought and as Communication,’ Sellars makes the rather radical suggestion that the activity of thinking out loud or indeed speaking one’s mind is “a form of meaningful speech which doesn’t consist in talking to anyone at all, even oneself, and hence is not, in any ordinary sense, talking” (Sellars 1969, 518). Echoing Sellars, Bar-On writes that spontaneous, self-ascriptive expressions are not made “with the aim of informing their audience of what is going on in them” and are not acts subjects undertake “to perform with a specific audience-directed goal in mind, such as convincing, informing, pleasing, etc. Like many non-verbal expressive acts, they may not even have any communicative point” (2004, 242). So when someone says “I think that P” or “I want X”, she is not informing or telling anyone anything but simply expressing her view that P or her desire for X. As Bar-On emphasizes, “the point of avowing an intentional state is not to provide a descriptive report of it, but rather to share it, or air it, or give it voice, or just to ‘vent’ it”, and our self-ascriptive expressions “may not even have any communicative point” (2004, 243).
Importantly, Bar-On claims this only holds for what she refers to as ‘avowals proper’, which Bar-On and Long (2001) define as “sincere, spontaneously volunteered, unreflective utterances (voiced or silent)” (Bar-On and Long 2001, 326). Avowals proper, they say, are “often evinced as an immediate reaction to something” and “are pressed from the subject and are not the culmination of the subject’s reflective truth-targeting act” (2001, 326). In what seems to me to be an important footnote, Bar-On contrasts the notion of ‘avowals proper’, which she characterizes as “purely expressive”, with so-called ‘mixed avowals’, which “have descriptive or reportive elements” (2004, 304). Later in the book she seems to suggest that we should think of our self-ascriptions as being on a spectrum, from highly secure ‘avowals proper’ at one end to “alienated or theoretical “I”-ascriptions, which are arrived at on the basis of evidence, inference, etc.”, at the other (Bar-On 2004, 338).

Bar-On is explicit about the fact that self-ascriptions made “on the basis of therapy, consultation with others, self-interpretation, or cognitive test results” do not qualify as avowals and do not belong to the “ordinary, everyday way” (2004, 194) in which we ascribe mental states to ourselves. Crucially, it is only when we offer avowals proper or when we “simply avow ... (as opposed to making a conjecture about my own state of mind, for example)” (2004, 193) that her expressivist account (fully) applies.7

My ultimate goal is to assess expressivism as a theory of self-knowledge, and, more specifically, to see whether it is plausible taken as an atomist theory of self-knowledge. In other words, whether speaking your mind is sufficient to acquire knowledge of your mind, all else being equal. But nothing I’ve said so far actually concerns self-knowledge. Should we understand expressivism as a theory of self-knowledge at all? Obviously, radical/moderate atomist expressivism is going to fail as a theory that gives us the sufficient conditions for self-knowledge if expressivism is, at the very beginning, not a theory of self-knowledge at all. So the first step is to address the question of how we might understand expressivism as a theory of self-knowledge.

---

7 This will be of relevance later, when I discuss expressivism as an atomist theory of self-knowledge.
4. Expressivism as a theory of self-knowledge?

4.1 Self-knowledge and first-person authority

Curiously, expressivists themselves have been hesitant to construe their theories as theories of self-knowledge. I say 'curiously' because an expressivist account understood epistemically seems to make a lot of intuitive sense. If I sincerely say "I'm so pleased to see you!" how could I possibly fail to know about my being excited to see you? Or if the waiter comes over and asks me what I want to drink, and I haven't really given it much thought, but I find myself saying "I'd like a cappuccino, please", what more would be needed for me to know whether I have a desire for a cappuccino? As we'll see, things aren't so simple, but at least prima facie expressivism may have a lot going for it epistemically speaking.

One reason why expressivism hasn't received much 'epistemic' interest might be due to the fact that self-knowledge is generally conceived as a type of second-order knowledge, and expressivists emphasize the spontaneous and unreflective nature of our self-ascriptions. Perhaps this combination is not taken to be very plausible. Also, even though contemporary expressivists allow that avowals are truth-evaluable, they still deny that avowals are reports of underlying states. Still, at the very least it seems plausible that self-knowledge would be a natural side-effect, if you will, of expressing ourselves self-ascriptively, even if when we avow some mental state we are not being 'detectives' or intending to report on what's on our minds.

But, as mentioned earlier, expressivists see their primary goal as explaining first-person authority rather than self-knowledge. Central to Bar-On's book is the following question:

What accounts for the unparalleled security of avowals? Why is it that avowals, understood as true or false ascriptions of contingent states to an individual, are so rarely questioned or corrected, are generally so resistant to ordinary epistemic assessments, and are so strongly presumed to be true? (2004, 11)

Explaining first-person authority involves explaining the special role that avowals have; in particular, it involves explaining why certain epistemic challenges to them

8 I should note that Bar-On does not like the term because she believes that notion implies that what's special about avowals is "a matter of epistemic advantage" (Bar-On 2004, 123).
are inappropriate. The basic idea is that when someone avows “I’m so tired” or “I hope the meeting will be cancelled”, we do not usually respond by saying “Are you sure?”. As Ryle writes, the avowal ‘I want X’ is “not used to convey information, but to make a request or demand. It is no more meant as a contribution to general knowledge than ‘please’. To respond with ‘do you?’ or ‘how do you know?’ would be glaringly inappropriate” (Ryle 1949, 164). Certain epistemic challenges are inappropriate, expressivists claim, because such challenges assume that you were ‘being a detectivist’ about your own mind, i.e. that your avowal was the result of trying to come to know your own mind. Thus understood, to say “How do you know?” as a response to my saying “I’d like a cappuccino” would be out of place because I wasn’t ‘claiming any knowledge’ in the first place. It’s a bit like, say, if you are strolling through a museum and then stop to look at a painting, and someone comes up to you and asks “Well, what’s the verdict?” when you weren’t in the business of trying to come up with a well-formed judgement about it in the first place.

Though the expressivist account of first-person authority has its strengths, I am not sure just how strong the link is between our actual intersubjective practices of (not) challenging avowals and what avowals actually are, i.e. the claim that they are non-reportive speech acts. There are all kinds of reasons for not challenging people's avowals, and it's not obvious that our reasons are grounded in the fact that the self-ascriptor was not ‘in the business’ of offering a self-descriptive report. Also, there are cases in which the audience seems equally disinclined to challenge someone’s speech act when this isn’t an avowal (e.g. “I had a headache this morning” or “I was imagining myself on a beach in Spain a moment ago”). Plus, there seem to be quite a number of cases in which we do challenge peoples avowals, and do so precisely for the reason that what we take them to be doing is making a claim about ‘how it is with them’, i.e. claiming to know how they feel or what they want. I address these issues in the next chapter. For now, the point is that even though expressivism can perhaps make sense of the absences of epistemic challenges, it’s not obvious that expressivism can also make good sense of the presence of such challenges.

But what explains our practices of (not) challenging each other’s avowals is not my primary concern here. The expressivist epistemology seems to have become buried under its account of first-person authority. Some even think that

---

9 These are not avowals because you cannot express, i.e. directly show or speak, from past mental states, only from your present mental states.
expressivism has not offered, or cannot offer, a theory of what allows subjects to know, by avowing, what mental state they are in. In his review of Bar-On’s book, Alex Byrne complains thus:

Now all this might seem a little disappointing. The subtitle of Speaking My Mind is ‘Expression and Self-Knowledge’, but self-knowledge is not the chief topic. We aren’t getting an explanation of why self-knowledge is “privileged”, but at best an explanation of why we presume this. (Byrne 2011b)

Similarly, Brie Gertler writes:

Because these [expressivist] views do not seek to explain our epistemic relation to our own states, they are not accounts of self-knowledge, *per se*. Since this book is mainly concerned with the epistemic dimensions of self-knowledge, we will not discuss [Wright’s] default authority or expressivist views further. (Gertler 2011b, 54–55 Gertler’s emphases)

I think this dismissal of expressivism is too quick, not only because I think there are good prima facie reasons for taking an expressivist theory of self-knowledge seriously but also because the expressivists that Byrne and Gertler discuss and dismiss have sought to explain our epistemic relation to our own states. It may be true that providing an account of self-knowledge is not the expressivist’s first concern, but that doesn’t mean such an account could not, or indeed has not, been provided.

In her book, Bar-On distinguishes the earlier non-epistemic question regarding first-person authority from the following question:

Do avowals serve to articulate privileged self-knowledge? If so, what qualifies avowals as articles of knowledge at all, and what is the source of the privileged status of this knowledge? (2004, 11)\(^{10}\)

---

\(^{10}\) Bar-On also adds a third question: “(3) Avowals aside, what allows us to possess privileged self-knowledge? How is it that subjects like us are able to have privileged, non-evidential knowledge of their present states of mind, regardless of whether they avow being in the relevant states or not?” This question, she writes, assumes that we can explain self-knowledge in abstraction from avowals altogether. Bar-On’s goal in Speaking My Mind is to give a non-epistemic answer to (1); to show that there are non-deflationary views consistent with (2); and to explain that there is something wrong with (3).
Even though Bar-On’s primary interest certainly isn’t self-knowledge but rather first-person authority, she devotes two chapters in the book to the question of how avowals might serve to express self-knowledge. Johannes Roessler has also recently defended a ‘modestly epistemic’ expressivist account of self-knowledge (J. Roessler 2013, 2015a, 2015b). In what follows, I explain the shape of an expressivist account of self-knowledge.

5. Avowals as expressing second-order beliefs

In order to make the activity of speaking one’s mind into an epistemic activity, we need that activity to provide (something like) justified true beliefs about one’s first-order states. As Anthony Brueckner writes, “if an avowal ‘represents’ self-knowledge, then there must be a justified, true belief associated with the avowal” (Brueckner 2011, 181). The justified true-belief model (JTB model) of knowledge is certainly not uncontroversial. Here, though, I use the JTB model merely as a benchmark, as does Bar-On (2004, 363). Applying JTB analysis to self-knowledge, the basic idea is that speaking one’s mind must first of all provide one with some sort of belief or judgement or some way of being cognitively related to one’s first-order state (belief, desire, feeling); second, the belief should be true; and third, the belief should be justified, i.e. should not have come about by chance or luck. I’ll discuss these elements one at a time.

The idea that avowals represent (second-order) beliefs might seem hard to square with the expressivist approach. After all, expressivists claim that an avowal is not a belief in the sense that it results from an activity of detecting or trying to recognize what state one is in. Bar-On, however, distinguishes between two different senses in which one may be said to believe that p, namely believing in the ‘opining’ sense and believing in the ‘taking as true’ sense (2004, 363). The first involves being able to give corroborating evidence or reasons for one’s belief, and this is the sort of belief that is incompatible with expressivism. For if I say “I really want some coffee”, then I am not ‘of the opinion’ that I want some coffee – this is why it would be out of place for you to ask, for example, what makes me think I want some coffee. If avowals were to express beliefs in the opining sense, then avowals are reports of one’s first-order states. If that were the case, then the distinctively expressivist element has been given up. But, Bar-On continues, there is a more liberal, dispositional sense of belief,
in which a subject believes that \( p \), provided (roughly) that she would accept \( p \) upon considering it. This holding-true sense, as we may refer to it, is the one we apply when we say that people have beliefs concerning matters they have not yet even considered. For example, I may not presently have any active opinion, formed on some specific basis, regarding matters such as the color of rain in Spain, or the sum of some numbers, or that a building within my field of vision is taller than a tree standing next to it. Yet, if suitably prompted, I would affirm the relevant claims. (Bar-On 2004, 363)

'Hold true' requires only that I "would accept \( p \) if I were to consider it" (2004, 365). This requirement is arguably too liberal, though. It seems odd to say that when someone avows that she is hungry, she might believe that she's hungry only if suitably prompted or if she were to "consider it".

However, Bar-On rightly observes that the expressivist can make room for a stronger notion of belief by appealing to the fact that speaking one's mind is an intentional action. Given that "we normally credit people with knowledge of what it is they are doing, in the course of doing it" (2004, 386), we can likewise say that we can credit the self-ascriber with knowledge of what she is doing. This feature of 'knowledge in action' is central to understanding the expressivist epistemic potential (as well as to some varieties of rationalism – but that's something I will discuss in later chapters). Roessler makes a similar claim: "As Hampshire puts it, statements about one's own future or current intentional actions have, in this sense, a 'double aspect' (Hampshire 1965, 72). It is not clear why a self-ascription of belief should not also serve to express two attitudes" (J. Roessler 2015a, rs8).

It's this intentional feature of self-expression that should make plausible what Bar-On refers to as the “Dual Expression Thesis”, which is the idea that "when I avow, unlike when I engage in naturally expressive behavior, I express not only the avowed condition but also my judgment that I am in that condition" (2004, 307). Brueckner formulates it as follows:

If (a) a speaker sincerely utters sentence \( S \), (b) the utterance is backed by competent linguistic understanding and (c) the utterance of \( S \) semantically expresses \( P \), then the utterance expresses the speaker's belief that \( P \). (Brueckner 2011, 181)
As Brueckner points out, Bar-On is reluctant to accept this principle (Brueckner 2011, 181). But I’m assuming her acceptance of this principle here for the purposes of assessing expressivism as an epistemic account of self-knowledge. For despite Bar-On’s reluctance, we have good reasons to accept the above principle as applying to avowals, because

if the Dual Expression Thesis is denied and the [above principle] is restricted to non-avowals, then this would constitute something close to a violation of Semantic Continuity. Whereas my utterance of “I am standing” would express my belief that I am standing, my utterance of “I am in pain” would not express my belief that I am in pain. (Brueckner 2011, 181)

This seems right. If the neo-expressivist wants to avoid collapsing back into simple expressivism, which sees avowals and non-avowals as semantically discontinuous, and if it’s plausible that ordinary assertions express beliefs, then expressivism ought to accept the idea that avowals also express one’s belief that one is in the self-ascribed condition. When I sincerely and spontaneously ‘vent’ my mental states by saying “I really want some coffee”, I have shown not only my desire for coffee but also my belief or judgement that I have a desire for coffee, and I have shown this not just to you but also to myself, the latter explaining why people’s avowals should represent attitudinal self-knowledge. Indeed, Bar-On in the end says, “All there is to having a self-belief in the case at hand, it might be held, is the (intentional) issuing of a self-ascription” (Bar-On 2004, 366). The upshot of this is that expressivism can satisfy the first-belief requirement for self-knowledge. The next question is what would make these second-order beliefs true and justified.

6. Avowals as expressing justified second-order beliefs

There are different ways in which an expressivist might consider second-order beliefs to be justified. Bar-On suggests that her expressivist view is compatible with different approaches. One possibility is to adopt an ‘Expressivist-Reliabilist’ view (Bar-On 2004, 369–73) and to say that someone has self-knowledge if her true self-
Knowing Your Mind by Speaking Your Mind

Ascription is reliably connected with the mental state she has expressed. This is not the view Bar-On herself prefers, though. Bar-On's own proposal is to understand the sort of self-knowledge that an agent acquires when speaking their mind by recourse to the fact that avowing is an intentional action. Johannes Roessler similarly appeals to the relevance of 'practical knowledge' to explain why speaking your mind might provide you with self-knowledge. He suggests that we may understand the epistemic role of people's expressions as follows:

Given that we normally enjoy "knowledge in intention" of what we are intentionally doing (Anscombe 1957), someone spontaneously (and sincerely) asserting that p can have knowledge in intention of stating her view that p. Beliefs are not actions. Still, one may be aware of an action under descriptions that do not fall short of a self-ascription of belief, such as "I am stating my view that p". Your assertion "the name is Langshaw", then, can make it intelligible that you are aware of your belief that he was called Langshaw: you may know your belief in "spontaneously" expressing it. ... [W]e might say that in making a spontaneous assertion one knows or is aware of—and in that sense confronted with—one's belief. (J. Roessler 2015a, 161)

Bar-On likewise stresses that 'avowing' something is "not merely something that happens to a subject, like the appearance of a rash, or a sneeze" but "is something she intentionally does: an act of speaking her mind". The character of such (intentional) actions, she suggests, "may yield a special epistemic warrant for avowing subjects" (Bar-On 2004, 383). Bar-On suggests that what makes someone's belief that she is in a particular state justified is the very state she speaks from:

What is epistemically unique about avowals is that the very same thing—one's being in M—provides both a rational reason for the avowal understood as an (expressive) act and an epistemic reason for the avowal understood as representative of the subject's self-judgment. (Bar-On 2004, 391)

---

11 This is what Bar-On refers to as the 'Low Road' approach. She also discusses a 'High Road' approach, which I will leave aside here and focus instead on her preferred 'Middle Road' approach.
What gives the subject epistemic warrant ... is the very same thing that gives her reason for the act of issuing the self-ascription: namely, her being in M. (Bar-On 2004, 393, see also 2012)

When I say “I want some coffee”, then the avowal is, as Brueckner explains, “grounded in, epistemically warranted by, the avowed state M itself, which is an epistemic reason”, which means that “the truth-maker for [my second-order belief] turns out to be identical to the justifier” (Brueckner 2011, 183). Fleming puts the point rather well when he says that when a man says he feels bored, then he “would say, surely, that he has got a justification for saying that he feels bored, namely, his feeling bored” (Fleming 1955, 620).

A consequence of this way of understanding the way in which one's second-order beliefs that are expressed in one's avowals are justified is that when we are dealing with a false self-ascription, one's second-order belief has no justification. For only in the 'good' case is there a mental state and one's avowal speaks from that state. Indeed, it's possible to claim that the avowal can be seen to be 'caused by' the mental state. But clearly, we need a very different (causal) story about what's happening when one issues a false avowal. Bar-On's solution is to appeal to disjunctivism (Bar-On 2004, 391ff), and she claims there is a fundamental difference between the justification one has to believe something in the good case versus in the bad case (i.e. in the case where there is a mental state and there isn't one respectively).

One of the most important questions, though, has not been considered so far, namely what actually makes someone's self-ascription true?

7. Avowals as expressing true second-order beliefs

What are the truth-makers of someone's avowal or self-ascription? On the face of it, the answer seems obvious: what makes someone's self-ascription true is just whether or not she is in the self-ascribed state. Just as we might say that what

12 A number of critics have taken issue with this move. Brueckner writes that the idea that there is no longer any "common epistemic factor between a case of a true avowal and a case of a false avowal" is counter-intuitive, and Aidan McGlynn notes that Bar-On's account of self-knowledge "rests on a very radical and implausible version of disjunctivism, and that her account of the security of avowals seems to push the question back in an unsatisfactory way" (McGlynn 2011).
makes someone’s self-ascription of having a headache true is her actually having a headache, we might similarly say that the question of what would make someone’s second-order belief about her desire/hope/belief true is just whether she actually does want/hope/believe that p. But this does not tell us very much, yet, for everything depends on what it means to ‘have’ a desire/hope/desire, and so on. On the face of it, the expressivist can just ‘plug in’ her preferred metaphysical account of what mental states are. In this section I therefore want to review some of the different metaphysical options that an expressivist (or any other theorist of self-knowledge) can appeal to.

We can broadly distinguish between ‘realist’ and ‘constructivist’ metaphysical accounts. Julia Tanney (2002) gives a helpful characterization of realism:

Much of modern and contemporary philosophy of mind in the “analytic” tradition has presupposed, since Descartes, what might be called a realist view about the mind and the mental. According to this view there are independently existing, determinate items (states, events, dispositions or relations) that are the truth-conferrers of our ascriptions of mental predicates. The view is also a cognitivist one insofar as it holds that when we correctly ascribe such a predicate to an individual the correctness consists in the discovery of a determinate fact of the matter about the state the individual is in—a state which is somehow cognized by the ascriber. (Tanney 2002, 37, see also 1996)

On a realist picture of the mental, the question of what your attitudes are (your first-order state) is independent of what you think your attitudes are (your second-order belief). The second-order state does not change the nature or ontological status of your first-order state. In other words, there are inner ‘facts of the matter’ that we can (fail to) know.

By contrast, on a constructivist view of the mental, the very nature of a person’s mental states is taken to be dependent on her relation to them, e.g. on her interpretation, acceptance or endorsement of them. Constructivists deny that a person’s first-level mental states and her judgements about them are ontologically distinct states, whereas a realist does not (see E. Fricker 1998, 173). On a constructivist view, then, a person’s first-order state and her second-order beliefs/judgements regarding that state are dependent on each other. Having a desire/belief or other attitude is partly constituted by a person thinking that she has a certain desire/belief or other attitude.
So-called ‘introspectionist’ views, according to which self-knowledge is modelled on perception, usually accept a realist view of the mental (e.g. Armstrong 1968; Lycan 1996; Goldman 2006). The core thesis of the inner sense view, as Gertler describes it, is that “introspection is a causal process and introspective beliefs qualify as knowledge because they are appropriately linked to the mental states they concern” (Gertler 2011b, 131). The so-called ‘rationalist’ account of self-knowledge, which I will discuss in later chapters, is, by contrast, a constructivist view. The important point for present purposes, though, is that expressivism can be construed either along ‘realist’ or ‘constructivist’ lines.

On a constructivist expressivist view, speaking your mind can be in part constitutive of what mental state you’re in. For instance, according to Charles Taylor’s (1985) ‘hybrid’ expressivist view (because he combines elements from expressivism, interpretivism and rationalism), our self-expressions are self-interpretations that may, or may not, constitute our emotions and attitudes. He argues that as ‘self-interpreting animals’ we are bound up in a process of articulating ourselves and that such self-interpretations shape us by informing and changing our attitudes.

This is not the sort of view Bar-On favours, however. The constructivist view, according to which “facts in the mental realm are determined in part by what self-ascribers take them to be”, Bar-On thinks, “seems to require settling for a certain irrealism about mental states, which many would find problematic” (Bar-On 2004, 412). In explaining what would make a self-ascriptive expression false (or true), Bar-On appeals to what she calls “robust commonsense realism” in relation to the mental (Bar-On 2004, 410ff). Bar-On does not tell us what she means by realism; indeed, she sees it as a strength of her view that she has “not committed myself to any particular view of what being in a mental state of these kinds consists in” (Bar-On 2004, 410). But, in any case, Bar-On criticizes the broader constructivist view by claiming that it is committed to what she takes to be a sort of irrealism about the mental, and claims it is “no part” of her expressivist view to maintain such dependence between first- and second-order states (2004, 254).

Why be anti-constructivist? The reason has to do with the fact that expressivism is often associated with some sort of crude behaviourism. There is a temptation, Bar-On says, to “associate all expressivist views of mental self-ascriptions with an implication that there simply are no independently existing mental states of subjects to ground the cognitive success or failure of mental self-ascriptions”, but this, she
writes, is a mistake. There is no reason “to saddle the [expressivist] view with the ontological denial, or with the intention of uprooting all cognitivist understanding of mentalistic discourse” (2004, 353–54).

If one adopts a so-called ‘Realist-Expressivist’ view, as Bar-On does, then this has consequences for how one thinks about failures of self-knowledge and, more specifically, of expressive failures, that is, cases in which one speaks one’s mind without acquiring self-knowledge. This will be central to the next chapter, so it is important to understand Bar-On’s way of dealing with expressive failures.

Bar-On claims that one might be in mental state M without showing the relevant expressive behaviour, and one might engage in certain expressive behaviour without actually being in M. She thus denies that her account does not commit her expressivism to the “crude behaviourist claim” that “naturally expressive behaviour is constitutive of mentality” (2004, 278). Bar-On goes on to discuss two types of ‘expressive failures’. In the first case, one expresses one’s own state, but one’s expressive behaviour is simply misperceived by others. An example might be a person who intends to show her mental state of joy, but whose expressive behaviour instead shows a strained grimace, e.g. because of a scar (2004, 281). In this case, the subject is in the relevant state and tries to show that state by expressing herself, but nonetheless fails to do so. So this is a case of expressive failure: someone engages in expressive behaviour, but her expressive behaviour does not show the mental state she’s in. In this case, the ‘mistake’ is not due to the subject, but rather the hearer.

Bar-On also discusses another, more interesting, type of expressive failure. In a second sort of case, the expressive behaviour is not misperceived by others, but expresses a state that the speaker is not actually in. Examples include saying “Ouch!” when one is not, in fact, in pain, e.g. when sitting in the dentist’s chair, anticipating the sting of the needle. In such cases, one cannot be expressing one’s own state, because one isn’t in the mental state, and hence it’s impossible that one speaks from that state.

Bar-On proposes that expressive failures of the second sort are instances in which one has still successfully expressed a mental state, but not one’s own mental state (2004, 320ff, see also 2012, 207, 2015, 145–46). This happens in cases of self-deception. Bar-On gives the example of someone who says “I find this painting very interesting”, but she does not really think this. The self-ascription was not insincere, but was brought about by “some kind of social or personal pressure to think well of the painting” (Bar-On 2004, 330). (This example does not strike me as particularly plausible, but
I’ll ignore this for now.) In cases of self-deception, one expresses a mental state (of thinking the painting interesting), without genuinely expressing one’s own state. Bar-On construes other examples in a similar way. For instance, she considers children who “display certain forms of expressive behaviour just for dramatic effect” or a talented actor who “produces a rolling laughter” (2004, 279). Bar-On writes:

In general, it seems possible to make expressive use of a tool that is semantically associated with a mental condition without being in the relevant condition, and thus without expressing one’s condition, even when one does not intend to deceive or engage in dissimulation or act playing. For this reason, I reject expressivist views that maintain that a subject’s avowing a mental condition conceptually guarantees her being in the condition (so the avowal must be true), as long as the subject is linguistically or conceptually competent and sincere. Such views fail to take account of the distinction between expressing M and expressing one’s M.

(2004, 325, see also 2012)

Bar-On’s suggestion is thus that in order to figure out whether someone’s speaking her mind is a way for her to achieve self-knowledge or whether it instead counts as an expressive failure, we must invoke a distinction between expressing M and expressing one’s M.

8. Some prima facie worries

Bar-On’s realist version of expressivism invites a number of worries and questions that I want to briefly turn to in this final section.

First, the distinction between expressing M and expressing one’s M appears to me to be rather an ad hoc move designed for the (sole) purpose of dealing with expressive failures. Also, it isn’t clear what it means, exactly, to express a mental state without expressing one’s own mental state. It now turns out that to express is still a ‘success verb’, but it does not necessarily successfully express the right thing.

This seems a counter-intuitive conception of ‘expression’, whereas arguably one of the strengths of the expressivist view is meant to be that it offers an intuitive account of what self-ascriptions are. What’s more important for present purposes, though, is that the entire weight of Bar-On’s account of when one has self-knowledge
via speaking one's mind and when one does not now falls on (her appeal to) mental realism. Whether one has expressed one's mental state or has merely expressed 'a' mental state – which makes all the difference for whether one does or does not have self-knowledge – in the end simply depends on whether or not one's self-ascription 'fits the facts'. In other words, the question of what is required for an (expressed) self-ascription to count as self-knowledge is simply outsourced to mental realism.

A first worry here is the metaphysical worry of whether we can even talk sensibly about mental ‘facts’ at all. Bar-On’s theory is dependent on a metaphysics of the mind that isn’t uncontroversial, as Tanney and others have shown. Another worry is a general epistemic worry, which various types of realism in general seem to face (including for example moral realism, see e.g. Sayre-McCord 2017) namely how do you know that your expressive (self-)reports ‘fit the facts’? What makes expressing oneself particularly reliable with regard to the facts?

Notice that on Bar-On’s realist construal of expressivism, it's no longer the case that the act of speaking your mind enables you to know your own mind, because for all you know, you were merely expressing a mental state – not your own. So when you speak your mind, there’s always the further question of whether you have really spoken your own mind. This seems to invite global scepticism about our avowals: how do we ever know whether or not our ascriptive self-expressions are true if their truth is something that's independent of our speaking our mind? Perhaps we could avoid such scepticism if we allowed that our relation to our own self-expressions co-determined what our attitudes are, i.e. by construing expressivism along constructivist lines as for example in Charles Taylor’s work, but this is not something Bar-On is willing to do. But if no room is made for such a constructivist element, then what would make it reasonable to think that by speaking our mind we acquire self-knowledge? This remains a mystery. Self-knowledge appears to be a (common) coincidence. Of course there's the possibility of saying that expressivism isn't and perhaps cannot be a theory of self-knowledge after all, but for reasons provided earlier, this does not seem a particularly desirable route to take.

Second, Bar-On’s rejection of constructivism appears to be based on an implausible conception of constructivism. Bar-On thinks that endorsing

---

13 Her reasoning seems to be that, apart from expressing a thought (so-called cogito-states), most self-ascriptions of attitudes are not self-verifying (Bar-On 2004, 220–21). This seems to be a particularly strong requirement, however. Attitudinal self-expressions, one might argue, can fail to be self-verifying and yet have an important constitutive role vis-à-vis the attitude expressed.
constructivism requires settling “for a certain irrealism about mental states” (Bar-On 2004, 412). This might be true of some constructivist views, but certainly not all. As Annalisa Coliva writes,

[Some] theorists have argued against constructivism on the grounds that it would entail the irreality of first-order mental states. The idea is that if such mental states do not pre-exist their self-ascription, they don’t have real, independent existence. Since this is implausible—the train of thought continues—constructivism is doomed from the start. (Coliva 2009)

But, as Coliva too notes, constructivists aren’t (necessarily) irrealists, anti-realists or relativists. This would only follow if one thought that for a mental state to be ‘real’ it must exist in the way that things like tables or cats exist or that having a belief would mean possessing “an internal representational token (perhaps a sentence in the language of thought) with the content $P$” (Schwitzgebel 2010, 536). In any case, if one thinks that constructivism is forced to accept some sort of irrealism, then the choice between constructivism and realism is not really a fair choice.

Bar-On’s main reason for resisting constructivism is that when it comes to “basic mental states”, such as “those we share with non-human animals and pre-cognitive children” (Bar-On 2004, 412), like being in pain or wanting an ice-cream, etc., it just does not seem plausible that such states are of the subject’s ‘own making’. That might be true, but then it seems we might need some other account of all of those mental states that we don’t share with animals and young infants, unless we restrict expressivism to these basic states, which makes the scope of the expressivist project rather limited. Also, it’s not uncontroversial to think of the basic mental states of linguistically competent adults as the same as the basic mental states of animals and young infants, which then invites a similar explanation. As McDowell claims, our

---

14 Daniel Dennett, among others, rejects the crude (Fodorian) realist view and defends a “milder sort of realism” according to which there is sometimes “no fact of the matter of exactly which beliefs and desires a person has” (Dennett 1987, 28), but denies that this should lead us to think that beliefs and desires, and so on, do not really exist. In Real Patterns, Dennett writes: “Philosophers generally regard ... ontological questions as admitting just two possible answers: either beliefs exist or they do not. There is no such state as quasi existence; there are no stable doctrines of semifrealism. Beliefs must either be vindicated along with the viruses or banished along with the banshees. A bracing conviction prevails, then, to the effect that when it comes to beliefs (and other mental items) one must be either a realist or an eliminative materialist” (Dennett 1991, 27).
subjectivity is not the same as those of cats and young infants "plus the conceptual garb" (McDowell 1998, 295).15

A third problem is that it's not clear how we should combine the realist claim that (1) what makes our avowals true is dependent on the existence of realist mental facts and that these mental facts are what causes an utterance or piece of behaviour (such as an avowal),16 with the expressivist claim that (2) mental states can be 'shown' and directly perceived. The worry is that the appeal to realism seems to pull us in the direction of thinking of expressive behaviour as signs or symptoms that reliably indicate a person's mental states, whereas thinking that mental states can be shown and directly perceived seems to push us away from that idea. The realist claim thus raises the question of how speaking one's mind would still be 'sufficient to show' one's mental states.17

A final worry is that the question of what it means to express 'one's own' mental state versus expressing 'a' mental state has just shifted to the question of what it means to (really) be in a mental state. In other words, unless we are provided with some more substantive account of realism, which Bar-On says she wants to remain neutral on, then we don't know what expressive failures are, and so we have no account of what it means in the end to know or fail to know one's mind by speaking one's mind.

Importantly, all of the above worries may or may not point to the possible limits of an expressivist account of self-knowledge, but this does not change the fact that expressivism can plausibly be taken as a theory of self-knowledge. Indeed, I will revisit some of these worries in the next chapter and will suggest that some of the above worries may lead us to consider a different (constructivist–interpretationalist) version of expressivism, such as Taylor's account briefly mentioned above.

However, of more immediate concern is the question of whether expressivism is plausible if construed along atomist lines. If expressivism is, as I have suggested,

15 Cf. "What is essential is to avoid the temptation to suppose that when, say, a cat, or a human infant, is in pain, what constitutes the relevant kind of episode in our inner lives is all that's in the cat's or infant's consciousness, barring only the ability to talk; or—to put the idea in connection with us—that our ability came to encompass our pains by way of our having our attention drawn to something that was already (sometimes) there in our consciousness—in the way that the cat's or an infant's pain is there in its consciousness—and being taught what to call it" (McDowell 1998, 294).

16 Bar-On invokes Sellars' notion of expression in the 'causal sense', defined as an utterance or piece of behaviour which "expresses an underlying state by being the culmination of a causal process beginning with that state" (Bar-On 2004, 216). For instance, someone's reaching behaviour may be caused by one's desire for a particular object; or one's shaking hands may be caused by one's pain or nervousness (ibid.).

17 In more recent work, Bar-On sets aside the causal sense of expression (Bar-On 2015, 149).
plausibly taken as a theory of self-knowledge, then there are at least two ways of understanding the view. One is that speaking one’s mind is sufficient for acquiring self-knowledge. This means that whenever one expressively avows believing that P or wanting Q, then one does believe that P/want Q, and by having expressed oneself in this way, one has come to know that one believes that P/want Q. Alternatively, on a moderate atomist version, speaking one’s mind normally provides self-knowledge, namely if one speaks one’s mind in the right or ‘normal’ circumstances of self-expression.

The relevant question then becomes what these circumstances would be and whether the expressivist can unproblematically appeal to certain (implicit) ceteris paribus clauses, such that the idea that one knows one’s mind by speaking one’s mind would still be informative. If, however, it turns out to be impossible to give an account of the circumstances in which speaking one’s mind actually is knowledge-conducive, because of what these circumstances can change from person to person, then this would mean an atomist conception of expressivism may have to be abandoned.

9. Conclusion

In this chapter I have introduced expressivism and have described more specifically how expressivism can be said to meet the relevant epistemic conditions to qualify as a theory of self-knowledge. The more or less standard contemporary expressivist view appeals to practical knowledge to explain both the sense in which one’s avowals represent second-order beliefs (as well as the first-order mental state it expresses) and what would make one’s belief justified. Whether or not this view is in the end plausible, it seems in any case unfair to dismiss expressivism outright by saying that it does not offer (or is not able to offer) a theory of self-knowledge.

Having shown how expressivism can meet the relevant epistemic conditions to qualify as a theory of self-knowledge, it’s now time to assess it as a theory of self-knowledge and, more specifically, as an atomist account of self-knowledge, according to which speaking your mind is always, or normally, sufficient for self-knowledge. But in what sort of circumstances might speaking your mind not provide self-knowledge? And are these circumstances minimal and straightforward enough to warrant putting them in an ‘all else being equal’ clause? This is the topic of the next chapter.
1. Introduction

On the expressivist view, the idea is that one directly shows or manifests one's mental state when one expresses or speaks one's mind. As discussed in the previous chapter, on Bar-On's account, it's possible to speak from 'a' mental state, but not from 'one's own' mental state. On the assumption that an expressivist theory of self-knowledge would not make any sense if it was a contingent matter whether, when speaking one's mind, one was expressing one's own mental states or rather 'a' mental state, it seems reasonable to suppose that when speaking one's mind, one expresses one's own mental state (hence I will simply, from here on, talk about mental states rather than 'one's own' mental states). If so, then a plausible understanding of expressivism is to see it as an atomist account of self-knowledge: if you speak your mind then you always (radical atomism) or normally (moderate atomism) come to know your own mind.

The key question is what is required to express or speak one's mind. There appear to be two conditions. First, one's self-ascription must qualify as what is referred to as an avowal. Call this the 'avowal condition'. If meeting the avowal condition is all that is required for one to acquire self-knowledge, then this amounts to a radical atomist version of self-knowledge, on my terminology. On this conception of self-knowledge, to answer the question of what makes someone's self-ascription deliver self-knowledge, all we need to do is zoom in on what the subject was doing when she ascribed a mental state to herself and answer the question of whether she was avowing her state or not. A moderate atomist, on the other hand, thinks that there are some circumstances (e.g. intoxication) in which one might speak one's mind without knowing one's mind. If that's the case, then we need to rule out
such circumstances. Call this the ‘normal circumstances’ condition. The moderate atomist version of expressivism relating to self-knowledge holds that only in normal circumstances does one come to know one’s own mind by speaking one’s mind.

The main goal of this chapter is to point out the limits of the moderate atomist version of expressivism, which I plan to do by exploring the plausibility of both above-mentioned conditions. First, I argue that there’s no unproblematic conception of the notion of ‘avowal’ that we can appeal to. I then argue that the strategy of ruling out exceptional circumstances in ceteris paribus clauses is not a legitimate move by pointing out that most ‘abnormal’ circumstances are actually quite ‘normal’ for human beings.

The structure of this chapter is as follows. First, I evaluate the plausibility of the avowal condition (Section 2), and then go on to address the ‘normal circumstances’ condition (Section 3). If both conditions are problematic, the result is that the question of what makes someone’s self-ascriptive act as self-knowledge cannot simply be answered by zooming in on the self-ascriptive act and figuring out whether, all else being equal, it was an ‘expressive’ self-ascriptive act. The interim conclusion is that perhaps we should consider whether expressivism can be construed along holist lines, and I make a proposal in this direction by suggesting that we should think of people’s self-expressions as self-interpretations, which amounts to understanding expressivism along interpretationist or inferentialist lines (Section 4). In the final two sections, I take a step back and offer a diagnosis by showing that the atomist version of expressivism can be traced back to the reluctance to see attitudinal self-ascriptions as genuine assertions (Section 5). I suggest that developing a holist account requires seeing self-ascriptions of our attitudes as genuinely semantically continuous with other ascriptions – a view that would be congenial to the interpretationalist approach suggested earlier. Given that reports regarding one’s height or hair colour are communicative speech acts, I suggest that if we want to respect semantic continuity and avoid the simple expressivist view, we should see our attitudinal self-ascriptions as communicative acts as well. I argue in the final section (Section 6) that this approach has significant theoretical advantages, in particular the fact that it paves the way for a holist account, but that it does not seem to be compatible with some of the expressivist’s core claims.
2. Ceteris Paribus I: assuming I am avowing

Central to expressivism is the notion of an ‘avowal’. So far, I’ve simply assumed that we have an intuitive grasp of what avowals are, and indeed this is what expressivists themselves typically do, too (Finkelstein 2014; Bar-On 2004). Bar-On, for instance, writes that she begins by describing “avowals in a very rough-and-ready way” and that her “starting point is a familiar phenomenon, which can be characterized ostensively as it were” (Bar-On 2004, 26). The question I want to address in this section is whether a definition of avowals can be provided that isn’t circular. In other words, whether it’s possible to give a characterization of what it means to avow something that does not build self-knowledge into the notion of what avowals are, for that would render the expressivist view non-informative (‘you come to know your mental states by avowing your mental states of mind if and only if avowing your mental states provides you with knowledge of your mental states’).

We can begin with the idea that avowals are ascriptions of states of mind, which means that not all self-ascriptions qualify as avowals. After all, I can sincerely self-ascribe this or that character trait (“I am timid” or “I am impulsive”), but such self-ascriptions are not epistemically privileged, because self-ascriptions of character traits such as being timid or impulsive are made on the basis of evidence (observation, theorizing, etc.). I cannot, by self-ascribing some character trait, show you this character trait of mine. To find out whether the self-ascription is true, both you and I need to look at the evidence over time. Self-ascribed character traits, habits, and so on are thus not epistemically special, and are ruled out as avowals. Avowals are pronouncements regarding one’s current mental states such as “hoping or wishing that p, whether I prefer x to y, whether I am angry at or afraid of z, and so on” (Bar-On 2004, 106).

However, this restriction won’t do, because self-ascriptions of mental states can also fail to qualify as avowals. Finkelstein, for instance, writes, echoing Wittgenstein, that the utterance “I hope he’ll come” can be understood as an ‘expression’ (avowal) but also as a report and that this can “vary from occasion to occasion” (Finkelstein 2014, 189). Bar-On makes a similar observation: “avowals contrast in their security with ascriptions that are grammatically and even semantically indistinguishable from them. ‘I am very happy’ can sometimes be an avowal and sometimes not” (Bar-On 2004, 26).

---

1 This also goes for other non-detectorist theories of self-knowledge, most notably, the rationalist account. So many of the worries voiced here can be read as being applicable to other theories that likewise appeal to the notion of avowal.
At this point, expressivists typically point out that whether a self-ascription is an avowal or not depends on the basis for and/or reasons that the subject had for issuing the self-ascription. More specifically, for a self-ascription to count as an avowal, there must be no such bases or reasons. An example of a self-ascription that isn't an avowal but is an ascription of a mental state, would be the self-ascription that I have certain implicit biases which I make, e.g. on the basis of having gone through a number of online implicit-association tests. Self-ascribed implicit biases have no special claim to truth because they’re made on the basis of evidence, and hence I could have made interpretative or inferential mistakes. Self-ascriptions like “I am bored”, “I am thirsty” and “I hope the meeting will be over soon” are different because they're not based on evidence, or so expressivists claim. So avowals are (1) ascriptions of one's mental states rather than, for example, character traits, where (2) these ascriptions are not based on evidence.

A further characteristic of avowals is that they are sincere self-ascriptions. Consider, for instance, Ryle’s remark:

Avowing “I feel depressed” is doing one of the things, namely one of the conversational things, that depression is the mood to do. It is not a piece of scientific premises-providing, but a piece of conversational moping. That is why, if we are suspicious, we do not ask “Fact or fiction,” “True or false,” “Reliable or unreliable?”, but “Sincere or shammed?” (Ryle 1949, 102)

But what is sincerity, exactly? As John Eriksson remarks, “the most common answer amongst philosophers is that a speech act is sincere if and only if the speaker is in the state of mind that the speech act functions to express” (Eriksson 2009, 213). But this is not a definition of sincerity that expressivists about self-knowledge can appeal to, on pain of circularity. If we want to explain expressing or speaking one’s mind in terms of avowals, and avowals in terms of sincerity (at least in part), then it seems we cannot explain sincerity in terms of expressing or speaking one's mind. But how else might we define sincerity?
In *Truth and Truthfulness*, Bernard Williams spends quite some time on these notions, especially in his chapter 'From Sincerity to Authenticity'. He writes:

Sincerity at the most basic level is simply openness, a lack of inhibition. Insincerity requires me to adjust the content of what I say. (Williams 2002, 75)

Regarding self-knowledge, Williams notes that "in the simplest case I am confronted with my belief as what I would spontaneously assert" (ibid.). Sincerity, he goes on, is to "speak frankly and spontaneously with others" (2002, 179) and is a form of "uninhibited expression or enactment, rather than reporting the findings of self-examination" (2002, 189).

In a somewhat similar vein, Gilbert Ryle, describes 'avowing' as a kind of "voluntary non-concealment" (Ryle 1949, 87). Ryle distinguishes between what he refers to as 'guarded' and 'unguarded' talk, and suggests that avowals belong to the latter category:

It is notorious ... that people are frequently insincere and talk in manners calculated to give false impressions. But the very fact that utterances can be guarded and studied implies that unguarded, unstudied utterance is possible. To be reticent is deliberately to refrain from being open, and to be hypocritical is deliberately to refrain from saying what comes to one's lips, while pretending to say frankly things one does not mean. In a certain sense of "natural", the natural thing to do is to speak one's mind, and the sophisticated thing to do is to refrain from doing this, or even to pretend to do this, when one is not really doing so. Furthermore, not only is unstudied talk natural or unsophisticated, it is also the normal way of talking. We have to take special pains to keep things back, only because letting them out is our normal response; and we discover the techniques

---

2 In this chapter, Williams gives an exposition of Rousseau's conception of authenticity that is implicit in his autobiographical *Confessions* – which Williams considers naively romantic – and contrasts it with the life and writings of Diderot. Williams, for instance, cites a passage from a letter Rousseau wrote to Malesherbes: "I shall show myself to you as I see myself, and as I am, for since I pass my life with myself I am bound to know myself, and I see from the manner in which others who think they know me interpret my actions and my conduct that they know nothing about them." Rousseau explains his decision to write the *Confessions* as follows: "I decided to make of it a work that would be unique because of its unparalleled truthfulness, so that once at least people would be able to see a man as he is within himself" (Williams 2002, 175). Williams' chapter from which this and subsequent passages originate is much richer than I can do justice to here.

3 For discussion, see also Roessler (2015a, 159-60) and Feldmann and Hazlett (2013).
of insincerity only from familiarity with the modes of unforced conversation that are to be simulated. (Ryle 1949, 162)

When talk is guarded, Ryle writes, we need to exercise “sleuth-like qualities”, that is, we have to “infer from what is said and done to what would have been said, if wariness had not been exercised, as well as to the motives of the wariness” (1949, 165). He compares a person’s guarded self-ascriptions to the pages of a sealed book, and contrasts this with spontaneous, unguarded talk, which he compares to the pages of an open book; he adds that “the fact that concealments have to be penetrated does not imply that non-concealments have to be penetrated” (1949, 165).

So if expressing your mental states means avowing your mental states and avowing one’s mental states means offering sincere mental self-ascriptions, then, following Williams and Ryle, we might understand this to involve speaking one’s mind in a spontaneous, uninhibited or unguarded way. If only avowals are the right sort of self-ascriptive speech acts to be included in the epistemic expressivist account, then only our spontaneous self-ascriptions can be known expressively.

The idea that avowals must be spontaneous or unguarded also plays a role in Bar-On’s account. She for instance, writes that “what I have been calling ‘avowals’ ... are to be contrasted with ‘alienated’ or ‘theoretical’ self-ascriptions” (2004, 172). She characterizes the latter by suggesting that these are ascriptions made, for example, “on the basis of therapy, consultation with others, self-interpretation, or cognitive test results” (2004, 194). The sort of self-ascriptions that are ruled out as avowals are, for instance, those that are offered in response to certain sorts of promptings. The reason for this, I take it, is because the subject might be in the business of answering a question rather than in the business of spontaneously speaking from her mental state. For expressivists, what makes avowals special (epistemically and otherwise) is that they are ‘groundless’ or ‘immediate’, i.e. that the avower has no “reasons or grounds for her avowal” other than the first-order mental state itself. If this were not the case, offering an avowal would not be similar to smiling – smiling directly ‘speaks from’ the underlying mental state and is not mediated by having a ‘reason’ to smile, or on the judgement that I am pleased to see you. Bar-On is explicit about this:

---

4 Moran (2000) similarly suggests that self-ascriptions made in the context of psychotherapy aren’t avowals (see e.g. Moran 2001, 86).
When avowing, as opposed to issuing a mental self-ascription on the basis of evidence, inference, analysis, or self-interpretation, a subject has no reason, or epistemic grounds, for affirming the various components of the self-ascription other than whatever reason or grounds she has for issuing the self-ascription as a whole. (Bar-On 2004, 226)

The question, though, is this: when we are actually dealing with unguarded, uninhibited, spontaneous self-ascriptions, and thus, avowals, is it even possible to distinguish real avowals – spontaneous, unguarded self-ascriptions – from seeming avowals, i.e. guarded self-ascriptions? How many of our self-ascriptions are really spontaneous and unprompted?

Interestingly, Ryle notes that “often we do not know whether [talk is guarded] or not, even in the avowals we make to ourselves” (1949, 164–65). Ryle thus realizes that there is an epistemic problem when it comes to being able to tell apart avowals and self-ascriptions that aren’t avowals. However, Ryle does not seem to recognize the implications that this has for theories of self-knowledge that appeal to the notion of avowals. It seems right to me that we often don’t know whether someone’s self-ascription is guarded or unguarded, including our own. But if that’s true, then this means that we often don’t know whether or not we are dealing with a sincere self-ascription, and so by implication we don’t know whether or not we are dealing with an avowal, which means we don’t really know which of our self-ascriptions are epistemically privileged and which of them aren’t. If that’s the case, then the (moderate atomist) slogan that you know your own mind by speaking your mind does not really tell us very much.

The emphasis on spontaneity or unpromptedness seems problematic because it rules out what appear to be perfectly normal instances of expressing one's mental states. Suppose, for instance, you’re at a workshop, and the coffee break has just begun. You’re standing next to someone and maybe you’re slightly uncomfortable, perhaps because you cannot get the conversation going, or because you’re tired of always talking about philosophy during coffee breaks, or because you’re dreading your own talk after the break. Then, at the end of the room, you see they’ve brought in the sandwiches. In order to avoid the uncomfortable situation, you say to the person standing next to you, “I’m hungry” (which is true) or “I’m going to grab a sandwich” (also true), and with a nod to the person next to you, you walk off to the buffet.
This seems to me a very ordinary scenario involving speaking your mind. However, though your self-ascription was sincere, it was clearly prompted, or in any case not particularly ‘spontaneous’. You offered your avowal partly in order to avoid talking to this person. It is not as if, in this example, you had “no reason, or epistemic grounds, for affirming the various components of the self-ascription other than whatever reason or grounds ... [you had] for issuing the self-ascription as a whole” (Bar-On 2004, 226). So it seems, to the extent that your self-ascription was, as Ryle would say, to some extent ‘guarded’, that your self-ascription does not belong to the class of speech acts that count as avowals. Or in other words, if your self-ascription was not a genuine avowal, then the expressive account of self-knowledge cannot apply to it. You don’t come to acquire self-knowledge by having avowed your mental state, because expressivists would not consider it to be a case of having avowed anything at all. You have either expressed or spoken your mind when you said “I’m going to grab a sandwich”, but you have not ‘avowed’ anything, which would be rather strange because it would require decoupling ‘speaking/expressing one’s mind’ and ‘avowing’ something, which I take it is something an expressivist would not be willing to do. Alternatively, you did not express or speak your mind and so did not avow anything. But if one cannot be said to speak one’s mind or express a desire when one says “I’m going to grab a sandwich”, then when does one speak one’s mind?

With the self-ascriptions that Bar-On rules out as avowals, she includes those made on the basis of “consultation with others” (2004, 194). The reason for this, I take it, is that Bar-On assumes that in conversational contexts, people’s self-ascriptions are never spontaneous but are always prompted, because, for example, they’re given in response to a question. For self-ascriptions to qualify as avowals they cannot be prompted. But suppose that during a conversation with a close friend, Harry self-ascribes the feeling of grief over the loss of his father.\(^5\) Harry clearly does not just ‘vent’ or ‘air’ his feelings; his self-ascription requires effort, and is only offered because his friend knows him well and feels confident to push him a little. It is only after a series of inquisitive questions that Harry finally has the courage to speak his mind.

What would an expressivist say about a case like Harry’s? First of all, Harry’s self-ascription is offered in the context of consultation with others, which appears to

\(^5\) This is intended to be a case of a self-ascribed feeling rather than, for example, a trait or a description of a pattern of behaviour.
be incompatible with avowing anything. Second, Harry has reasons or grounds for issuing his self-ascription other than speaking directly from his feelings: he offers the self-ascription because he wants to feel better and because he wants to stop his friend giving him a hard time, etc. Third, it's also not unlikely that Harry's self-ascription is made on the basis of evidence. Maybe, during the conversation, he has, if only very briefly, considered his feelings over the last couple of months. Maybe he imagined his father and noticed a distinct pang of sadness, on the basis of which Harry inferred that it is grief that he had been feeling (cf. Lawlor 2009). Or maybe it's only when Harry hears himself say the words that he is still grieving following the death of his father that everything locks into place and that Harry finally realizes that this has been what was on his mind, or what was bothering him, and only now does he discover his feelings.

Despite the fact that Harry seems to have spoken his mind, nearly everything seems to be going wrong, from an expressivist's point of view. Again, we have two options. Either Harry spoke his mind but has not 'avowed' anything, which is implausible for the reasons given above, or he did not speak his mind and so did not offer an avowal, which again seems implausible.

The more general problem regarding the notion of avowals is that whereas the notion of an avowal is meant to be more or less intuitive, it turns out to be a technical one.6 Worse, it seems that only expressivists can tell us what it means to (truly) avow your mental state. This is illustrated, for instance, by Bar-On's discussion of so-called 'negative avowals'. Negative avowals involve the ascription of the absence of a mental state, such as "I am not feeling hungry", "I do not believe that p", "I am not angry with you". Such ascriptions can, it seems, be true, and they are in many ways relevantly similar to positive avowals of the sort I've been discussing (e.g. "I am feeling hungry").

The problem is that when it comes to negative avowals, there is no corresponding first-order state, and therefore they seem to pose a problem, for there is nothing for the subject to express or speak 'from'. To deal with this problem, Bar-On first of all claims that negative self-ascriptions are "not self-ascriptions of absence but rather ascriptions of 'complementary' mental states" (Bar-On 2004, 334). So "I do not think P or "I do not want X" is, she suggests, a positive avowal in disguise, namely "I think not-P" or "I want not-X". This strategy is not entirely satisfactory,

6 Snowdon (2012, 246) makes the same point against Crispin Wright.
though, because the meaning of the complementary positive self-ascription is not necessarily the same as the negative avowal. As Brueckner writes, when I avow “I do not believe there are exactly ninety-nine wombats in the Ukraine” then I am not thereby expressing a belief that there are not ninety-nine wombats in the Ukraine. What I do, rather, is suspend judgement on the Ukrainian wombat population” (Brueckner 2011, 186). It seems plausible that there will be negative avowals for which we cannot find a semantically equivalent positive complement (“I am not angry with you” or “I do not feel jealous” seem to be good candidates). This is where Bar-On turns to a second strategy and responds by saying that if there are examples of where we cannot find a semantically equivalent complement of a negative avowal, then these aren’t avowals; rather, they are alienated or theoretical self-reports in disguise. She writes, “I suspect that actual cases fitting this bill will not be easy to find. But if we find them, I see no compelling intuitive reasons to regard them as avowals” (Bar-On 2004, 334). These ‘seeming’ avowals are actually self-interpretations camouflaged as expressive self-ascriptions. This is strange. After all, these self-reports in disguise are phenomenologically identical to avowals.

My main worry here is chiefly methodological: what justifies our seeing a particular ascription as a self-interpretation rather than an avowal, if what it seems like from the subject’s point of view does not provide us with the answer? Aidan McGlynn (2011) claims that the fact that Bar-On cannot handle expressions in the absence of some mental state is the “biggest problem” that expressivism faces. Although I think McGlynn is absolutely right that it’s important for expressivism to be able to deal with negative avowals, what seems more problematic to me is that expressivism turns out to rely on a rather technical notion of avowals – one that, coincidentally, is congenial to the expressivist thesis – and that only the expressivist seems to be able to distinguish between seeming and genuine avowals. In other words, the expressivist can ‘handle’ expressions of the absence of some mental state, but can only do so at the cost of making the concept of avowals, and so expressive self-ascriptions, rather elusive.

The overall point is that it’s unclear what avowals are supposed to be. To avow a mental state is, apparently, not just to speak your mind sincerely but to do so in a very specific way, i.e. doing so in a way that is non-interpretative, not prompted, not in conversational circumstances, and so on. Bar-On does mention, in a footnote, the contrast between ‘avowals proper’ and what she calls ‘mixed avowals’, i.e. avowals that have “descriptive or reportive elements” (2004, 304). The question is how we
know whether we're dealing with a mixed avowal or with a self-report. Only the first, after all, delivers privileged self-knowledge. If we don't know when our self-ascriptions are expressive, or qualify as avowals, and when they are not, then we don't, on closer inspection, know which of our self-ascriptions represent privileged self-knowledge and which of them do not. The more fundamental problem is not just an epistemic one – of how to tell the two apart – but a metaphysical one. Are there really such things as avowals (proper), to be distinguished from other types of self-ascriptions? We may have reason to be sceptical.

One option, of course, is to ignore all of the ‘mixed’ cases and restrict the expressivist account of self-knowledge to avowals proper, that is, avowals that are truly unprompted, spontaneous, unguarded, and so on. It seems we do sometimes offer these sorts of self-ascriptions, e.g. when I run into you and immediately I say “I'm so happy to see you!”, or if I see a piece of carrot cake and I cannot help myself from saying “I want that!”. The problem with restricting expressivism to avowals proper, though, is that the scope of expressivism is very narrow, because many of our self-expressions aren't ‘proper’ in the intended sense. Arguably, the sort of self-ascriptions that do fit the ‘avowal proper’ bill involve rather primitive ways of expressing ourselves. We can, and often do, control when we speak our minds, how we speak our minds or to whom we speak our minds. Pace Ryle, this is not the “sophisticated” but rather the natural thing to do – the less natural thing to do would just be to vent or air our thoughts and feelings whenever and wherever they came to us. My understanding is that one of the things that distinguishes adults from infants is precisely the fact that their self-ascriptions can be guarded to greater and lesser degrees, respectively.

One of the strengths of contemporary expressivist accounts is insisting that speaking our minds is an intentional act. But, ironically, this strength is arguably also its weakness. For, given that speaking our minds is intentional, this also means that we can, to some extent, control whether we speak our minds at all, or choose not to do so. In other words, if we see avowals as intentional actions, which seems plausible enough, then this has as its consequence that usually our self-ascriptions aren't as spontaneous, unprompted and unmediated in the way that non-mixed avowals are supposed to be. If we exclude all of our quasi-mediated or quasi-prompted self-expressions, then nothing much is left for the expressivist to give an epistemic account of.

Where does this leave us? Understanding expressivism along moderate atomist lines means that, in normal circumstances, you come to know your mind by
speaking your mind. What follows from the above is that it isn’t clear what this
comes down to, because speaking one’s mind turns out to require that one ‘avows’
one’s mental states, and the notion of avowal turns out to be a technical notion.
The expressivist seems to be required to say that many cases of what appear to be
perfectly good and sincere instances of speaking one’s mind either are not, despite
appearances, instances of speaking one’s mind after all, or they are instances of
speaking one’s mind but are not instances of ‘avowing’ something. Both options,
though, seem to conflict with the idea that the expressivist’s “starting point is a
familiar phenomenon, which can be characterized ostensively as it were” (Bar-On
2004, 26).

This has consequences for the atomist approach to expressivism, because if we
don’t know whether or not someone’s self-ascription counts as an avowal, then we
don’t know which of our self-ascriptions provide privileged self-knowledge and
which do not. To the extent that it is clear (‘avowals proper’), there’s the worry that
the scope of the expressivist theory only covers rather primitive self-ascriptions,
which would make expressivism a lot less plausible in terms of proposing to offer a
‘theory of self-knowledge’.

In the next section I assume, for the sake of the argument, that we can help
ourselves to the notion of an avowal and that one offers an avowal when expressing
or speaking one’s mind. The question I address is whether speaking one’s mind
always delivers self-knowledge, as the radical atomist view claims, or whether
it only does so in normal circumstances. This raises the question of what those
circumstances might be and whether the exceptions don’t end up becoming the
rule.

3. Ceteris Paribus II: assuming I am sane, sober and not asleep

The idea that there are abnormal situations in which one’s avowals do not provide
self-knowledge seems widely assumed but it is rarely made explicit, let alone
explained. But, for instance, in his article ‘Expressing’, William Alston writes
that if a self-ascription is made “in a fit of abstraction, its indicative value will be
impaired if not altogether lost” (Alston 1965). Roessler agrees with Alston’s main
idea, to which he adds that the epistemic value of self-ascriptions is undermined not
just in a fit of abstraction but also if made “during sleep, or in a state of advanced
intoxication" (J. Roessler 2015a). And Victoria McGeer writes (though not in the context of expressivism), “If I claim to be upset or happy about something or to have a yearning for plum pudding, then, all things being equal (i.e. assuming I am sane, and sincere, and not deeply distracted), the appropriate default presumption is that such claims are true” (2007a, 81). So, in general, we might say that in the presence of a ‘normal consciousness’, expressing one’s mental state amounts to knowing one’s mental state.

What this shows, first of all, is that even if you take an atomist approach to self-knowledge and think that the question of what makes a person’s self-ascription true can more or less be answered in terms of the procedures that the self-ascriber did or did not follow, you are not committed to the implausible view that the self-knowledge procedure delivers self-knowledge in all (psychological) circumstances. Atomists, including expressivist atomists, can after all allow that psychological circumstances, like those mentioned above, matter to whether or not a self-ascription qualifies as self-knowledge by claiming that the relevant procedure only provides self-knowledge ‘all else being equal’ or under the ‘normal’ circumstances of self-ascription. According to the expressivist view understood along atomist lines, then, all you need to do in order to have self-knowledge is to follow the relevant procedure – to speak your mind – and to be sincere, sane and sober.

The *ceteris paribus* clauses mentioned by Alston, Roessler and McGeer can be taken to be uncontroversial because the circumstances that could undermine self-knowledge are highly exceptional or abnormal and so typically irrelevant for thinking about self-knowledge. Only in a handful of abnormal or extreme cases, such as ‘advanced intoxication’ or when one is ‘deeply distracted,’ do expressive

---

7 As discussed in the previous chapter, Roessler (like Alston) explains the epistemic privileges of expressive self-ascriptions by appealing to the idea that expression is an intentional action. On Roessler’s account, the value of a self-ascription of an attitude as a source of knowledge of the speaker’s attitude depends on the utterance being intentional, with the speaker realizing what she is doing. The idea here is that in such circumstances, one arguably does not “know what one is doing”, and hence the self-ascription fails to express self-knowledge. The question thus becomes in what circumstances one fails to know what one is doing.

8 In all fairness, McGeer here is making a claim about first-person authority, not self-knowledge. But given that many philosophers including McGeer do not properly distinguish between the two (or they start out by doing so, but end up conflating them again), I will not discuss the distinction here and will assume here that if we cannot assume someone is sane, then “the presumption that [a person’s self-ascriptions] are true” (2007a, 81) no longer holds. Characterizing McGeer’s remarks as having this implication is somewhat unfair, but doing so is instrumentally convenient so that the issue can be brought to a head in this chapter. For what it’s worth, I don’t think it’s plausible that ‘sanity’ should be a condition for having first-person authority, either.
self-ascriptions fail to deliver self-knowledge. I'll call these circumstances – intoxication, or being asleep or deeply distracted – 'exceptional circumstances'. The question I want to address is whether such an appeal to the 'normal' psychological circumstances is clear enough and so legitimate or not. If not, then this raises a challenge for the moderate atomist view of self-knowledge.

To start, I want to observe that the sort of psychological circumstances mentioned so far do not necessarily threaten the epistemic value of a self-ascription. Apart from the folk wisdom of the saying in *vino veritas* (which suggests that being intoxicated might, in some cases, give you an insight into what you care about most or whom you love, or indeed might give you the confidence required in order to come to know it), there seem to be quite a number of things that one can unproblematically know while being in a state of 'advanced intoxication', for instance the desire to lie down or have a glass of water or that one is experiencing a headache and wants an aspirin. When a heavily intoxicated stranger (sincerely) tells you he wants to marry you, you might rightly take the epistemic value of his self-expression to be 'impaired if not altogether lost'. But if he tells you he is feeling sick and needs to throw up, you'd better believe him. In other words, he seems to know his own mind in spite of being severely intoxicated. Also, 'being asleep' need not threaten (all types of) self-knowledge. A sleepwalker might still know that she is hungry while (sleep)walking towards the fridge.

Nor does it seem plausible that being 'severely distracted' is a condition that we should necessarily include in *ceteris paribus* clauses. If someone is in a fit of anger, say, and is inclined to sincerely, but falsely, express herself in ways she will regret afterwards ("I wish you were dead", "I want a divorce", and so on), but then is suddenly deeply distracted by something (she witnesses a car accident, say), her being distracted might actually allow her to regain self-knowledge by allowing her to cool off and put her resentful thoughts and feelings into perspective. Distraction can be a good thing if your current thoughts or feelings do not represent your actual attitudes. Consider the following passage from Iris Murdoch:

I am looking out of my window in an anxious and resentful state of mind, oblivious of my surroundings, brooding perhaps on some damage done to my prestige. Then suddenly I observe a hovering kestrel. In a moment everything is altered. The brooding self with its hurt vanity has disappeared. There is nothing now but kestrel. And when I return to thinking of the other matter it seems less important. (Murdoch 1970, 67–68)
Sheila Mullett (1987) suggests that Murdoch is going through what she refers to as an 'allocentric' shift—a shift away from her own inner states towards the world—and thereby comes to see her own attitudes anew and learns something about herself and her attitudes in the process (1987, 317). Maybe this is not what was going on at all in Murdoch’s case, but the idea that distraction is necessarily a circumstance in which one cannot acquire self-knowledge is certainly too strong.

My main concern with the appeal to ruling out exceptional circumstances is not that these exceptional circumstances aren’t necessarily undermining, as suggested above, but rather that the circumstances that threaten the epistemic status of our sincere self-ascriptions are not particularly exceptional, i.e. that the list is (far) from sufficient. The idea would seem to be that if you speak your mind, provided that you are sober, sane and sincere, then you know your own mind. But being sober, sane and sincere is no guarantee of expressive self-knowledge, for there are many circumstances that are not particularly exceptional or extreme in any way but which still raise important challenges regarding the epistemic status of people’s self-ascriptions.

As we’ve seen, Alston mentions ‘severe intoxication’ as a potentially undermining circumstance. But what about just being tipsy? You might be out for drinks with a colleague, and you end up having one drink too many, which leads you to self-ascribe, rather impulsively and naively, the intention to be at work on time tomorrow. You’re not drunk, let alone severely intoxicated, and yet your psychological condition makes it such that your self-ascription of your intention is implausible. Having had too much coffee can also lead people to speak their minds where doing so fails to deliver them with self-knowledge. Suppose I sincerely express feeling nervous for a meeting or a telephone call later in the afternoon, but come to realize that I have probably drunk too much coffee. In such a case, I have misinterpreted some of the bodily cues that are similar to both anxiety and having had too much caffeine. Having had too much coffee can hardly be considered an exceptional psychological condition, but still, in at least some cases, being ‘high on caffeine’ might lead to expressive failures.

Being sleep-deprived is another possible ‘self-knowledge defeater’. Suppose Sally suffers from insomnia and hasn’t had a proper sleep in days. When she wakes up, she feels reluctant to get out of bed and to go outside, and she shrinks from the thought of having to go to her friend’s dinner party. Sally knows her occurrent thoughts, but does she have knowledge of what she wants? Does her sincere self-
ascription “I don’t want to go to the dinner party” accurately represent her attitude? It does not seem implausible to think that not having had enough sleep might make Sally think that not getting out of bed is what she really wants, even though it’s not. Perhaps one doesn’t need to be sleep-deprived for one’s self-expressions to fail to provide self-knowledge – perhaps one simply has to get out of bed on the wrong side.

Being hungry is another potentially epistemically relevant psychological condition. For some people, being hungry makes it hard to know what they believe, feel or want. Consider Hungry Tess. When we ask Hungry Tess whether she wants to go, say, to the cinema later this evening, she expresses her desire to stay at home. If we ask her again, she will probably get irritable. After having had some lunch, however, Tess feels and thinks differently about the subject matter. Not only does she regret her being irritable but she now thinks that she didn’t see things clearly before: she didn’t know what she was saying when she self-ascribed the desire to stay at home rather than go to the cinema. When Hungry Tess self-ascribed the desire to stay at home, it was, as we might say, not Tess but rather ‘the hunger talking’. Again, hunger is hardly an ‘exceptional’ circumstance or a condition that we can just rule out from the outset. The following statement is, after all, somewhat odd: ‘speaking your mind allows you to know your mind, unless you’re hungry’.

It’s also plausible that whether or not someone acquires self-knowledge by speaking her mind might depend on the stakes or the costs of being wrong. This seems plausible, in particular for many of our more complex attitudes: you’re more inclined to make expressive failures if a lot is at stake when you are speaking your mind. But even in relation to much more simple self-ascriptions, such as self-ascribed sensations, we can imagine that the question of whether someone has self-knowledge is context-sensitive. An excellent example of such a case is provided by Paul Snowdon (2012):

Imagine you are having troubling pains in your stomach. The doctor after an examination says: “It is one of two complaints. There is a test to determine which. I shall apply pressure to a region and it will cause pain in two areas—one

---

9 Again, just as in Murdoch’s case, this isn’t necessarily the case. You can be sleep-deprived and genuinely have a desire to stay in bed, and know that you do. The point is only that it is possible in at least some cases that your original self-ascription was in fact false and that your occurrent thoughts and feelings misrepresented your actual attitudes.
left and one right—and which ever hurts most tells us what the problem is.” He then applies the pressure and you judge that the left area hurts most. The doctor then says: “Oh dear, that means that you have a very serious condition, requiring drastic treatment. If it had been the right side then it would have been minor.” Confronted with this, do you say: “Since my judgment is incorrigible go ahead with the drastic treatment?” Or might you say: “Do you mind if we repeat the test? Maybe I was a little bit hasty—I would like to make sure and check things”? (Snowdon 2012, 359)

Snowdon rightly points out that in such a ‘high-stake’ case, we would opt for the second sort of response, not the first. Snowdon’s example is meant to challenge the idea that beliefs about our sensations are supposed to be incorrigible, but for present purposes, we can also take his example to illustrate the fact that many of our self-ascriptions—not just the more complex self-ascriptions of our attitudes but also more simple self-ascriptions of our sensations—can be importantly context-sensitive or ‘stake-dependent’. It seems plausible to think that in the sort of example given by Snowdon, the initial avowal (e.g. “My left side hurts the most”) did not provide the subject with knowledge of her sensations. And yet it seems implausible to include these sorts of cases in the original ceteris paribus clauses. It’s implausible to say, for instance, that speaking your mind provides you with self-knowledge, assuming the implications of your self-ascription do not determine whether or not you’re seriously ill.

The overall message should be clear: human beings tend to get tipsy, distracted, confused, sick, nervous, angry, depressed, uptight, drunk and sleepy, become euphoric, jealous, foolhardy and insecure, and fall in love. In a great many cases, these circumstances—hardly exceptional—cause us to think, act and express ourselves in ways such that our self-ascriptions do not amount to self-knowledge. Certain circumstances make us speak our minds too easily (e.g. being in love, high on adrenaline, enthusiastic, the subject in an experimental study), the result of which is that our self-expressions do not provide self-knowledge. Also, making up our minds in certain circumstances might even lead us to (greater) self-deception (e.g. depression or insecurity, being under stress or dealing with hormonal fluctuations).

Interestingly, the idea that non-exceptional circumstances can impede self-knowledge is widely accepted when we consider the sincere self-ascriptions that young infants make. Children are more often liable to make expressive failures.
When considering young infants, the epistemic impact of being tired, thirsty, angry, and so on is especially salient. Parents are typically aware that their children are not fully competent self-ascribers yet, and they sometimes know what state their child is in better than the infant, particularly by taking the circumstances into consideration. For instance, imagine a toddler fretting and crying for a while, and finally saying “I’m thirsty!” The toddler’s self-expression is sincere, even spontaneous, and yet his or her parents might not want to take the toddler’s self-report at face value because it doesn’t express the toddler’s real or genuine desire—which is, simply, to get some sleep.

In the case of young infants, we find it intuitive to give circumstances their due when it comes to whether or not our expressive self-ascriptions amount to self-knowledge. Outside philosophy, we seem to take the epistemic relevance of certain circumstances seriously as well. But these considerations have had surprisingly little impact on the philosophical debate on self-knowledge, including the literature on expressivism. My suggestion would be to realize that there isn’t a clear division between the epistemic value of infants’ self-expressions and the epistemic value of self-ascriptions of sincere, sober, sane and mature adults. Epistemically undermining psychological circumstances were relevant when we were young, and they remain relevant when we grow older.

I don’t want to suggest that, on closer inspection, we hardly ever know what we want, think or hope for, because of this fact of human nature. The point, rather, is that we should not be mistaken in thinking that there is some kind of neutral psychological state that would provide the ‘normal’ circumstance of self-knowledge. There is no determinate list to be given of what the relevant ceteris paribus clauses are. The reason is that what counts as epistemically undermining circumstances is something that needs to be assessed in relation to the person who issues a self-ascription and the life that she leads. Whereas for Hungry Tess being hungry might be epistemically undermining, it might not be epistemically undermining for Hungry Tim. Indeed, being hungry might be epistemically undermining for Hungry Tess on Tuesday, but not on Thursday. Likewise, being tipsy might lead Peter to issue self-expressions that lead to self-deception, whereas being tipsy might lead Petra to make self-expressions that instead lead to self-revelation.

All this points in the direction of a holist rather than an atomist expressivist view. A holist version of expressivism would hold that the question of what makes someone’s self-ascription true cannot simply be answered, assuming everything else is equal, by zooming in on the self-ascriptive act (is it an expression, i.e. was it...
sincere and 'unguarded'?). On a holist view of self-knowledge, the question of what makes someone's expressive self-ascriptive true (or false) is something that depends at least in part on the attitude type, the content of the attitude, the self-ascriptor's character traits and her earlier actions (history), as well as on her dispositions (future) and the context in which the self-ascriptive was made. A holist view would say that self-ascriptions have their proper place within the life of a person and that the question of what makes someone's self-ascripted attitude knowledgeable must be answered relative to that life and not just relative to her expressive act.

An expressivist might object that the criterion for expressive self-knowledge isn't that one simply acquires self-knowledge when expressing oneself, but, more specifically, when one speaks one's own mind. My response to this objection is twofold. First, the distinction between expressing 'a' mental state versus expressing one's 'own' mental state is not unproblematic (see previous chapter). Second, even if we can talk meaningfully about expressing one's own mental state rather than 'a' mental state, then this still leaves the question of what counts as 'my' mental state.¹⁰

One might propose to answer this question of ownership by appealing to some sort of mental realism, as Bar-On proposes. But this, as discussed in the previous chapter, seems to be a somewhat evasive strategy, because the question of what is required for an (expressed) self-ascriptive to count as self-knowledge is simply outsourced to an appeal to mental realism, the latter of which is usually not elaborated upon much.¹¹

The question of ownership becomes particularly pressing when we consider that, first, many of our attitudes are not restricted to the here-and-now, but instead have "a reasonable life expectancy" (M. Fricker 2009, 52) and second, that one can have a belief that P, hope for Q or desire for R without consciously considering P, Q or R at a specific time. This challenges the idea that one only acquired self-knowledge when one expresses one's 'own' mental state, because what the subject's own mental state is, is not something she needs to be actively considering or consciously thinking about. To say what makes a belief or desire 'mine', we need to say something about the me whose attitude it is. This is just another way of saying that one cannot explain the idea of what it means for a mental state to belong to a subject, or what it means for a mental state to be the subject's own, without mentioning the subject's (mental) life.

¹⁰ For a different but related discussion regarding the so-called experience of 'mineness' of one's mental states, and a suggestion of how to explain such mineness in coherentist-narrative terms, see Slors and Jongepier (2014).
¹¹ I return to the claim regarding the relevance of metaphysics for questions concerning self-knowledge in the final chapter.
The point is that self-expressions as such don’t necessarily give us self-knowledge. One might after all express oneself rather badly. In order for self-expressions to allow for self-knowledge, one needs to make an extra step. More specifically, one needs to make an interpretative step from one’s self-expression (which functions as material for self-interpretation) to the conclusion that one is (likely to be) in a mental state. For example, I might infer on the basis of the fact that I spontaneously expressed the desire to jump in the fountain that I have the desire to actually jump in the fountain. But I might come to a rather different conclusion on the basis of the same ‘expressive evidence’—an equally possible conclusion is that the thought merely occurred to me to jump in the fountain. I don’t have a genuine desire to jump in the fountain. In other words, our self-expressions are best understood as forms of self-interpretation, and so expressivism is best understood along ‘interpretationist’ or ‘inferentialist’ lines (Ryle 1949; Bem 1972; Gopnik 1993; Cassam 2014, 2015a, Lawlor 2003, 2008, 2009). In the next section, I elaborate on what an interpretationist version of expressivism, which would be compatible with a holist approach, might look like.

4. Self-expression as self-interpretation

Of particular interest in this context is Charles Taylor’s (1985) expressivist view. He argues that as ‘self-interpreting animals’ we are bound up in a process of articulating ourselves and that such self-interpretations shape us by informing and changing our attitudes. On Taylor’s account, expressing ourselves involves ‘articulating’ what we feel and think. Our linguistic articulations of our thoughts and feelings enable us to see whether or not our attitudes or emotions become more focused, intelligible or coherent to us as a result (or the opposite). The point of self-expression is “to get clear about how I feel by letting my emotions take shape and find a voice in what I say and do” (Carman 2003).

The central interpretationist claim is that self-knowledge is a matter of interpreting what attitudes we have on the basis of the available evidence. As Quassim Cassam points out, there are many different kinds of evidence: “One kind of evidence is behavioural but there are other possibilities; you can discover your own standing attitudes on the basis of your judgements, inner speech, dreams, passing thoughts and feelings” (Cassam 2014, 138). Krista Lawlor (2009), for instance, makes a convincing case for the idea that knowledge of one’s desire is derived from inferences of what she calls ‘internal promptings’. Internal promptings can include
simple sensations but might also include one's emotions, imaginings and goings-on in inner speech.\textsuperscript{12} Inference from internal promptings, she says, is "a routine means by which we know what we want" (2009, 48).

The suggestion would be to simply add 'self-expressions' to the list of 'promptings'. For instance, articulating one's feelings -- e.g. expressing one's anger or desire -- might give one "a sense of ease or settledness" (Lawlor 2009, 57). This feeling might be a (good) indication of whether you really are angry or whether you really do have the self-expressed desire. Or you might order a glass of champagne (by avowing "I want some champagne") and immediately realize that "what you actually want is a vodka martini" (Cassaro 2014, 143). The same, Cassaro points out, may well apply to one's beliefs: "you say you believe the present government will be re-elected but the minute you say the words you realize they don't ring true" (ibid.). Even though beliefs and desires aren't feelings, "what you feel can sometimes tell you what you believe or desire" (Cassaro 2014, 144). Lawlor's and Cassaro's suggestions seem congenial to the sort of view Taylor defends, according to whom we come to experience the accuracy of our new self-expressions on the basis of what these words stir up in us.

If avowals are self-interpretations, then it seems we would have to give up on the idea that expression is a 'success notion' and would instead have to see expression as an "evidential" notion (Davis 2003, 47; Eriksson 2009, 217-19). This means that self-expressions are (good) 'indications' on the basis of which one can infer what one's attitudes are. According to Alston, if $x$ has "indicative value" relative to $y$ then "from $x$ one can (fairly safely) infer the existence of $y$" (1965, 20). Footprints in the snow, for instance, give us a good reason to think that someone was there before us. Applied to expressive self-ascriptions, the idea is that a self-ascription, such as "I feel bored", gives me a good reason to think that I am genuinely in the mental state of feeling bored. As Eriksson remarks, 'expression as indication' is a weaker notion than 'expression as showing':

\textsuperscript{12} Lawlor gives the following example: "Katherine, a young woman, stands by her son's crib, watching him sleep. 'Have another'--she hears the words in her head. She is startled. Was she thinking about that? She supposes she was, but she wonders: was that a directive from some part of her? Or was it a question? Or was it something else?" (Lawlor 2009, 47). Lawlor imagines her to acquire self-knowledge in the following way: "[H]ow will Katherine find out what she wants? Now that the question has been called, Katherine starts noticing her experiences and thoughts. She catches herself imagining, remembering, and feeling a range of things. Putting away her son's now-too-small clothes, she finds herself lingering over the memory of how a newborn feels in one's arms. She notes an emotion that could be envy when an acquaintance reveals her pregnancy. Such experiences may be enough to prompt Katherine to make a self-attribution that sticks. Saying 'I want another child', she may feel a sense of ease or settledness" (Lawlor 2009, 57).
it is possible to indicate, in the sense outlined, that one has a particular state of mind without actually having it, i.e., one can express, e.g., the belief that p without having the belief that p. ... Asserting that p gives us reason, albeit not conclusive reason, to think that the speaker believes that P. (Eriksson 2009, 218)

The above also holds for the speaker himself. When Larry speaks his mind and expresses his desire to stop studying for his PhD, then this gives him a good reason, though not a conclusive reason, to think that he wants to stop studying for his PhD. Differently formulated, Larry’s spontaneous and sincere self-ascriptions expressed in private or in public form a good basis on which Larry can come to know what his desires are. So when we speak our minds, we do not necessarily show our states; rather, we give (very good) evidence of what state we’re in.

There are different ways of understanding expressivism as a version of interpretationism. For instance, one might think of one’s mental states, of which one’s self-expressions form (good) evidence, either along realist or along constructivist lines. On a realist view, one might think that there are certain facts of the matter as to what one’s attitudes are, and that speaking your mind isn’t constitutive of them. Alternatively, one might think that one’s self-expressions and other ‘promptings’ can only be evidence of one’s attitudes, depending on how one sees, conceives, feels about or interprets these promptings. In other words, a person’s self-expressions would be very peculiar ‘indications’; they’re not like footprints in the snow, because the subject’s relation to her own self-expressions is, at least in part, what would make them reliable indications in the first place. This, I take it, is Taylor’s view, according to whom “our interpretation of ourselves and our experience is constitutive of what we are” (C. Taylor 1985, 47). He suggests that we must understand the notion of ‘self-interpretation’ as a form of self-understanding, where such self-understanding is reflexive in the sense that we relate our attitudes to our other attitudes and projects and what we care about or find important. Taylor proposes that “our attempted definitions of what is really important can be called

---

13 Lawlor seems to lean towards a realist approach, as her understanding of the notion of self-interpretation seems to suggest. Self-interpretation, for Lawlor, is a causal notion. She writes, “In interpreting the import of these rehearsals, one does a special sort of interpretive work: ... one interprets one’s imaginings for their likely cause, for the trace of desire they carry. One engages in causal self-interpretation” (Lawlor 2009, 72–73, see also 2008). Lawlor continues, “Assigning a cause is part and parcel of identifying the mental attitude that underlies the imaging”, because doing so is “a way of identifying one’s imaging as the product of an intending, a remembering, or a desiring, to a specific effect. Assigning a cause is knowing one’s mind” (Lawlor 2009, 73).
interpretations” (1985, 75). On Taylor’s account, the question of what our attitudes
are is not independent of our expressive efforts to interpret them and thereby try to
make sense of them. Taylor suggests that expressing our minds allows us to relate
to our mental states in a new and different way, simply because we have made our
thoughts or feelings explicit.

However, whatever interpretationist view one favours, it’s clear that construing
expressivism along interpretationist and, as I have suggested, holist lines requires
giving up on what appears to be one of the most fundamental if not the most
fundamental expressivist claim, namely that when we speak our minds, we make our
mental states directly visible to others and that speaking our mind involves “showing
us the conditions they express” (Bar-On 2004, 271). The question, then, is whether
there’s anything still distinctively expressivist about an interpretationist expressivist
view, or indeed a holist expressivist view according to which the question of whether
someone’s sincere self-ascriptions provide knowledge can only be answered by
seeing how the self-ascription relates to the specific (psychological) circumstances
the person is in, the relation her self-expression has to her other attitudes, values,
self-conception, character traits, and more generally, how the subject relates to
herself.

So far, I’ve only considered a moderate atomist version of expressivism, and
have suggested that this view isn’t plausible; if that is true, it would be a reason
for considering what a non-atomist expressivist view would look like. However, the
alternative is to go radical: to deny that there are undermining circumstances at all
and to say that whenever one avows being in mental state M, one is necessarily in
the relevant mental state. This involves a retreat to what I referred to as the radical

---

14 This also seems to be a problem for Taylor. In addressing the question of what is meant by ‘expression’, Taylor
writes, “I think it means roughly this: something is expressed, when it is embodied in such a way as to be made
manifest. And ‘manifest’ must be taken here in a strong sense. Something is manifest when it is directly available
for all to see. It is not manifest when there are just signs of its presence, from which we can infer that it is there.,
such as when I ‘see’ that you are in your office because of your car being parked outside” (C. Taylor 1985, 219).

15 By the latter I mean that I expect that any holist construal of expressivism, or any holist account of self-knowledge
more generally, must include a ‘reflexive’ component regarding the subject’s self-understanding, for example how
she relates to her own beliefs, intentions, personality, and so on. David Velleman characterises reflexivity as follows:
“Some activities and mental states have an intentional object: they are mentally directed at something. Of these,
some can take their own subject as intentional object; they can be mentally directed at that which occupies the state
or performs the activity. Of these, some can be mentally directed at their own subject conceived as such—conceived,
that is, as occupying this very state or performing this very activity. A reflexive mode of presentation is a way of
thinking that directs an activity or mental state at its own subject conceived as such” (Velleman 2006, 2; see also
Döring and Düwell 2015).
atomist version of expressivism. But in that case, we would have to say that all the non-exceptional circumstances that I have mentioned, e.g., anger, intoxication, sleep, hormones, and so on, have no impact on the epistemic value of a self-ascription. Faced with this alternative, it seems better to accept the holist-interpretationist alternative just sketched.

In the next two sections, I want to argue that the atomist version of expressivism can be traced back to the reluctance to see avowals or sincere self-ascriptions (I shall use these notions interchangeably) as genuine assertions, and I suggest that self-ascriptions are communicative acts—a view congenial to the holist-interpretationist approach suggested above.

5. Diagnosis: self-expression and communication

Contemporary expressivists like Bar-On want to retain a delicate balancing act between seeing so-called avowals as instances of ‘airing’ or ‘venting’ on the one hand and seeing them as truth-evaluable and so semantically continuous with other ascriptions on the other. The first element of the balancing act is what makes the account expressivist, whereas the second element is what prevents one’s account from collapsing into simple expressivism. In this section, I want to explain why this balancing act cannot, in the end, be pulled off.

Recall that Bar-On is reluctant to see avowals as assertions. To express one’s mental state, on her view, is to “show the conditions they express, rather than merely communicate information” (Bar-On 2004, 273). The contrast is thus between showing something on the one hand and communicating something on the other, and Bar-On’s suggestion is that avowals are thus not instances of communicating information.

When you make use of a truth-evaluable sentence, such as “I’m so pleased to see you!”, “I’d like a cappuccino” or “I really need a holiday”, Bar-On claims you are not aiming to report or describe your mental state and are not in the business of informing your audience what mental state you’re in. In his article ‘Language as Thought and as Communication,’ Sellars makes the radical suggestion that the activity of thinking out loud is “a form of meaningful speech which doesn’t consist in talking to anyone at all, even oneself, and hence is not, in any ordinary sense, talking” (Sellars 1969, 518). Echoing Sellars, Bar-On writes that self-ascriptive
expressions are not made "with the aim of informing their audience of what is going on in them" and are not acts subjects undertake "to perform with a specific audience-directed goal in mind, such as convincing, informing, pleasing, etc. Like many non-verbal expressive acts, they may not even have any communicative point" (2004, 242). So when someone says "I think that P" or "I want X", she is not (intentionally) informing or telling anyone anything but simply expressing her view that P or her desire for X.

Avowals, then (whatever they are), are pseudo-assertions at best. Avowals are continuous with our ordinary means of communication only to the extent that we use the same semantic vehicles. The basic idea is that expressive self-ascriptions are not (proper) assertions because in issuing them, one has no 'communicative intention' to report what state one is in.16

However, I don't think the appeal to (the absence of) communicative intentions is a good reason to deny that avowals are assertions. Even though the basic insight that often our spontaneous and sincere self-ascriptions are not made with the intention of communicating something or informing anyone about 'how it is with one' seems plausible enough, I don't see how this should imply that what expressivists call avowals could not still be assertions or, more specifically, communicative acts. To say that most of our self-ascriptions are not made with any communicative intention is not the same as saying that they have no communicative point, unless one thinks that one can only be in the business of communicating something when one has the

16 But, one might ask, might it not be possible that the emphasis on the passage(s) just quoted is on the notion of aiming or intending to report or describe? After all, to the extent that Bar-On says that avowals are truth-evaluable (in order to avoid collapsing into simple expressivism), avowals must be reports or descriptions of some sort, or so it seems. And so all that she denies is that avowals are speech acts that one intends to be reports or descriptions – this does not mean she denies that they are assertions. However, looking at the text, what's more plausible is that Bar-On defines reports, descriptions and assertions as things for which one necessarily has to have an intention (Bar-On 2004, 294; 312). And so, given that when we speak our minds we often have no intention of stating to ourselves or others how things are with us, the self-ascription is not, therefore, an assertion or report. So avowals are not assertions in the end (rather than avowals being assertions, we merely have no prior intentions to assert). There's an important reason that counts in favour of this interpretation of Bar-On's account, i.e. seeing people's 'avowals' as instances of venting rather than assertions, which is that apart from the fact noted above that this is what seems to make Bar-On's account genuinely expressivist, it is also only on this understanding of avowals that Bar-On's account of first-person authority can be maintained. For if avowals are assertions, then epistemic challenges ("How do you know?"?) seem to be appropriate (I discuss this in more detail below). One of Bar-On's key claims is her explanation of why certain epistemic challenges are inappropriate, which she establishes by pointing out that avowals are not assertions. If the speaker wasn't making a claim or assertion about 'how it is' with him in the first place then it isn't the hearer's business to correct the speaker, because the speaker wasn't in the business of communicating or asserting anything. If he was asserting, then the challenges are appropriate. This is what I'll try to argue for.
intention or aim of communicating something, which seems false. I might stand in front of a painting and spontaneously say “This is the best thing I've ever seen!”, thereby not only expressing my mental state but also communicating to you that you should come and see it, even though I did not have any intention of doing so.

This raises the following question: what are assertions? Generally speaking, assertions are speech acts in which something is claimed to hold. The more specific question is what asserting anything essentially consists in and what a speaker is doing when she asserts something. One option that was already implicitly considered in the foregoing is to say that when one asserts something, one must have a communicative intention of informing the hearer about something (cf. Grice 1957). On a Gricean model of assertion, a subject $S$ asserts that $P$ by the utterance $u$ if and only if there is a hearer $H$ such that

- $i$. $S$ intends $u$ to produce in $H$ the belief that $p$
- $ii$. $S$ intends $H$ to recognize that i)
- $iii$. $S$ intends $H$ to believe that $p$ at least partly for the reason that i).

When I’m at the museum, and I point to a painting and say “That’s a Francis Bacon”, my utterance can be understood in Grice’s sense. I am saying (i) “That’s a Francis Bacon” in order to make you believe – or let you know – that this painting is a Francis Bacon, which (ii) I intend you to recognize, and (iii) I intend you to come to know that this is a Francis Bacon via my telling you so. As applied to self-ascriptions, this view of assertions seems prima facie implausible (as Bar-On claims). Indeed, even applied to non-ascriptive speech acts, the Gricean model is not uncontroversial because it’s (meta-)representationally demanding (see e.g. Bar-On 2013).

The question, though, is at what level we should understand the Gricean model generally and, as applied to avowals, specifically. As Bart Geurts and Paula Rubio-Fernández (2015) point out, drawing on Marr (1982), we can distinguish between the level of what agents do and why (the W-level) and the level of explanation that’s concerned with how agents do whatever it is they do (the H-level). As Geurts and Rubio-Fernández write, “Whereas a W-level theory describes a system from the outside, so to speak, an H-level theory deals in internal processes” (2015, 453). The
suggestion here would be to take Grice’s account (also) as an account that’s located on the W-level.  

When it comes to our expressive self-ascriptions, one might think this distinction isn’t of much help. When I say “I’m so happy to see you!” and we want to explain why I said this, it seems wrong to think that I am saying this in order to make you believe, or let you know, ‘how it is with me’, let alone that I intend you to recognize that I am saying this in order to make you believe how it is with me. And yet, I think, there is a sense in which we cannot properly explain why we ‘speak our minds’ at all, if it weren’t for letting others (and ourselves) know what we’re thinking and feeling. The reason why the idea that it seems wrong to think that I am saying this in order to make you believe, or let you know, how it is with me, I suspect, owes a lot to the H-level reading of the proposal. 

To see why avowals can plausibly be taken as assertions, as I want to suggest, what would help is to further distinguish different levels within the W-level. When I say “I’m so happy to see you!” and you ask me why I said that, it would be odd if I were to respond by saying “I intended you to believe that I am pleased to see you”. According to Geurts and Rubio-Fernández, a W-level theory describes a system from the outside. The why-question in this case was directed at my specific speech act, and at that level, the Gricean answer is odd. But I think we can, and must, zoom out further if we want to explain ‘why’ I was speaking my mind at all. Zooming out involves recognizing that the why-question concerns not the specific interaction between myself and my hearer, but rather concerns the place that expressing our minds has in our broader communicative practice, if, indeed, it has a place at all. 

One answer would be that on the zoomed-out W-level, what we should say is that I have no reason whatsoever. I’ve just blurted it out by accident, perhaps similar to the way a reflex works. This answer is clearly false, as contemporary expressivists also claim, given that they think that avowals are intentional actions. An avowal is a natural expression, and a natural expression is something “a creature does, or allows to happen, which is potentially susceptible to voluntary control, or at least subject to modulation in response to input that is directed at the behavior itself, even if it cannot be produced at will” (Bar-On 2004, 268). Saying “I’m so happy to see you!” or “I want some of that carrot cake” is something that is subject to voluntary control: I could not speak my mind, or I could speak my mind differently. 

17 NB: “The two perspectives will have to mesh, but how they will mesh is very much an open question” (Geurts and Rubio-Fernández 2015, 467).
I think this fact about how we express ourselves makes the following question salient: why do we speak our minds at all? A plausible answer, it seems to me, is this: we speak our minds at least in part to get something across, or to let others (and ourselves) know what we’re thinking and feeling. Seeing my self-ascription as a form of social interaction, more specifically as a form of communication, can explain precisely the intentional nature of our sincere self-ascriptions, however spontaneous or impulsive they might be. So I find much to agree with when Brueckner writes:

Though Bar-On does not say so, I take it that I may have communicative reasons [for avowing] as well ... I (1) intend to get my audience to believe that P, and (2) intend that he gets that belief as a result of recognizing my intention (1). Or so the Gricean story goes. (Brueckner 2011, 183)

Denying this would seem to constitute something close to a violation of seeing avowals as semantically continuous with other ascriptions. As discussed in the previous chapter, we should see both my utterance “I am standing” and my utterance “I am in pain” as expressing a (second-order) belief. But the desideratum of semantic continuity goes further than this. If it's true that my saying “I am standing” is meant to induce in you the belief that I am standing, and that I intend you to have this belief, then if we deny that avowals are communicative acts, what we get is that my saying “I am in pain” is not meant to induce in you the belief that I am in pain, which would mean that avowals and non-avowals are importantly semantically discontinuous.

---

18 There are arguably also indirect assertions, or, more generally, assertions that do not count as ‘assertions proper’ (e.g. “This is a Francis Bacon”). An example of an indirect assertion would be “Could you pass me the salt?” or an assertion that is implicit in requests, commands, rhetorical questions or ironic remarks. When I ask “Can you pass me the salt?”, my question, taken literally and directly, concerns your abilities (hence the response “Yes, I could” is unsatisfying), but indirectly, I'm claiming that there is salt on the table and am asking you to give it to me. And when I say, cynically, “Another lovely day!” when it's raining cats and dogs, then I can be said to be indirectly conveying information about the awful weather. If we do not find the W/H distinction compelling, we could at least model avowals on indirect or ‘weak’ assertions of this kind. This doesn't mean that any speech utterance is an assertion, of course. When I'm testing whether Skype works, and I say “It's raining” just to test it, and I have no idea what the weather is, then I'm not making an assertion, neither in its strong nor in its weak sense.

19 Rather than making the avowal follow the same norms as the non-avowal, the other option would be to say that in neither case does the Gricean story apply. Taken at the H-level, this seems right to me, but again, from a W-level perspective, I don't see what reason I have for intentionally expressing myself in the way that I do if it weren't for the reason of trying to communicate something.
The W-level approach seems to be one way of making sense of the way in which self-ascriptions are semantically continuous with non-avowals. Offering self-ascriptions is, it seems, not derivatively but intimately connected to one's trying to get something across, making oneself be understood, inviting the audience to think with the speaker or to correct her, and so on. All this can be taken on board without having to say that instances of speaking your mind are "acts that subjects deliberately undertake to perform with a specific audience-directed goal in mind, such as convincing, informing, pleasing, etc." (Bar-On 2004, 242). If we adopt a W-level approach to the practice of communication and the notion of communicative intentions, we do not seem to have any good reason to think that self-ascriptions don't serve the purpose of communication and so to deny that they are assertions. 20

An 'assertive speech act' account of self-ascriptions is suggested by Noel Fleming (1955), who takes issue with Ryle's expressivist claims. As we've seen, according to Ryle, we are tempted to mistakenly take avowals to be self-descriptions:

[I]n its primary employment "I want . . . " is not used to convey information, but to make a request or demand. It is no more meant as a contribution to general knowledge than "please". To respond with "do you?" or "how do you know?" would be glaringly inappropriate. Nor, in their primary employment, are "I hate . . . " and "I intend . . . " used for the purpose of telling the hearer facts about the speaker. (Ryle 1949, 164)

---

20 One question is what the norm of assertion is. A plausible view, it seems to me, is Timothy Williamson's view, which is to say that assertions are not claims to believe (truly and justifiably), but are simply claims to know, and that when one makes an assertion, one ought to follow the 'knowledge rule', that is, "one must assert p only if one knows p" (Williamson 2002, 243). Applied to self-ascriptions, this means that when I self-ascribe some mental state, I claim to know I am in that state. This is compatible with the fact that the rule can be broken, because I might claim to know that I am in some mental state or other when I'm not. And so, even though my attitudinal self-ascriptions are claims to know - i.e. assertions - they can still fail to amount to knowledge. As Williamson notes, "When one breaks a rule of a game, one does not thereby cease to be playing that game. When one breaks a rule of a language, one does not thereby cease to be speaking that language; speaking English ungrammatically is speaking English. Likewise, presumably, for a speech act: when one breaks a rule of assertion, one does not thereby fail to make an assertion. One is subject to criticism precisely because one has performed an act for which the rule is constitutive. Breaches of the rules of a game, language, or speech act may even be common. Nevertheless, some sensitivity to the difference - in both oneself and others - between conforming to the rule and breaking it presumably is a necessary condition of playing the game, speaking the language, or performing the speech act" (Williamson 2002, 243).
But according to Fleming, when someone says "I feel bored" then he is "giving information about himself, telling us how he feels, or what he wants to do, or whom he dislikes, much in the way that, in saying 'I have brown hair,' he is telling us the colour of his hair" (1955, 614). Fleming goes on to say that "we can learn something from the avowal in the way that we can learn something from 'I have brown hair,' and not just in the way that we can learn something from a yawn or a sigh, or from 'hurrah' or 'goodness me'" (1955, 618).

The assertive take on expressions has, I think, a number of advantages. One crucial advantage is that if we think of expressive self-ascriptions as assertions, then it makes a lot of sense to regard self-ascriptions, including one’s own, as possible interpretations or articulations of the state one is in. This, in turn, would appear to make room for a holist approach.

There are other advantages, too. I’ve already mentioned one such advantage, which is that it makes better sense of the fact that self-ascriptions, however spontaneous they might sometimes be, are still intentional (speech) acts. This leaves the place that avowals have in our broader communicative practices less of a mystery. But there are a couple of other noteworthy advantages, which I turn to below.

6. Avowals as assertions: theoretical advantages

The first advantage of the assertive approach is obvious: there is a natural way of accommodating the ways in which people’s self-ascriptions, e.g. "I am hungry", are semantically continuous with other ascriptions, such as "She is hungry", and thus provides a straightforward way of avoiding what I referred to in the previous chapter as a ‘simple’ variety of expressivism that is committed to treating them as semantically different. Semantic continuity is not, pace Bar-On, a desideratum of an expressivist approach, because on the assertive view, it’s built in from the beginning.

Second, the assertive take on self-ascriptions makes room for the idea that the subject and her hearer acquire knowledge of S’s mental state by virtue of S’s self-ascriptive act. More specifically, it does so without construing the subject as having privileged access to something that is in principle unavailable to her audience, or, differently put, without construing the audience’s knowledge as a type of testimonial knowledge, in the sense that the subject is, when ascribing some state
to herself, telling the hearer something that only the subject knows and has access to and that her audience lacks. Roessler (2015a) makes a convincing case for the idea that it is possible to claim (1) that the knowledge an audience acquires of a speaker's belief depends on the speaker's self-knowledge, without (2) committing to the traditional model according to which the audience 'inherits' such knowledge by 'being told' about something to which only the speaker has access. If we consider self-ascriptions to be assertions, then there is no longer a good reason to think that the audience gains knowledge of S's mental state by thinking of S as offering something like an eyewitness report. Roessler suggests that the audience's and the speaker's knowledge may share a common explanation, which lies in the speaker's sincere self-ascription. Roessler points out that both ways of knowing may be seen as "complementary roles, or as interdependent aspects of a single shared capacity for communication" (J. Roessler 2015a, 163).

Third, thinking that our sincere self-ascriptions serve a communicative purpose would explain the fact that one might speak one's mind more often, or more openly, to someone one knows well and/or trusts, compared to strangers. The fact that speaking one's mind is audience-relative in this way can be taken as evidence for the idea that sincere self-ascriptions are not just intentional actions but are intentional communicative actions. Harry's voicing his frustration about work is not just an act of venting his feelings but is probably also an act of informing Sally how he feels; perhaps it's a request for support, even if this is not at the forefront of Harry's mind. This, at least, would explain why Harry would not express himself in this way if Sally were not around (not even silently to himself). So when Ryle claims that the "natural" thing to do is "to is to speak one's mind, and the sophisticated thing to do is to refrain from doing this" (Ryle 1949, 162), perhaps he got things the wrong way round. In many cases, when you don't want to communicate your feelings (to a stranger on the street, say), the natural thing is keep things back, instead of just venting your thoughts and feelings ("I want coffee!", "I feel miserable!", "I need to go to the bathroom!"). The fact that, as Ryle puts it, being "spontaneous, frank and unprepared" in one's self-ascriptions would be 'natural' is only plausible assuming the speaker feels that she can speak freely in the communicative situation in which she finds herself.

Fourth, the assertive approach makes better sense of the way in which we actually treat our own and other peoples self-ascriptions in everyday life. It's certainly

21 Admittedly, Ryle himself begins by saying "[i]n a certain sense of 'natural'" (Ryle 1949, 162).
true that my saying “I am hungry” is different, epistemically, from my saying “I have blonde hair”, and philosophers have been right to try to figure out how this difference should be explained. But the preoccupation with such puzzles may have caused philosophers to overstate the differences in practice between self-ascriptions and third-personal self-ascriptions. For in everyday life we often do challenge and correct people’s sincere self-ascriptions, including our own, especially when they concern our emotions, desires, hopes and fears, and even, I think, (some of) our beliefs. On the view that does not see self-ascriptions as assertions, when someone says she is in some mental state, then it’s “inappropriate to subject her to such criticism” (Bar-On 2004, 200). Indeed, Bar-On thinks that if avowals are like natural expressions which likewise merely serve to express the subject’s mental states, then avowals

belong in the wrong “grammatical” category for any epistemic assessment. It is as inappropriate, conceptually speaking, to assess their epistemic credentials as it would be to assess a moan or a cry or a laugh in terms of its evidence, correctness, or reasons. (Bar-On 2004, 231)

Such a view, it seems, has a hard time explaining why avowals sometimes appear to be of precisely the right (grammatical) category for epistemic assessment. The non-assertive view leaves it somewhat of a mystery why, in at least some if not many cases, it does seem appropriate to assess the epistemic credentials of someone’s avowal. By contrast, an assertive approach can make much better sense of what is going on when we are dealing with so-called ‘expressive failures’. If I make an assertion, or indeed a knowledge claim, then this means that all sorts of challenges are in principle legitimate. For when someone makes an assertion, certain questions seem prima facie appropriate, or so at least Austin claims in Other Minds (1946):

When we make an assertion such as “There is a goldfinch in the garden” or “He is angry”, there is a sense in which we imply that we are sure of it or know it ... though what we imply ... is merely that we believe it. On making such an assertion, therefore, we are directly exposed to questions (1) “Do you know there is?” “Do you know he is?” and (2) “How do you know?” If in answer to the first question we reply “Yes”, we may be asked the second question, and even the first question alone is commonly taken as an invitation to state not merely whether but also how we know. (Austin 1961, 77)
If people’s sincere self-ascriptions are assertions, that is, if they involve ‘informing’ or ‘telling’ others ‘how it is with them’, then it seems that we are in principle entitled to ask, in response to someone’s self-ascription, whether she is sure that, or how she knows that, she wants X. Fleming gives a good example:

Suppose someone says “I dislike deep-sea fishing,” and you have good reason to believe that he has never been on the ocean, let alone fished in it; here your natural retort will be “How do you know you dislike deep-sea fishing? You’ve never done any”; and he may have to withdraw his avowal, or say that he meant he disapproves of deep-sea fishing. Nor is this challenge limited to such outrageous cases. If someone has just begun playing squash, and in the middle of a game says to you “I don’t like this game very much,” you may reply “How do you know? You haven’t played enough to find out”; or if he has just arrived in Oxford for his first visit, and avows “I want to spend the rest of my life here,” you may ask him how he knows this, for as yet he has barely seen the town. (Fleming 1955, 624)

There are many more examples in which raising epistemic challenges is not inappropriate or out of place at all, pace the view that likens self-ascriptions to natural expressions.

Notice that saying that because we are entitled to ask questions like “How do you know?” when we are confronted with an assertion does not mean that the self-ascriber is able to answer the question, or indeed to give the right answer. It also does not have the implication that just because self-ascriptions are best conceived as assertions, the question “How do you know?” is always an appropriate one to ask and could not in practice be impolite, rude or even disrespectful. When I say I have a bad headache, then given that my self-ascription is (also) a statement about the state I am in, you are entitled to ask me how I know that I am in pain or whether I am sure that I am. I might well be surprised or even offended by your question, and the assertive view does not need to deny this. In this regard, self-ascriptions of mental states considered as assertions might not be all that different from perceptual reports, such as my saying “The cat is under the plum tree”. In the latter case, I

---

22 I understand Fleming’s example of deep-sea fishing to be an example of a self-ascribed attitude, not a self-ascribed character trait.

23 Snowdon (2012, 260) makes a similar observation.
am also directly exposed, as Austin puts it, to questions like “Do you know that the cat is under the plum tree?” and/or “How do you know?”. In the perceptual case, such questions might likewise take the subject by surprise, but this does not mean that the initial report was not a knowledge claim. Of course, perceptual reports and ascriptions of mental states differ in many ways. For instance, a possible answer in the perceptual case would be “Because I see that there is”, whereas such an answer seems not to be available when I am reporting on an attitude of mine. But the fact that the self-ascriber cannot give an answer to the question does not mean that there is not an answer; at most, it tells us that the answer is difficult, and is perhaps one for which we need to turn to science rather than the subject issuing the self-ascription (cf. Snowdon 2012, 260–61).

All in all, I think the expressivist faces a dilemma and has to make a choice. If self-ascriptions are instances of venting and serve no communicative purpose, then they're not genuinely semantically continuous with other speech acts after all. Alternatively, if avowals are semantically continuous in this way, then our self-ascriptions aren’t just instances of venting but are communicative acts for which one normally has reasons and intentions and for which epistemic challenges are (in principle) appropriate. This would be the ‘holist’ route. The second route forms a horn of the dilemma, however, because expressivists want to claim one does not have reasons or intentions to avow.

Nonetheless, I hope to have been able to suggest that the latter option is to be preferred. In fact, the second option makes room for a more plausible view of what sincere self-ascriptions are. Sincere self-ascriptions of mental states involve ‘informing’ or ‘telling’ others and oneself ‘how it is with one’, and, as such, these self-ascriptions are to be understood as self-interpretations that can be the object of challenge and questioning, including one’s own challenges and questioning, and things that may serve as, and stand in need of, reasons.

Bar-On and other expressivists seem reluctant to construe avowals as bona fide assertions and so are reluctant to move towards holism, because they want, by any means, to avoid the detectivist view according to which our self-ascriptions are the result of recognizing or ‘reporting on’ some first-order mental state. But rejecting the recognitional or detectivist conception of what underlies people’s self-ascriptive expressions is compatible with the claim that our sincere self-ascriptions are assertions. The problem is that the latter, interpretationalist–holist, approach, it seems, is not compatible with expressivism.
7. Conclusion

If we understand expressivism along atomist lines, then, assuming expressivism meets the epistemic condition, discussed in the previous chapter, the idea is that in normal circumstances one can come to know one's mental states by speaking one's mind. In this chapter I have aimed to point out the limits of this moderate atomist version of expressivism. I hope to have shown that both the 'avowal condition' and the 'normal circumstances condition' do not stand up to scrutiny. The overall upshot of this chapter is that a moderate atomist take on expressivism is not very promising, because we cannot appeal to 'normal' or epistemically 'good' circumstance of self-ascription. This is because what counts as a 'good' circumstance of self-knowledge is something that can only be determined by relating a specific self-ascription to the subject whose self-ascription it is, and the life that she leads.

I have suggested that a holist version of expressivism, according to which the question of what makes someone's self-ascribed attitude knowledgeable must be answered relative to that life and not just relative to her expressive act, has been overlooked because expressivists have been reluctant to see sincere self-ascriptions as genuine assertions. Expressivists have tried to maintain a balancing act between seeing avowals as instances of venting that serve no communicative point, on the one hand, and seeing them as truth-evaluable and so semantically continuous with other ascriptions, on the other. This position, I've argued, is ultimately unstable. The reason the expressivist thinks that when one avows being in some mental state, one is not communicating something is because this would (1) secure the neo-expressivist account of why we are usually not prepared to challenge or correct people's avowals, and (2) allow the account to remain genuinely expressivist. However, if avowals really serve no communicative purpose in the end, then we must abandon the claim that avowals are semantically continuous with non-avowals; it also remains a mystery why we speak our minds at all.

Taking the other horn of the dilemma is to be preferred, or so I have tried to suggest. What we should say is that avowals are truth-evaluable and semantically continuous with other ascriptions, which commits us to seeing self-ascriptions of our mental states as similar to self-ascriptions of our hair colour or height and to seeing both as communicative acts through which we are informing someone about as aspect of ourselves. I have suggested that the communicative/assertive take on avowals has a number of benefits.
However, for the expressivist it will be hard to accept, because it requires giving up on the expressivist account of first-person authority (because epistemic challenges are no longer inappropriate if avowals are assertions), and it also seems that the assertoric view is no longer distinctively ‘expressivist’. If avowals are assertions, then it seems we will have to give up on the idea that expression is a success notion whereby subjects show the conditions they express, and must instead consider expression as an ‘evidential’ notion and self-expressions as ways of communicating information.
4

Knowing Your Mind by Making Up Your Mind

1. Introduction

On Richard Moran's (2001) influential account of self-knowledge, the 'capacity to make up your mind', or the capacity to judge, is taken to be fundamental to acquiring self-knowledge. I will refer to this approach as the 'rationalist' account of self-knowledge, because on the sort of view Moran defends, knowing one's attitudes regarding X involves answering the question of whether X is true (or desirable, etc.) in a "deliberative spirit" (Moran 2001, 61), that is, by considering what counts as a reason for or against the truth (desirability, etc.) of X. In this chapter, my primary goal is to explain the rationalist approach to self-knowledge, and, in particular, the so-called 'transparency procedure' that is fundamental to it. Given that this view can be interpreted in different ways, and because it introduces a number of new concepts that need careful consideration, it will be important to be clear about the larger rationalist project before evaluating them more critically in the chapters to come.

I proceed as follows. I begin by describing the more general 'agential' view that I take rationalism to be a sub-species of (Section 2). Next, I turn to rationalism and discuss the so-called 'transparency procedure' (Section 3.1) and the claim that following the transparency procedure delivers immediate or non-inferential self-knowledge (Section 3.2). Next, I briefly discuss the claim that if one does not arrive at one's self-ascription 'transarently', then this leads to alienation (Section 3.3). In Section 4, I discuss existing criticism of the rationalist approach, which I suggest has mostly been concerned with the question of whether following the transparency procedure is necessary for self-knowledge and has been less concerned with the

1 I borrow the label 'rationalism' from Cassam (2014) and Gertler (2011b).
more radical question of whether, even if you do follow the transparency procedure, it provides the subject with attitudinal self-knowledge. In Section 5, I consider a possible way of responding to the existing criticism that has sometimes been hinted at but which has not been systematically considered so far. The strategy is to say that 'transparency' is not an actual method or procedure in the first place, but rather a 'normative demand', which would render the rationalist account immune from most of the existing criticisms. I argue why such a non-epistemic or a 'transcendental' reading of transparency, despite its potential, is not plausible if we take rationalism to be an epistemic theory of alienated and non-alienated types of self-knowledge. In the final section, I turn to the question of whether rationalism is 'a version of' inferentialism or not, and why this question matters to the relevance of the circumstances of deliberation, which in turn is relevant to atomist/holist conceptions of rationalism.

2. The agential account of self-knowledge

In order to understand the rationalist view, it may be helpful to consider rationalism as a version of a broader 'agentalist' approach to self-knowledge (e.g. Tugendhat 1986; McGeer 1996, 2007a; Moran 2001; Hieronymi 2009; Boyle 2009b, 2011a; Bertolotti 2009; O'Brien 2007). On the agential account of self-knowledge, acquiring self-knowledge requires taking up an 'active stance as an agent'. The notion of the 'self as agent' is one that goes back to the writings of Ernst Tugendhat, an early defender of the agential account of self-knowledge:

What does it mean to ask oneself what one wants? Since this is a practical question, the issue is not to ascertain what I in fact want, what my inclinations are, but to determine what position to take toward my inclinations. (Tugendhat 1986, 1996)

Tugendhat criticizes the tradition of thinking that characterizes self-knowledge as analogous to an observer looking at an object (most notably, introspectionism). For Tugendhat, our attitudes or the states we are in are always conceptually tied to the practical question of who we are and who we want to be, and so knowing our attitudes likewise requires considering who we are and who we want to be.
On the agential view, self-knowledge involves a kind of 'maker's knowledge'. Maker's knowledge can be defined as "the special knowledge someone has of something, in virtue of making that thing" (Langton 2009, 289). With respect to self-knowledge, the idea is that the subject does not discover but instead constitutes her attitudes. The relation in which the subject stands to her attitudes is not an evidential but rather a constitutive relation. Just as car designers can have maker's knowledge of next year's models (Langton 2009, 303), subjects can have maker's knowledge of their own attitudes: when they self-ascribe an attitude, they 'make it the case' that their self-ascription is true. As Victoria McGeer writes, the agential account traces the special self-knowledge that subjects have,

not to an agent's capacity for epistemic accuracy in self-ascription, but to a capacity to shape or determine her own states of mind. The agent has a privileged authority in self-ascribing intentional states because it is she who makes it the case that she deserves to be ascribed these states; she has 'maker's knowledge'. (McGeer 2007a, 82)

McGeer writes that our self-ascriptions have an "actively constructive ingredient" and that our self-expressions are used "actively to tell a constantly updated story about ourselves that we also act upon to make true" and concludes that "'knowledge' of our own minds is peculiarly dependent on our role as agents" (McGeer 1996, 514–15). Matthew Boyle, whose view I will describe in detail in Chapter 6, claims, in a similar vein, that subjects have "an ability to know our minds by actively shaping their contents" (Boyle 2009b, 134; for discussion, see Cassam 2014, 113ff)

In this sense, then, agential theories of self-knowledge are similar to so-called 'constitutivist' theories of self-knowledge. According to constitutivists, thinking that you are in mental state M somehow makes it the case, i.e. (partly) constitutes your being in M, and/or vice versa. More specifically, constitutivists think that self-knowledge is "guaranteed to hold a priori, as a matter of conceptual necessity" (Coliva 2009, 369; Burge 1988, 2013; Wright 1989b). However, whereas 'traditional' constitutivists think that this holds as a matter of conceptual necessity, agentialists deny this is so and instead think that self-knowledge requires (at least some) cognitive work. Differently put, constitutivism is what Paul Boghossian (1989) would refer to as a "cognitively insubstantial" theory of self-knowledge, because constitutivists claim that we can have attitudinal self-knowledge without having
to base our second-order beliefs on observation (inner perception) or inference from (inner or outer) evidence. Take the judgement “I am here now”. I can have knowledge of this judgement without having to infer it or direct my inner eye at anything. Constitutivists claim that knowledge of our thoughts and beliefs is cognitively insubstantial in this way.

Agentialists claim that you don’t get self-knowledge just by being conceptually competent. In other words, they think that it is wrong to think that observation on the one hand and making inferences on the other exhaust the options regarding what would make a theory of self-knowledge ‘cognitively substantial’. Thinking that these two options are exhaustive is something that appears to follow only if one accepts a specific metaphysical conception – a realist, fact-of-the-matter conception – of what mental states are and what it means for something to count as ‘cognitive’ or a ‘cognitive achievement’.

So what does the subject need to do? According to McGeer, for instance, the capacity for self-regulation is crucial. She claims, “We know our own minds because we have been trained to take on the responsibility, as only cognitively and linguistically sophisticated agents can, for suiting our words to our deeds and our deeds to our words” (McGeer 1996, 515, see also 2007a; de Bruin, Jongepier, and Strijbos 2015).

Moran (1994, 1997, 1999, 2001, 2012) has a different view, though. What Moran puts in place of ‘observation’ and ‘inference’ is practical reason. More specifically, he claims subjects have the ability to determine or constitute their attitudes by considering the reasons in favour (or contra) the object of the attitude. Briefly, you don’t get self-knowledge via third-personal observation or interpretation of various forms of evidence, nor do you get it for free by merely being conceptually competent. Instead, you get self-knowledge by deliberating about worldly affairs. I explain Moran’s account, and in particular the appeal to practical deliberation, in detail in the subsequent section.
3. The rationalist account of self-knowledge

3.1 The transparency procedure

Central to Moran’s (2001) philosophical account of self-knowledge is the idea that knowing our own mental states, e.g. finding out what we believe or want, is a matter of actively making up our minds by deliberating on their particular subject matter. The special sort of authority that we have when making a self-ascription (believe that P, want Q, fear R) is due to the fact that we have the capacity for practical deliberation, and that exercising this capacity can make a metaphysical difference with regard to one’s attitude, that is, a difference to what one’s attitude is.

Moran sets out to explain the special nature of first-person authority in the wider context of (moral) human agency. He criticizes accounts of self-knowledge that seek to explain first-person authority exclusively in epistemic terms (2001, 1). Such ‘epistemic’ accounts typically model first-person knowledge on perception or theory-based observation (Armstrong 1968; Gopnik 1993; Nichols and Stich 2003; Carruthers 2011). In cases of perception or theory-based observation, the perceiver or observer typically has a passive role in relation to the object perceived or observed, in the sense that the object leads an existence independently of the epistemic subject. Moran, by contrast, argues that such passivity is alien to our conception of ourselves as self-knowing agents. To use one of Moran’s examples (2001, 26), imagine asking someone whether she intends to pay back the money she borrowed. Suppose she answers, “As far as I can tell, yes”. What makes this response particularly disturbing is that it appears to be issued from an onlooker’s perspective, as if she were talking about someone else. We generally do not accept such answers, precisely because they signal a lack of first-person involvement. We demand that others play an active part in coming to know their own mental states; we demand that they make up their mind, i.e. decide whether they will pay back the money, perhaps after deliberating about the matter and endorsing the intention to (not) do so.

Moran thus criticizes purely ‘epistemic’ accounts of self-knowledge for modelling the self-relation as an exclusively third-person or ‘theoretical’ stance towards one’s own mental states. Thus, he states,

---

2 Some of the material in this section and the next is adapted from Strijbos and Jongepier (forthcoming).
What is left out of the Spectator's view is the fact that I not only have a special access to someone's mental life, but that it is mine, expressive of my relation to the world, subject to my evaluation, correction, doubts and tensions. This will mean that it is to be expected that a person's own awareness of his mental life will make for differences in the constitution of that mental life, differences that do not obtain with respect to one's awareness of other things or other people. For this reason, introspection is not to be thought of as a kind of light cast on a realm of inner objects, leaving them unaltered. (Moran 2001, 37 emphasis in original)

Moran often uses the phrase that someone's attitudes are in an important sense "up to him". By this, Moran does not mean that I have direct control over my attitudes - I cannot believe that pink elephants exist, for instance, by an act of will. What he means is that my attitude regarding P arranges itself to the conclusion of my exercises of practical reasoning. In other words, what my attitudes are is not independent of my reasons. They are up to me in so far as my reasons are up to me.  

According to what Moran, following Gareth Evans, calls the 'transparency condition' of first-person statements of, for example, one's beliefs, one should treat the question of one's belief about P as equivalent to the question about the truth of P. This means that one defers answering the self-directed question "Do I believe that P?" by answering the world-directed question "Is it the case that P?". This is Evans' formulation:

> 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)

It's not that it would be wise or good for me to attend to certain outward phenomena, but rather, as Johannes Roessler points out, that "the 'must' here is naturally interpreted as a 'must' of practical necessity" (J. Roessler 2013, 8). This ties in with Moran's interpretation of transparency, according to whom the world-directed question is a deliberative question:

---

3 As we'll see, this formulation reveals a potential problem for the rationalist account, for to what extent are my reasons really up to me? I will discuss these issues in Chapter 7.
I address myself to the question of my state of mind in a deliberative spirit, deciding and declaring myself on the matter, and not confront the question as a purely psychological one about the beliefs of someone who happens also to be me. (Moran 2001, 63)⁴

That one question is ‘transparent’ to another is meant to bring out the fact that we can answer one question in terms of another. Moran stresses that transparency should be understood as a “stance taken toward oneself” and is not just “a logical feature” of self-ascriptions (Moran 2001, 66–67).

The answer to a deliberative question should thus be understood as “the formation of a desire or a belief” (2007, 56 emphasis in original), that is, the outcome of taking up a deliberative perspective regarding P is “a form of conviction and hence brings me to a new state of mind” (Moran 2007, 73). Moran explicitly mentions that his account is meant to give us knowledge not just of our beliefs but of “various standing attitudes of the person, such as beliefs, emotional attitudes, and intentions” (e.g. Moran 2001, 9; see also 43ff; 52; 63; 78ff, 2002). For instance, he gives the example of someone wanting to “change jobs, or learn French, or avoid being seen”, all of which can be understood to be ‘judgement-sensitive’ in that they “depend on certain beliefs about what makes these various things desirable” (Moran 2001, 115). Regarding these other judgement-sensitive attitudes, we must simply reformulate the relevant word-directed question. For instance, the idea would be that in order to know whether I have a desire to change jobs, I could ask whether ‘changing jobs is desirable’. When I want to know whether I want to learn French, I answer the question of whether ‘studying French is a good idea’, etc. (I turn to (worries regarding) the scope that the transparency procedure is meant to apply to below.)

⁴ There are other interpretations of transparency, such as Robert Gordon’s (1995, 2007) so-called ‘ascent-routine’ account. On Gordon’s view, the passage from Evans quoted earlier takes central stage as well: “In answering a question concerning our belief about some matter, what we generally do is to step down a semantic level and answer the corresponding question about the matter the belief is about. To answer the question concerning my belief about the weather, I don’t in general search for a telltale feeling or other experiential mark of belief. Nor do I examine my recent behavior in the light of a theory. I simply recast the question as, ‘Is it raining outside?’” (Gordon 2007, 153). One of the differences between Gordon’s and Moran’s accounts is that for Gordon, the “question-answer form is not essential to the notion of an ascent routine” (Gordon 2007, 154). Gordon’s main point is not that in answering a question about one’s belief about something, one can step down a semantic level. Rather, his main point is that “whether in answer to a question or not, people optionally step up a semantic level from an assertion that p to a self-ascription of a belief that p” (Gordon 2007, 154 emphases added). Second, the answer for Gordon doesn’t have to be rational, as in Moran’s case. The answers we give to world-directed questions are not made on the basis of rational deliberation but simply on the basis of training (Gordon 2007, 158ff; cf. Michael 2010).
I take ‘transparency’ not (just) to be a transcendental claim about, for example, the nature of human (rational) agency or what it is to be a believer but as an actual way of acquiring attitudinal self-knowledge (though I consider the ‘transcendental reading’ in more detail below). It is, I believe, no accident that Moran talks about transparency as a “route to knowledge” of one’s attitudes (e.g. 2001, 68). Hence, I will refer to the activity of making up one’s mind as ‘following the transparency procedure’. I should stress, though, given that the notion of a ‘procedure’ may not be the most suitable in this context, that I understand the notion of procedure (or method) in a broad sense, as referring simply to a ‘way’ of arriving at a self-ascription, or coming to believe that one has a certain belief, hope, intention or desire.5

As I read Moran, making up your mind is equivalent to coming to a judgement, and this judgement constitutes, determines or settles your attitude. Therefore, in what follows, I take ‘following the transparency procedure’ to be equivalent to (1) ‘making up one’s mind’, (2) ‘answering the world-directed question’ and (3) ‘coming to a judgement by reflecting on the reasons in favour of the subject matter’.6 The central idea, then, is not just that we have the capacity to answer world-directed questions but also a metaphysical-epistemic one: reflecting on the reasons in favour of P issues in a judgement, and that judgement is constitutive of one’s belief, one thereby knows about. Following the transparency procedure involves, first, asking oneself whether one has a certain attitude (“Do I believe that P?” or “Do I want Q?”). Second, one tries to answer the corresponding world-directed question (“Is P true?” or “Is Q desirable?”). The second step involves deliberating over the subject matter, i.e. considering the reasons for and against. Third, one comes to a conclusion regarding P, in other words, the deliberative process ends in a judgement (e.g. whether there will be a third world war). This judgement (a mental act or event) constitutes one’s attitude (a mental state).

At this point, though, nothing much has been said about self-knowledge. As Alex Byrne wonders, “Suppose that I examine the evidence and conclude that there will be a third world war. Now what?” (Byrne 2011a, 203). For the transparency procedure to have anything to do with self-knowledge, we need to add an epistemic condition such that the subject also knows about the attitudes that are so constituted.

---

5 I return to the question of what the status is of ‘transparency’ (a procedure, a condition, a constraint?) in the final chapter.
6 Surprisingly, the notion of ‘judgement’ does not play much of a role in Moran’s work. But I see no other way of understanding the ‘outcome’ of answering world-directed questions, i.e. one’s conclusion (for a judgement-based take on transparency, see e.g. Boyle 2011a).
How exactly can one be said to know one's attitudes when they are constituted in a transparent way? The thought is that if one judges that P, one knows that one judges that P, i.e. one judges 'knowingly'. If one judges that P, then that judgement provides an epistemic basis, or grounds, for one's self-ascription (second-order belief), which explains the belief's status as knowledge. This interpretation of why making up one's mind provides one with self-knowledge invites an inferentialist explanation, though, namely that what is going on is that one infers that one believes that P from the fact that one has judged that P (Byrne 2011a; Cassam 2011, 2014).

Such an 'inferentialist' construal of the transparency procedure, however, would be problematic, because one of Moran's key claims is that transparent self-knowledge is 'immediate'. Moran claims that one's self-ascriptions are “not based on the observation of the person's sayings and doings” (2001, xii) and that “the basic concept of first-person awareness that we are trying to capture is that of awareness that is not based on evidence, behavioral or otherwise” (2001, 11 emphases added). The idea is that one's knowledge or second-order belief regarding one's first-order state (one's belief that P, say) has no 'grounds', or is not 'based' on anything. Indeed, according to Moran, if someone had to rely on (behavioural) evidence to report on his mental states, then this "would suggest something wrong with him, some state of dissociation, and would raise doubt, about the rationality of those attitudes of his which are not accessible to him in the normal 'immediate' way", i.e. in "a way that does not rely on inferences from anything inner or anything outer" (Moran 2001, 68).

3.2 Alienation
Moran recognizes that sometimes what my attitudes regarding some object are and what my reasons regarding that object are can come apart. In other words, sometimes one cannot come to know one's own attitudes via 'making up one's mind' or following the transparency procedure. Central to Moran's account is that if one cannot answer the question of what one believes by directly reflecting on the content of that belief, then one is alienated from that attitude.

In this context, Moran discusses an example of someone who learns, by going through psychotherapy, that she feels betrayed but who is unable to endorse this description of herself and is only able to attribute this state to herself from a third-person perspective rather than avow it first-personally. Moran writes,
The person might be told of her feeling of betrayal, and she may not doubt this. But without her capacity to endorse or withhold endorsement from that attitude, and without the exercise of that capacity making a difference to what she feels, this information may as well be about some other person, or about the voices in her head. From within a purely attributional awareness of herself, she is no more in a position to speak for her feelings than she was before, for she admits no authority over them. (Moran 2001, 93)

Her feeling betrayed is, Moran says, “detached from her sense of the reasons”. If one is unable to learn of one’s attitude by reflection on the content or object of that attitude, and can only take a ‘theoretical’ or ‘empirical’ stance on oneself, then this leads to alienation. The fact that Moran explicitly remarks that the theoretical or alienated route to self-knowledge is also a route to self-knowledge means that transparent avowal and opaque attribution should be, as he says, “seen as different routes to knowledge of the same thing” (Moran 2001, xviii; 89). Not all self-knowledge is transparently acquired, and even knowledge of one’s attitudes can be acquired in a non-transparent way.

What the language of alienation is meant to brings out is that despite the fact that alienated self-knowledge is also self-knowledge, it’s not the ‘right’ or ‘proper’ way of gaining self-knowledge. This means that we cannot have a proper understanding of Moran’s project if we ignore the fact that his project is irreducibly normative. This is evident from Moran’s claims that there is something “wrong” (Moran 2001, 68) with the person who can only rely on behavioural evidence to report on his or her mental states. It is only by putting to use our capacity for practical reason, if we avow our own attitudes, that they amount to ‘genuine’ self-knowledge where the latter means a type of self-knowledge that is non-alienated. 7

7 I will return to the relevance of the notions of alienation and ownership in Moran’s account and their connection to the notions of authenticity and autonomy in Chapter 7.
4. Is the transparency procedure necessary and sufficient for self-knowledge?

4.1 Common objections to rationalism

Moran himself is mostly concerned with showing that making up your mind is necessary for (genuine) self-knowledge, which is illustrated in the fact that he is arguing specifically against introspectionist views or other 'spectatorial' approaches which construe self-knowledge in third-personal terms and which overlook the first-person perspective. The introspectionist view, he says, “underdescribes the differences between self-knowledge and the knowledge of others” (Moran 2001, 37). Moran's main claim regarding transparency thus seems to be that transparency is necessary for (non-aliened) self-knowledge, even if it isn't necessary for every instance of acquiring self-knowledge, and that any account that fails to mention the perspective of the judging first person only tells half of the story at best. In this section, I want to review, briefly, some of the common objections and worries about the rationalist account of self-knowledge.

Much of the criticism of the rationalist approach takes issue with the claim that following the transparency procedure is indeed necessary for self-knowledge (the 'non-aliened' clause is mostly forgotten about somewhere along the way). Most critics have been concerned with pointing out that having to deliberate over the reasons in favour of P as a way of finding out what one's attitudes regarding P are is a highly demanding and overly intellectualistic requirement for self-knowledge, and that we can often do without it (McGeer 2007a; Shah and Velleman 2005; Shoemaker 2003; Heal 2004; Cassam 2014). This is clear, for instance, when it comes to beliefs we already have, i.e. regarding things we don't have to make up our minds about. Sydney Shoemaker notes that, often, “one would be at a loss to say what one's reasons are for believing what one does” (Shoemaker 2003, 395). This is true, he says, for instance, of one's perception-based knowledge of one's current surroundings, and one's memory knowledge of one's recent past. And it is true of one's knowledge of, and belief in, the vast number of historical and scientific facts that one has learned and retained in memory. (Shoemaker 2003, 395)
In these cases, following the transparency procedure becomes a rather laborious, indeed superfluous, task. As Shoemaker rightly points out, even though I know that I am wearing trousers, and thus believe that I am,

it is hard to think of circumstances, other than those of a dream, in which it could be a question for me whether I believe this. I would also have a hard time saying what reasons I have for believing it. And I cannot think of any good sense in which it is “up to me” whether I believe it. (Shoemaker 2003, 396).

Or consider David Finkelstein’s argument:

I dislike goat cheese intensely. But I don’t believe goat cheese to be bad or undesirable or something that one ought not to like. I understand that fine people of excellent taste love the stuff. ... When I tell you that I dislike goat cheese ... I am speaking with first-person authority even though I am not, in Moran’s sense, avowing anything. (Finkelstein 2008, 163;165)

Cassam points out that answering the world-directed question regarding the content of my attitude is often more difficult than simply answering the mind-directed question about what my attitudes are (Cassam 2014, 104). The question of whether the reasons require me to want carrot cake or intend that I buy apples is much harder than simply answering the question of whether I do in fact have the desire for carrot cake or the intention to buy apples.8 In order for me to know that I want a slice of carrot cake, it does not seem necessary to answer the world-directed question (e.g. “Is carrot cake desirable?” or “Are there convincing reasons to have a slice of carrot cake?”).

The critique that rationalism is too intellectualistic and demanding leaves intact the idea that if one were to follow the transparency procedure, it would provide attitudinal self-knowledge. It merely claims that it’s an odd way of acquiring self-knowledge. We might call this the ‘soft’ critique on rationalism. While I take it to be clear that Moran’s goal is to show that transparency is fundamental to (non-alienated) self-knowledge, I also think it’s fair to say that a natural reading of his

---

8 Notice that this (‘whether the reasons require me’-condition) is a specific interpretation of the transparency procedure (Cassam follows Finkelstein 2012, 104). I address this below.
view is one that claims that if you do make up your mind, what you end up with is attitudinal self-knowledge. In other words, one reading would be to think that there is nothing else you need to do in order to know your mind other than answer the world-directed question. Of course, there will be certain minimal conditions that the subject must meet, such as being minimally conceptually competent, having the concept of belief and being able to deliberate at all. But apart from these minimal and relatively trivial conditions, it appears plausible to understand rationalism as saying that for someone to have self-knowledge, all she has to do is follow the transparency procedure. The more radical critique would be to question whether answering the world-directed questions actually delivers knowledge of her attitudes.

The radical objection to rationalism would then involve showing that even if you do follow the transparency procedure, this does not necessarily give you knowledge of your attitudes. The radical objection has been made by those who have argued that it is not possible to acquire knowledge of some (or many) types of mental states by answering world-directed questions. Cassam, for instance, takes issue with the possible scope of rationalism, and argues that in some or even many cases, the world-directed question is also the wrong question to ask in order to acquire self-knowledge:

The answer to the question “Do the reasons require me to adore my dog?” is plainly no even though the answer to the question “Do I adore my dog?” is plainly yes. In trying to answer the latter question by answering the former question I would be barking up the wrong tree. (Cassam 2014, 105)

Our emotions, in particular, are often unresponsive to reasons. Consider the case of phobias. Moran brings them up as an example of what he too considers a familiar fact about some emotional states: they can be irrational and “do not alter when

---

9 My concerns here are not principally exegetical. I think this reading is plausible even if, in the end, the sufficiency claim is not one Moran actually makes, or would be willing to make.

10 Cf. “If your doctor has told you to cut down on your drinking then you have a good reason not to want a gin and tonic, but that doesn’t alter the fact that you want one. So consideration of what you ought rationally to want won’t be a good guide to what you actually want... If your desire for a gin and tonic is impervious to the reasons for not wanting one then reflecting on those reasons won’t tell you what you want you ought not to want one but you do. Other attitudes are no different. You shouldn’t fear the spider in your bathroom but you do, and it would be remarkable if every one of your beliefs is one that you ought rationally to have” (Cassam 2015b).
the beliefs on which they are based are sincerely denied by the person in question” (Moran 2001, 54). But it’s hard to see why someone would not know that she is afraid even if her reasons don’t determine what her attitude is. As Taylor Carman (Carman 2003, 404) claims, we often express ourselves in various ways, not “to decide how I ought to feel, but to get clear about how I do feel by letting my emotions take shape and find a voice in what I say and do”.

There is another, more serious concern about Moran’s exclusive focus on transparent deliberation in accounting for self-knowledge. This worry does not concern the deliberative or rationalistic aspect of Moran’s transparency condition for having non-alienated self-knowledge, but rather the requirement of transparency itself.

4.2 Possible responses
There are different ways in which the rationalist could respond to the more radical objections. One response would be to backtrack somewhat and say that it was never the rationalist’s ambition to explain the ‘full scope’ of self-knowledge anyway (cf. Boyle 2009b), and hence we should not be surprised that the transparency procedure cannot handle all our attitudes in a satisfactory way. But as we’ve seen, Moran and other rationalists mention explicitly that the transparency procedure is meant to explain the sort of knowledge we have of other attitudes, including our “emotional attitudes, and intentions” (Moran 2001, 9), and indeed the suggestion seems to be that for these attitudes, making up your mind is (normally) sufficient for acquiring knowledge of them. So accepting that the scope of rationalism in the end only applies to our beliefs, indeed only applies to the sort of beliefs that we don’t already have and that don’t require making up our minds, is quite a concession.

A response that I believe has more potential is to take issue with the way in which the transparency procedure is understood by (some) critics. Cassam and Finkelstein, for instance, define transparency, and criticize rationalism on the basis of, the following definition:

The question of whether I believe that p is, for me, transparent to the question of what I ought rationally to believe—i.e., to the question of whether the reasons require me to believe that p. I can answer the former question by answering the latter. (Finkelstein 2012, 103)
If this is how we should understand the transparency procedure, then it's no surprise that rationalism gives the 'wrong result' when applied to our feelings, desires and emotions, because we can want and feel things that we ourselves may think we "ought rationally" not to want or feel.

The solution, then, is to reformulate the world-directed question that the mind-directed question (what I believe/want/feel/care about, etc.) is transparent to. The rationalist could suggest that instead of saying that the mind-directed question (e.g. "Do I want X?") is transparent to the question of whether I ought rationally to want X, we should instead say that it is transparent to the world-directed question, e.g. "Is X desirable/lovable/good/worthy/beautiful?". In other words, the transparency procedure gives the wrong result in a (large) number of cases only if we operate with the wrong world-directed question; what we need to do is formulate the right world-directed questions for different kinds of attitudes.

But reformulating the world-directed question seems problematic. For one thing, the less rationalistic formulation (which is not formulated in terms of what one ought to think/feel) seems to face similar problems, because one might concede that something is desirable/good/beautiful without having the relevant desire, and vice versa. I take this to be Finkelstein's point: he does not believe that goat cheese is bad or undesirable, and yet he doesn't want to try any. So the same problem of answering world-directed questions as a way of coming to have knowledge of your attitudes remains.

What the above examples show is that focusing on the reasons in favour of P might give an answer to the question of what people in general want, or should want, but it does not seem to be a good way of finding out what it is that I want (hope for, feel, etc.). Finkelstein could, and does, recognize that goat cheese is not undesirable 'as such'. Rather, it's just that he does not have the relevant desire. Likewise, when I answer the world-directed question of whether, say, pursuing a career in academic philosophy is a good idea, I will conclude, on the basis of the reasons that are available (the publication rat race, the growing pressure to produce societally relevant research, the endless temporary positions, and so on), that it isn't. But this does not answer the question of what I want, or what I believe is the right course of action. As Naomi Kloosterboer argues, Moran's idea that rational agency consists in deliberating about the truth of some subject matter "neglects important aspects of our practical agency, in particular that we are agents to whom things matter, with certain projects, relationships, vulnerabilities and peculiarities" (Kloosterboer 2015, 265). What we would minimally need, then, is a 'self-referential'
reformulation of the world-directed question. I will address (the problems of) the 'self-referential' take on transparency and the general problem with reformulating the relevant world-directed question in more detail in the next chapter.

For now, I want to observe that there might be a more radical response available to the rationalist, which would at the same time take issue with the 'soft' objection regarding the intellectualistic flavour of the transparency procedure. The more radical response is to say that transparency is not a procedure in the first place and that all of the existing criticism is premised on the assumption that it is. I address this response below.

5. A transcendental take on transparency?

Are Moran's critics taking 'transparency' too literally, as an actual psychological procedure when this is not how transparency should be understood? Some of Moran's own remarks seem to plead for a different conception of transparency:

For a desire to belong to the "judgment-sensitive" category it is, of course, not necessary that it formed as the result of deliberation. For very few of our desires come into existence as the conclusion of an explicit exercise of practical reasoning. Equally, however, very few of our beliefs about the world arrive as the conclusion of any explicit theoretical reasoning that we undertake. It is nonetheless essential to the category of belief that a belief is a possible conclusion of some theoretical reasoning. (Moran 2001, 116 emphases in original)

Moran explicitly claims that it is not as if "one normally arrives at one's beliefs (let alone one's fears or regrets) through some explicit process of deliberation", and instead suggests that there must be "logical room for such a deliberative question" (Moran 2001, 63 emphases added). In more recent work, in replying to his critics, Moran stresses that his claim that the concepts of "rationality, responsibility, and agency" that he takes to be fundamental to self-knowledge are sometimes "understood to mean something sweeping in the exaggerated claims for the actual governance of rationality or responsibility in our lives", and clarifies that his goal is to try and be "as minimalist as possible in commitments about how far our empirical lives are governed by, or explainable by, the normative structures of these concepts and other ones" (Moran 2004a, 455 emphases in original). He describes his project as follows:
In the book I emphasize that the role I give to the deliberative stance is not meant to suggest that most of our beliefs are actually formed through explicit deliberation or reasoning. We'd end up with many fewer beliefs for coping with the world than we actually have if we could only acquire them through explicit reasoning or deliberation. Here again, I'm trying to keep to a claim, and show that it has surprising consequences for understanding self-knowledge. The modest claim is that, while most of our beliefs and other attitudes either never arrive at consciousness at all, or only do so from we know not where, the fact remains that it is possible for a person to draw a conclusion, reach a finding, determine his belief about something on the basis of his assessment of the reasons supporting it. Put this way, I take this to mean something the denial of which would be equivalent to denying that people ever actually reason to a conclusion, or act or hold beliefs or other attitudes for reasons. (Moran 2004a, 458 emphases added)

There are at least two different ways of understanding the 'modest claim': an implicit one and a transcendental one. An 'implicit' version of the modest claim comes down to saying that, even though you do not arrive at your beliefs through explicit deliberation or reasoning, you arrive at them through implicit deliberation and reasoning. And even though answering the question of what your attitude is by answering a corresponding question about the world is not something you do consciously, it can be said to be something you do implicitly. But I don't think this way of construing rationalism is very plausible (exegetically and otherwise). Does it mean that every time one 'avows' being in some mental state, one is implicitly following the transparency procedure? This seems hard to believe, and does not seem to be compatible with the overall rationalist approach, either, which makes no mention of unconscious or implicit processes of deliberation. Instead, it's more plausible to think that when Moran denies (and in this respect agrees with his critics) that we hardly ever follow the transparency procedure 'explicitly', this doesn't mean that we do so implicitly but rather that following the transparency procedure is something that it is possible for us to do. 11

---

11 There's an analogous discussion regarding a well-known passage of Christine Korsgaard, which runs as follows: "[O]ur capacity to turn our attention on to our own mental activities is also a capacity to distance ourselves from them, and to call them into question. I perceive, and I find myself with a powerful impulse to believe. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse doesn't dominate me and now I have a problem. Shall I believe? Is this perception really a reason to believe? I desire and I find myself with a powerful impulse to act. But I back up and bring that impulse into view and then I have a certain distance. Now the
On this understanding of the modest claim, answering the question of what we believe, intend, desire, and so on by answering a corresponding question about the world is not meant to be an accurate psychological or empirical description about how we acquire self-knowledge; rather, we might take it to be a description of what’s essential to human agency. More specifically, we might understand the capacity of being able to follow the transparency procedure as a transcendental condition of the very having of mental states, and this is how we should understand the sense in which transparency is supposed to be ‘fundamental’.

This reading of the project seems to be suggested by Boyle, who recognizes that Moran could not have taken ‘fundamental’ in an empirical or psychological sense, but adds, “My own view is that there is a sense of ‘fundamental’ on which Moran’s claim is true” (Boyle 2009a, 140). He explains that transparent self-knowledge is fundamental, because without it, one could not understand oneself first-personally at all (Boyle 2009b, 155). Transparency understood in this way is something that is fundamental to being “a thinker and an agent” (Boyle 2009b, 161). And so we might understand the modest claim regarding transparency to be a transcendental claim in which the capacity of being able to follow the transparency procedure is a requirement for being a believer at all.12

A transcendental take on transparency has a number of advantages. First, this reading would make sense of the fact that Moran never talks about transparency as a method or procedure (as I and others have done). He never talks about transparency in this way in Authority and Estrangement, nor in more recent work. Instead, he talks about the “Transparency Condition”. Whereas a procedure or method is something one can or should follow, which has cognitive and/or epistemic connotations, it may be argued that a condition is something to be obeyed— it is rather more like a command or something one ought to do. Conforming to transparency, Moran says, is a “normative demand” or “a normative requirement on rational agency” (Moran

---

12 A transcendental take on transparency would see Moran’s project as one that is in line with other Kantian projects, such as those that are meant to offer a ‘transcendental psychology’ (Boyleveld and Ziche 2015; see also Düring and Düwell 2015).
Knowing Your Mind by Making Up Your Mind

2001, xvi–xvii). Second, the transcendental reading of transparency would render irrelevant the criticism that the transparency procedure would be incomplete, inefficient or unreliable.

However, the transcendental reading is also problematic, for a number of reasons. First, Moran himself stresses, as we've seen, that his account of transparency is more than 'merely logical'; he claims that one might "lose the right" to assume that one's judgements constitute one's attitudes (Moran 2003, 406) and thus that rational deliberation can fail to settle one's beliefs. These claims seem to go beyond a transcendental reading of transparency. I thus agree with Johannes Roessler, who says that "Moran is sometimes interpreted as offering a 'non-epistemic' explanation of self-knowledge ... This, it seems to me, is misleading. The stated aim of Moran's theory is to 'explain and vindicate avowal as a privileged form of knowledge of oneself' (Moran 2001, 134)" (J. Roessler 2013, 7).

Second, a transcendental take on transparency is not just modest; it might simply be too weak if, that is, it's meant to explain how we acquire self-knowledge, i.e. that it is a way of coming to know one's own attitudes. Recall that all we need is for there to be 'logical room' to answer a world-directed question. As Taylor Carman notes, "[I]t is important to recognize how weak that claim is" (Carman 2003, 404). It tells us very little in terms of how we acquire self-knowledge. The claim that human beings have the 'capacity' to make up their minds merely means that it should be possible for anyone to answer the question of what she believes by answering the question of what is the case. A transcendental reading of transparency is compatible with the fact that we don't 'follow' transparency regularly – indeed, it's compatible with the idea that we don't follow it at all. What's required is that it should be possible for someone to do so. But how could such a weak claim, psychologically speaking, offer an explanation of how we know our own minds, which is what it was meant to offer?

Third, the transcendental reading of transparency is incompatible with some of the fundamental claims of Moran's account, namely his claim that 'transparency' matters in relation to (avoiding) alienation and that it matters to the mental health of the person – a claim Moran makes at various points in his book (Moran 2001, 8, 35, 60, 107–8, 136–37). Moran claims that "clearly something is wrong if [someone] cannot consciously avow the first attitude and can only ascribe it to himself on the evidence" (Moran 2001, 86). Rationalism is meant to make sense of specific instances of adopting the deliberative stance and 'obeying transparency'. This is evident, in particular, in Moran's discussion of the 'analysand', the person who, through the
course of therapy, learns that she believes/feels X. She does so on the basis of evidence, namely because of what her therapist tells her. Unless she is able to avow her mental states, i.e. conform to the transparency condition, she will remain alienated with respect to her own mental states. But if transparency is all about having a particular capacity, then it’s not clear what would be ‘wrong’ with people who have this capacity but who don’t exercise it. On a transcendental take on transparency, there would be nothing wrong with those who rarely follow the transparency procedure or with those who don’t make up their minds in certain scenarios – there’s only something wrong with those who are altogether incapable of doing so. It is one thing to claim that human beings have the capacity to ‘conform to the transparency condition’, and another to say that this is what would be ‘normal’, ‘better’ or ‘healthier’ for us to do, which seems to be a stronger claim that Moran wants to make. For these reasons, I think transparency should not be understood (only) transcendently.

All of this should make clear that the transcendental take on transparency is perhaps plausible as a theory of what’s essential to human beings, but taken as something that’s relevant to the question of what explains (non-alienated) self-knowledge, it cannot be quite right. The question for the rationalist is whether there is room in between (1) the modest claim that it should be possible for someone to answer the mind-directed question in terms of a world-directed one and (2) the claim that transparency is a genuine procedure of sorts that has its place in the actual (psychological) lives of persons and is a capacity they ought to exercise to avoid alienation. What rationalists want, it seems, is a middle position between (1) and (2), but it’s hard to see what such a middle position would be. Despite the fact that I think the transcendental or capacity-reading of transparency is plausible—indeed, ironically, I believe it may even be more plausible than any psychological rendering of it—I don’t think it’s compatible with the rationalist’s key ambitions with respect to self-knowledge.

In what follows, I therefore assume that we should take transparency to be a procedure. Understood in this way, it seems critics are right to take transparency to be a procedure. And, further, it seems they are right to argue that it’s not a procedure we often use, and also that when we do, it doesn’t necessarily provide us with knowledge of our attitudes and indeed might lead to self-deception.

In the next section, I want to return to the question bracketed before, namely whether the transparency procedure is an inferential procedure.

---

13 For discussion of the ‘normative’ reading of transparency, see Cassam (2014, chap. 7).
6. Is the transparency procedure an inferential procedure?

As mentioned earlier, a natural way of explaining why following the transparency procedure provides self-knowledge at all is to understand rationalism along inferentialist lines. In order for me to gain knowledge of my belief that P, I need more than my judgement that P. The reason for this is that I also need to assume that my judgement (necessarily, normally, typically) constitutes my beliefs. Moran recognizes the fact that in order for the transparency procedure to provide self-knowledge, some such assumption is needed:

What right have I to think that my reflection on the reasons in favor of P (which is one subject-matter) has anything to do with the question of what my actual belief about P is (which is quite a different subject matter)? Without a reply to this challenge, I don’t have any right to answer the question that asks what my belief is by reflection on the reasons in favor of an answer concerning the state of the weather. And then my thought at this point is: I would have a right to assume that my reflection on the reasons in favor of rain provided an answer to the question of what my belief about rain is, if I could assume that what my belief here is was something determined by the conclusion of my reflection on those reasons. An assumption of this sort would provide just the right link between the two questions. And now, let’s ask, don’t I make just this assumption, whenever I’m in the process of thinking my way to a conclusion about some subject-matter? (Moran 2003, 405)

The transparency procedure thus only works “insofar the person is entitled to the assumption that, e.g., what he believes about something on reflection is determined by what he has reason to believe” (Moran 2004a, 457). So for one to be justified in believing that P on the basis of judging that P, one has to make the assumption that one’s belief is indeed determined by the conclusion of one’s reflection on the reasons in favor of P.

The assumption that judgements and beliefs are linked in the appropriate way seems to make the rationalist account inferential (Cassam 2014; Byrne 2011a). In this context, it is helpful to first address the prior question of what would make knowledge in general inferential. Following Jim Pryor (2005), Cassam proposes to distinguish between psychological and epistemic immediacy. When we want to know whether someone’s belief (including one’s second-order belief that one has a certain attitude) is ‘immediate’, then
On an epistemological reading the issue is what kind of epistemic support you have for P, and not whether you arrived at P by inferring it from other propositions you believe. Whether you infer that P in the latter sense is only relevant to whether your knowledge or justification is inferential in a psychological sense. For your knowledge that P to count as non-inferential in an epistemological sense it must be the case that your justification for believing P does not come from your justification for believing other propositions. (Cassam 2015a)

On the topic of what it means for knowledge to be ‘non-inferential’, John McDowell similarly writes,

Surely the distinction should be epistemologically, not just psychologically, significant. It should concern the character of a state’s credentials, not the process by which it was arrived at. (McDowell 2010)

So, then, when we are talking about the causes of one’s belief, how one arrives at one’s belief or how one’s belief is formed, as well as talking about (temporal) structures, we are considering the psychological features (histories, pathways) of a person’s state. If, by contrast, we are concerned with what makes it appropriate for one to believe that P, or what one’s grounds or bases are for believing what one does, then we are in the epistemic domain.

Returning to the relation between judgement and belief, we should now be able to see why Moran’s account appears to be inferential. For in order to know what one’s

14 McDowell makes this remark in the context of criticizing the way Robert Brandom distinguishes inferential from non-inferential knowledge. Brandom considers a physicist who begins by having to infer claims about certain unobservable sub-atomic particles (mu-mesons) from premises about observable goings-on, and later learns to “respond to cloud-chamber phenomena with claims about mu-mesons directly, without needing to take an inferential step”. McDowell goes on: “As Brandom uses the idea, she has now acquired a capacity for noninferential knowledge about mu-mesons. But is that the right way to distinguish inferential and noninferential knowledge?”, which is followed by the passage just cited. What’s noteworthy is that, apparently, McDowell takes the physicist’s way of responding (as requiring no explicit making of inferences) as sufficient for psychological immediacy. This suggests that McDowell equates psychological inferences with something like conscious inferences, and that for something to be psychological immediate, it will be enough for the person not to have to draw any explicit inferences. The alternative is to see ‘psychological’ inferences as cognitive inferences, whether or not they are conscious. In which case, the physicist’s knowledge might still be inferential. If a cognitive construal of psychological immediacy is to be preferred, then it seems more accurate to say that McDowell is discussing what we might call phenomenological immediacy rather than psychological immediacy.

15 For a similar distinction, see, for example, Roche (2016).
beliefs are on the basis of what one judges, one has to assume, as Moran recognizes, that judgements and beliefs are linked in the appropriate way. This is not a trivial assumption, nor does it seem plausible to think it's a transcendental assumption, because Moran allows that one can “lose the right” to assume that one's judgements constitute one's beliefs when he says that

even in our more self-constituting moments, we know ourselves to be rationally frail creatures, and in various circumstances we can lose the right asserted in the Transparency condition (just as the Kantian Transcendental Assumption of Autonomy does not involve the denial that there can be what he calls “alien influences” on the will, or that their hold on the person can fatally compromise the claim to autonomy). (Moran 2003, 406 first emphases added)

For someone's knowledge of the belief that \( P \) to count as immediate, it must be the case that one's justification for believing that \( P \) does not come from one's justification for believing other propositions. But, as it turns out, one's knowledge does come at least in part from believing other propositions, namely from one's belief that judgements and beliefs are related in the right way, or what Cassam refers to as the ‘linking assumption’:

Assuming that my belief concerning \( P \) is determined by the conclusion of my reflection on the reasons in favour of \( P \) is therefore equivalent to assuming that my belief concerning \( P \) is determined by whether I judge that \( P \). Call this the linking assumption (LA). Moran represents the linking assumption as one which I actually make, and am entitled to make, when I am in the process of thinking my way to some conclusion. The specific role of this assumption is to connect the inward-directed and the outward-directed questions in such a way as make it intelligible that I am entitled to answer the former by answering the latter. (Cassam 2011, 12)

As Sacha Golob puts it, “Moran’s position is not epistemically immediate because the subject needs to be aware of the principle that when she judges \( P \) it follows that she believes that \( P \)” (Golob 2015, 12).

---

16 Sacha Golob makes an interesting claim when discussing the linking assumption: “to make the self-ascriptive move I must assume that my judgments have this power. But, as we have seen, so must Moran himself. In most cases,
If this is true, then this is somewhat problematic for the rationalist account because one of its central claims regarding immediacy must be given up. The fact that we need to make the above assumption regarding the relation between my judgements and my beliefs is precisely what makes the transparency procedure epistemically inferential rather than immediate. In any case, it seems that Moran's claim that "the basic concept of first-person awareness that we are trying to capture is that of awareness that is not based on evidence, behavioral or otherwise" (2001, 11) is too quick. Transparent self-knowledge is based on evidence, even if it isn't behavioural evidence.

I don't think that understanding rationalism along the inferentialist lines suggested above presents a very serious problem, unless one takes Moran's claims about immediacy to be his central claim. But this does not seem plausible to me. The central claim seems to be that the practical standpoint of the first person is fundamental to understanding non-alienated self-knowledge and that I can come to know what my attitudes are by deliberating on some subject matter. This claim could still be true (for some of our attitudes, at least).

What's of particular interest, though, is that the discussions regarding the linking assumption raise the question of when one would be entitled to make that assumption. Usually? Always? In some cases? Moran does not seem to take seriously the possibility that there could be circumstances in which one would not be entitled to assume that one's judgements constitute one's beliefs, given that Moran thinks asking a rhetorical question suffices (cf. "And now, let's ask, don't I make just this assumption, whenever I'm in the process of thinking my way to a conclusion about some subject-matter?" (Moran 2001, 405)). This in turn appears to agents will in fact move between judging that P and self-ascribing the corresponding belief automatically, where they are aware of countervailing forces, they will rightly be hesitant (consider belief ascription by agents who have just been prompted by reading the implicit bias literature)" (Golob 2015). The sentence in brackets is quite surprising, and has far-reaching consequences that I don't think Golob himself takes seriously enough. For if Golob is right that the linking assumption can be overruled in the above sort of circumstances, which indeed does not seem unlikely, then this would mean that one is arguably not entitled to make the linking assumption if one is prompted by reading implicit bias literature, or, which seems equally plausible, if one simply 'knows about' this literature and this is one of the things one considers when answering the question of whether P is true. I think this is quite radical, for it means that when answering the question of whether one believes all human beings are equal, say, one of the things to take into account when answering this question is (research on) implicit bias, in which case it would be conceivable that the reasons one considers when answering the world-directed question of whether P is true can include insights from scientific psychology. I imagine that this is what a holist version of rationalism might want to accommodate, but I will not pursue the relevance of empirical psychology to the linking assumption here.
suggest that one is always entitled to assume that one's judgements and one's beliefs are related in the right way (the right way being judging that P determines one's belief that P). But mightn't there be circumstances (e.g. psychological or social) in which we should not assume that our judgements constitute our beliefs, and so a 'gap' occurs between the two? If so, what sort of circumstances might those be?

The reason for calling attention to these questions is because it matters to our understanding of rationalism construed along atomist lines - a topic I will take up in the next chapter.

7. Conclusion

In this chapter I have described the key points of the rationalist account and have discussed some of the more common objections to it and ways in which the rationalist might respond to them. I considered in detail the claim that says we should understand 'transparency' not as a procedure but as a transcendental condition, which I argued is a promising response in some respects, but is incompatible with Moran's epistemic claims about self-knowledge and normative claims regarding alienation. In the final section, I turned to the question of whether rationalism is a version of inferentialism or not, which I suggested is relevant to the question of whether or not rationalism is a version of atomism, which I will address next.
5
The Limits of Rationalism: Failures of Transparency Part I

1. Introduction

Consider the Angry Spouse. The Angry Spouse is having a fit of anger and asks herself “Do I believe a divorce is a good idea?”. To answer this question, the Angry Spouse decides to answer the corresponding world-directed question “Is a divorce a good idea?”. She deliberates over the reasons in favour of the subject matter \( P \) (“a divorce is a good idea”). But, angry as she is, the only reasons in favour of \( P \) that are salient to her are ‘angry reasons’ (e.g. “This is the third time this month we’ve got into a fight!”, “I’m not happy”, “Being afraid of not being together is no reason not to get divorced,” etc.). Reflecting on her angry reasons, then, the spouse’s deliberative reflection results in a positive judgement: yes, a divorce is definitely a good idea.

The question is, does the Angry Spouse, when making up her mind, have self-knowledge, according to the rationalist? It does not seem evident that she does. A well-known piece of anger-management advice, after all, is to count to ten before you decide to do anything. Maybe we should give the Angry Spouse similar advice (“Don’t believe what you judge”). It’s not hard to imagine that when the Angry Spouse has cooled off, she will see the world quite differently, and so will be inclined to answer the world-directed question differently, too. Maybe she will no longer think that getting a divorce is a good idea after all – it’s just something she thought in the heat of the moment. However, it’s not evident that the Angry Spouse necessarily lacks knowledge of her beliefs and desires, either. Maybe her judgement that getting a divorce is a good idea really does reflect her attitudes.

1 By ‘angry reasons’ I simply mean reasons that are salient to one when one is angry. Presumably, it is not just the content of her reasons that is affected by her anger but also her reasoning style.
The question of whether or not the Angry Spouse has self-knowledge is hard to answer, I want to suggest, because the rationalist's proposed way of acquiring self-knowledge is simply 'deliberate'. But deliberate how? Deliberate when? Deliberate with which reasons? In other words, the dictum 'make up your mind' does not tell us very much about what making up your mind comes down to in practice. This, in turn, means that it's not clear when making up your mind is conducive of self-knowledge.

Moran and other rationalists have not paid much attention to the question of what exactly is required for making up your mind to be attitude-constitutive and therefore knowledge-conducive. In this chapter, I therefore evaluate the rationalist account of self-knowledge, doing so by concentrating on how it deals with the circumstances of deliberation. I address the question of what counts as 'failures of transparency', i.e. cases in which following the transparency procedure is not attitude-constitutive and so not knowledge-conducive. I propose that regarding the question of when and in circumstances making up your mind delivers self-knowledge, there are different possible interpretations of rationalism. One is the radical atomist rationalist view (hereafter radical rationalist), which is the view that by making up your mind you always acquire self-knowledge, no matter what the circumstances. In other words, the radical rationalist thinks that all circumstances are 'good' circumstances for deliberation. Another option is a moderate atomist rationalist version of rationalism (hereafter moderate rationalist), according to which making up your mind is only self-knowledge conducive in normal circumstances of deliberation. In other words, there might be circumstances, such as anger, in which answering world-directed questions and coming to a judgement is not sufficient for constituting one's attitudes and hence is not sufficient for acquiring self-knowledge. This view requires making a non-arbitrary distinction between what counts as the normal or right and what counts as abnormal or wrong circumstances of deliberation. Given that deliberation is so fundamental on the rationalist's view, the least s/he can do is recognize that we sometimes deliberate rather poorly, and address the question of whether or not this has any consequences for his/her account of self-knowledge.

I proceed as follows. First of all, in Section 2, I describe what I'll refer to as the 'Anscombean Constraint' (2.1) and describe the radical rationalist view in terms of it (2.2). I present some prima facie problems and challenges for this view in Section 3. One such problem is what I call the 'garbage in, garbage out' objection, which is the objection that if someone makes up her mind in the 'wrong' sort of circumstances, i.e. if 'bad' reasons go into the transparency procedure, then
the transparency procedure may likewise have a ‘bad’ outcome: self-deception instead of self-knowledge. In Section 4, I describe moderate rationalism, the view according to which making up one’s mind is only self-knowledge conducive if one makes up one’s mind in the right or normal circumstances, and briefly address the (exegetical) question of where we are supposed to place Moran’s writings in the radical/moderate rationalist landscape. Next, I turn to what seems to be a straightforward response to the ‘garbage in, garbage out’ objection, which is to filter out the wrong circumstances or the ‘bad input’ to the transparency procedure (Section 5). I argue that this solution fails to be convincing by arguing that the exceptions seem to become the rule (5.1) and that one’s ‘considered reasons’ are neither necessary nor sufficient for acquiring self-knowledge (5.2).

Before I begin, I need to make an important qualification, namely that I will be working on the assumption that a subject is actually able to follow the transparency procedure/make up her mind about some subject matter. After all, rationalism does not tell us what to say about subjects who are altogether unable to deliberate or reach a conclusion by reflecting on some subject matter because of whatever circumstance. I will thus ignore contexts or moods in which the subject fails to be able to deliberate, judge or avow anything at all. Such a subject, I imagine, would not acquire self-knowledge by following the transparency procedure, but only because she would not be following the transparency procedure in the first place. I will concentrate on cases in which a subject is (capable of) making up her mind and not in those circumstances that make it hard or impossible for someone to make up her mind and prevent her from acquiring transparent self-knowledge. The latter is, though, I think, an important issue that is underexplored and certainly deserves more attention.

2 An objection with the same name has been raised with respect to Rawls’s method of reflective equilibrium (Jones 2005, 66, 74; de Maagt 2016). Using the same term to describe the objection to the transparency procedure is deliberate. This is because the way in which the rationalist could respond to the objection is similar to the way in which advocates of reflective equilibrium might (or indeed have) responded, and it is problematic for similar reasons.

3 As with the notion of ‘procedure’, I take the notion of ‘input’ very broadly as simply referring to ‘the material one deliberates over’.

4 I should add that, since I’ve already discussed why rationalism has trouble dealing with knowledge of attitudes other than belief, my focus here will be on beliefs, more specifically on whether and when one would be entitled to assume that one’s judgement that P determines one’s belief that P. Beliefs are the ‘home base’ of rationalism, and hence seeing whether and how the transparency procedure works for beliefs is of particular relevance.

5 For an interesting discussion regarding the relevance of self-respect for being able to deliberate at all and so come to acquire transparent self-knowledge, see Beate Roessler’s (2015) paper ‘Autonomy, Self-Knowledge and Oppression’. See also Jan Branden, who argues that self-love is a necessary condition for self-knowledge (2015), and Katrien Schaubroeck (unpublished manuscript) for an insightful take on the role that self-confidence plays with respect to self-knowledge.
2. Radical rationalism and the Anscombean Constraint

2.1 The Anscombean Constraint

The rationalist's key claim is that the first-person question about one's attitude is "answered by reference to (or consideration of) the same reasons that would justify an answer to the corresponding question about the world" (Moran 2001, 62). But what reasons might those be? Can they include my friends' reasons? Other people's reasons? Or must they be 'my own' reasons? If that's the case, then might these also be my future self's reasons? What about my past self's reasons? If they must be my 'current' reasons, then what are those? Should we understand a person's current reasons simply to be 'belief-desire pairs'? Can feelings or emotions be reasons?

To begin answering the rationalist's reply to (some of) these questions, I want to start by considering a specific passage from Moran in which he appeals to Anscombe's work on intention, and which I take to be a good illustration of what Moran takes 'reasons' to be, which is as follows:

In belief as in intentional action, the stance of the rational agent is the stance where reasons that justify are at issue ... Anscombe's question "why" is asking not for what might best explain the movement that constitutes the agent's action, but instead is asking for the reasons he takes to justify his action, what he is aiming at. ... To do otherwise would be for him to take the course of his belief or his intentional action to be up to something other than his sense of the best reasons, and if he thinks that, then there's no point in his deliberating about what to do. Indeed, there is no point in calling it "deliberation" any more, if he takes it to be an open question whether this activity will determine what he actually does or believes. To engage in deliberation in the first place is to hand over the question of one's belief or intentional action to the authority of reason. (Moran 2001, 127)

Other rationalists and agentialists have made similar appeals to the Anscombean account of intentional action being based on the person's reasons for acting (Hieronymi 2008; Boyle 2009a, 2011a, Hamilton 2000, 2008, J. Roessler 2013, 2015b). On the Anscombean view, we have 'knowledge without observation' of our own intentions. The knowledge isn't perceptual, nor is it inferential; it's a type of 'practical' knowledge (Setiya 2011). There is a "privileged relation" between
The Limits of Rationalism: Failures of Transparency Part 1

the question of what an agent is doing and what she takes herself to be doing (this formulation is from Moran’s paper on Anscombe (cf. Moran 2004b, 64)). The latter is “not just another description, side by side with all the others, but has some claim to determine what the action itself is” (Moran 2004b, 44). Crudely formulated, we would not come to know ‘what the agent was doing’ when she acted the way she did if we did not take her own answer to the question of why she did it as giving us the (exhaustive) answer. On Anscombe’s view, then, “you cannot be paying, hiring, or marrying unless you are doing so intentionally”, and “in order to act intentionally, one must have practical knowledge of what one is doing” (Setiya 2011, 159). Such knowledge is, as Kieran Setiya points out, both necessary and sufficient because it constitutes the action it represents. Even if we did not consider the reasons she herself gives, whatever they are, in answer to a why-question, we could still give some account or ‘explanation’ of her doings, but it would not be one that would rationally explain her doings, i.e. would not explain her doings understood as an intentional action.

Anscombe describes the sort of knowledge that one has in acting intentionally as a form of practical knowledge, which she explains by using the (somewhat elusive) phrase that an intention is “the cause of what it understands” (Anscombe 1957, 87). Boyle explains the phrase as follows:

where A-ing is something I am doing intentionally, it is only insofar as I understand myself to be A-ing that I am A-ing, and it is only insofar as I understand myself to be A-ing on account of X, Y, and Z that I am A-ing on account of X, Y, and Z. My understanding makes these things the case, not by causally precipitating certain events whose unfolding does not itself require any contribution from my understanding, but by governing my activity as it unfolds, in light of the concept: A-ing, conceived as in something to be done. (Boyle 2009a, 140)6

6 Cf. Hieronymi: “Anscombe ... noted that whenever one intentionally Φ ‘s ... one can rightly be asked, ‘Why did you Φ?’ (or ‘Why are you Φ-ing?’) where this question looks, not for an explanation of how it came about that one o-ed, but rather for the agent’s reasons for Φ-ing.... A similar question is given application by belief whenever one believes that P ... one can rightly be asked, ‘Why do you believe that P?’ where that question looks, not for an explanation of how it came about that one believes, but rather for considerations that one takes to bear positively on whether P” (Hieronymi 2008, 259). What’s curious about Hieronymi’s take on Anscombe is that she formulates the why-question in the past tense (“Why did you Φ?”), whereas it appears to be fundamental to Anscombe’s approach that we have the special sort of knowledge while we are acting, and so the relevant question is (only, and not also) “Why are you Φ-ing?” (see also Boyle 2009a, 138).
The rationalist approach to self-knowledge can be understood as applying the Anscombean account of the type of knowledge we have of our own intentions regarding our other attitudes, most notably, our beliefs. The idea is that there is some privileged relation between the question of what an agent \( q \) \( P \) (believes, desires, hopes for, and so on) and what she takes herself to \( q \), the latter of which requires specifying her reasons regarding \( P \), whatever they might be. Again, the thought would be that we would not come to know what the agent’s attitudes were if we did not take her own answer to the question of why she believes or wants what she does as giving us the (exhaustive) answer.

The fact that Moran says rather little about what he takes ‘deliberation’ and ‘reasons’ to be – and provides no answer to the sort of questions with which I began this section – is, I think, indicative of the fact that he is or wants to be neutral or non-committal about what deliberating or reasoning involves and what reasons one should deliberate ‘with’, so to speak. So instead of saying, as I have, that the problem is that Moran does not say much about these issues, what we should say is that he thinks rationalism does not need to say much, or anything, about them.

2.2 Radical rationalism

If we return to the question of what qualifies as the ‘ingredients’ for the transparency procedure, such that it delivers self-knowledge, one reading of the rationalist project is to see rationalism as posing only one constraint, which I will refer to as the ‘Anscombean Constraint’. The Anscombean Constraint holds that the (only) material that can be used to answer world-directed questions in a way that results in that procedure having self-knowledge as its outcome is ‘deliberative’ material, i.e. the subject’s own reasons, or considerations that the subject draws on to the question whether \( P \) is true, desirable, cute, fearful, something to be hoped for, and so on. \(^8\)

The radical way of understanding rationalism comes down to saying that in order to acquire self-knowledge by following the transparency procedure, all you need to satisfy is the Anscombean Constraint. No constraints are put on the circumstances

---

7 For Boyle, this is his explicit goal: “my purpose here is, not to defend Anscombe’s view in its own right, but to draw a comparison between her understanding of our relation to our own actions and what we have seen about our relation to our own beliefs” (Boyle 2009a, 140).

8 I am sceptical about finding the right world-directed question for attitudes other than beliefs (see previous chapter), but here I assume for the sake of the argument that we can find the right formulations for such questions. I address some of these issues below.
in which one makes up one’s mind. In other words, on this reading of rationalism, it is irrelevant what the psychological context is or what mood the subject is in while considering her reasons that she takes to bear positively on whether $P$ is true or desirable. For instance, whether a subject follows the transparency procedure ‘angrily’, ‘nervously’ or ‘reluctantly’, and so on, or even while depressed or intoxicated, is irrelevant for the epistemic potential of the transparency procedure. Also, it does not matter epistemically what induces or causes a subject to engage in transparent question-settling (was it a question, a feeling, a conversation, a divorce, a change of careers, a romantic dinner, a therapeutic scenario or perhaps no apparent cause at all?).\footnote{I take it that the context in which subjects follow the transparency procedure can also include various social or political contexts, which, likewise, should be irrelevant. These moral-political circumstances require a separate, fuller discussion, however, which I will address in the next chapter.} If all you need to meet is the Anscombean Constraint, then all you need to know is ‘your reasons’ regarding the subject matter of your attitude, where ‘your own reasons’ are those reasons you can here and now come up with in support of the subject matter. Rationalism understood in this way would thus appear to place no constraints on how and/or whether the reasons that the subject uses to answer a world-directed question coheres with the subject’s other reasons or other attitudes (at that same time or at other times) or coheres with her plans or (life) projects, her character traits, personality, strengths, weaknesses, values or cares or her self-conception. Nor should it matter if the reasons a subject gives for wanting or believing something express a fundamental lack of respect for herself and/or violate the moral rights and/or duties she has to herself (cf. Hill 1991).

In short, what matters is not how the subject follows the transparency procedure or what material (which reasons) go into the procedure, but simply that she follows it. Whatever reasons the subject gives, these reasons will determine what her attitude is, so long as they are ‘her’ reasons, where the possessive pronoun refers simply to the reasons that are salient to her in whatever moods she is in, at whatever time, in whatever context and whatever their content. The constraints the subject must meet are formal ones, not substantive ones regarding the content of the subject’s reasons or how these reasons stand in relation to her values, self-conception or other attitudes she also has, nor indeed the frequency of deliberation (too much, too little).
If it were true that rationalism puts only one constraint on the transparency procedure to be successful, i.e. to deliver attitudinal self-knowledge, then that would explain why, in the context of rationalism, little has been said about what reasons are and whether or not all reasons are in principle good candidates that function as input for the transparency procedure, that is, to allow for the transparency procedure to issue judgements that are epistemically trustworthy with regard to self-knowledge.

It seems that Anscombean assumptions of the above kind explain why Moran does not seem to take seriously the possibility that there could be circumstances in which one would not be entitled to assume that one's judgements constitute one's beliefs. Recall that Moran thinks asking a rhetorical question suffices: “And now, let’s ask, don’t I make just this assumption, whenever I’m in the process of thinking my way to a conclusion about some subject-matter?” (Moran 2001, 405). This passage makes clear that whenever I answer some world-directed question, i.e. whenever I judge, and so, crucially, however I judge, I always have the right to assume that that judgement constitutes my attitude(s). This would (only) follow if one is an Anscombean regarding what should ‘go into’ the transparency procedure.

I am not trying to give an accurate description of rationalism or point out what it is necessarily committed to. The point is merely to consider one of the ways in which we might understand what ‘deliberation’ involves and what the rationalist appeal to making up your mind by reflecting on ‘the reasons’ comes down to, given that nothing much has been said about these issues. Having said that, it seems to me that the radical rationalist view faces some prima facie problems, which I turn to now.

3. Prima facie challenges and problems for radical rationalism

If all one has to meet is the Anscombean Constraint, then the first worry is that it appears to follow that we can give a straightforward answer to the question of whether the Angry Spouse knows her own mind when answering the world-directed question while in a fit of anger. Her angry reasons would appear to satisfy the Anscombean Constraint after all: they are ‘her’ reasons and are reasons that are available from her practical, first-person perspective. Assuming, with Moran, that one is entitled to assume that one’s judgements constitute one’s attitudes whenever one reasons one’s way to a conclusion about some subject matter, it appears we must conclude that the Angry Spouse settled her attitudes and acquired knowledge of what she
believed and wanted. This conclusion does not appear to be very nuanced. This does not mean we have to draw the opposite conclusion, that is, that the Angry Spouse is entirely self-ignorant when she makes up her mind. But there are many other options. I will discuss the metaphysics of belief in greater detail in Chapter 7, but for now, the prima facie worry is just that the radical rationalist view lacks nuance.

A second worry is that it’s not evident that the Anscombean account of self-knowledge of our intentions can be applied to our attitudes in general. In other words, the worry is that the Anscombean Constraint might work for Anscombe but maybe not for Moran. What arguably makes the Anscombean explanation of how we know our own intentions plausible is that what it is meant to explain is something that I am doing here and now. It seems prima facie plausible that I can know that I am now, for example, watering the flowers or baking a cake, and that I might know this even if I’m angry, jealous, tired, insecure, etc. All I need to know is that I want to water the flowers now, and to answer the why-question all I need to do is come up with the reasons that are available to me right now. Boyle rightly stresses this temporal/enduring aspect of Anscombe’s approach:

Something that is striking in Anscombe – something that sets her apart from many subsequent action theorists – is her resolute focus on action in progress. Anscombe’s “why?”-question is first and foremost “Why are you doing A?” ... Her assumption, in effect, is that to understand the nature of intentional action, we must describe the specific character of the subject’s relation to it as it unfolds. (Boyle 2009a, 138–39; but see also Setiya 2011)

The problem with many of our (other) attitudes, though, is that their existence is not restricted to the ‘here and now’. My beliefs, desires, hopes, expectations, and so on must have “a reasonable life expectancy” (M. Fricker 2009, 52). It’s not obvious to
think of beliefs and other attitudes as things that ‘unfold’, as Boyle puts it, in the way that a person’s actions unfold. Despite these prima facie metaphysical differences between intentions-in-action and things like beliefs and desires, the rationalist suggests the procedure is the same: I answer the question of what my attitudes are by coming up with reasons that are salient/available to me at this moment. By implication, this means that I should be able to answer the questions of whether I believe a career change is a good idea or whether I believe I should start learning French (these examples are from Moran (2001, 115)) by giving reasons in support of the subject matter I reflect upon here and now. But arguably I do not have the belief that ‘a career change is a good idea’ in the same sense as I have the intention to water the flowers when I’m watering the flowers. The temporal restrictions on the availability of my reasons may be unproblematic for ‘knowing what one is doing’, but arguably this is different for knowing (most of) one’s other attitudes.

A related worry is what we might call the ‘Nietzschean worry’. According to Nietzsche, we often lack awareness of our motives, indeed, he claims that it’s a “universal madness” to think that we generally know what we are doing; that “actions are never what they appear to be ... all actions are essentially unknown” (cited in Katsafanas 2015, 117, see also his 2012). That’s probably a bit strong, but the Nietzschean worry does bring out the fact that if we have reason to doubt that we generally know what we are doing, then these worries will automatically carry over to those theories of self-knowledge, such as Anscombean accounts, that rely on some idea that one normally or always knows what one is doing. This is a potential weakness of the view.

Finally, if the rationalist claims that the Angry Spouse ‘has what it takes’ to know her own mind, then this might render the rationalist account trivial. For if everything the rationalist wants to explain in terms of self-knowledge is present in the Angry Spouse’s case, then it’s hard to see how ‘transparent’ self-knowledge could still have “a special importance to the overall psychic health of the person” (Moran 2001, 136) and “matters to the well-being of the person” (Moran 2001, 137).

---

11 Paul Katsafanas also mentions the following passage from Nietzsche’s *Genealogy*: “Just as the common people separates lightning from its flash and takes the latter to be a deed, something performed by a subject, which is called lightning, popular morality separates strength from the manifestations of strength, as though there were an indifferent substratum behind the strong person which had the freedom to manifest strength or not. But there is no such substratum; there is no ‘being’ behind the deed, its effect and what becomes of it; the doer is invented as an afterthought,—the doing is everything. (GM I.13)” (cited in Katsafanas 2015, 157; for discussion of Moran on modelling knowledge of our beliefs on knowledge of our actions, see also O’Brien 2003, 38ff)
It's not unlikely that it would be better for the psychic health and well-being of the Angry Spouse to realize that her own avowals aren't to be trusted. If she were to trust her own angry judgements, that may well lead to greater self-deception. The problem with the Angry Spouse, then, is that if only the Anscombean Constraint matters, following the transparency procedure appears to be compatible with what we would ordinarily consider to be a possible case of self-deception. The thought is that if 'bad' reasons go into the transparency procedure, we should expect 'bad' judgements to be its outcome. I will call this the 'garbage in, garbage out' objection, which the Angry Spouse intuition is an instantiation of.

Clearly, the 'garbage in, garbage out' objection is only plausible given our intuitions about the Angry Spouse, and so presupposes some sort of standard. But appealing to our 'intuitions' about specific cases is certainly not always unproblematic: we cannot always take our immediate intuitions at face value, if indeed it's clear what, exactly, intuitions are (this is something I will discuss in more detail in Chapter 7). But the more fundamental point is that I think it seems reasonable to expect of any theory of self-knowledge, including rationalism, that it is able to either (a) accommodate the intuition that there's something unsatisfying about saying the Angry Spouse has self-knowledge or (b) explain why the intuition is misguided.

In Chapter 7, I will return to radical rationalism. In what follows, I concentrate on the first strategy — how the rationalist might accommodate the intuition.

4. Moderate rationalism

Moderate rationalism, as I understand it, can be understood as having the ambition to provide a response to the 'garbage in, garbage out' objection. The moderate rationalist thinks that saying that the Angry Spouse has knowledge of her attitudes is problematic. But in order to rule out cases like the Angry Spouse, the rationalist must do more than claim that the reasons one deliberates with when making up one's mind must meet the Anscombean Constraint. It involves recognizing that in certain circumstances, such as particular states of mind (e.g. anger), following the transparency procedure does not deliver attitudinal self-knowledge. According to the moderate rationalist, then, following the transparency procedure will be attitude-constitutive and knowledge-conducive if (1) the Anscombean Constraint is met and
(2) one follows the transparency under normal circumstances. If, for instance, you are very angry, like the Angry Spouse, then making up your mind does not deliver self-knowledge, because your judgement about the world is untrustworthy and fails to settle your actual attitudes. The advantage of a moderate rationalist view is that it would allow us to give a (somewhat) more nuanced answer to the question of whether the Angry Spouse has self-knowledge or not. The challenge, though, for the moderate rationalist, is to provide a non-arbitrary cut-off point between normal and abnormal circumstances.

One might wonder why the rationalist would have to appeal to ‘normal’ or ‘right’ processes of deliberation or judgement. In other words, the question is what would motivate an understanding of rationalism as requiring more than meeting the Anscombean Constraint? One reason is the Angry Spouse. But I think there’s a more important reason, which becomes salient if we consider the fact that rationalism appeals to the notion of ‘rationality’, which is a normative notion. If the ‘good’ case is being rational, then presumably there are two different sorts of ‘bad’ cases, namely being irrational and being a-rational. These three options (rationality, a-rationality, irrationality) can in turn be applied to rationalism about self-knowledge:

1. someone is rational: she follows the transparency procedure ‘rightly’, or ‘correctly’
2. someone is a-rational: she does not take up the deliberative stance at all but rather answers the question of what her attitudes are in a theoretical, empirical spirit
3. someone is irrational: she follows the transparency procedure ‘incorrectly’ or ‘badly’

When discussing failures of transparency, we can take such failures to mean either that someone does not follow the transparency procedure at all or that she follows the procedure in a bad way, because she judges badly or deliberates badly. And so appealing to ‘normal’ or ‘right’ circumstances of deliberation and judgement seems to be built in to the rationalist project from the very beginning. The Angry Spouse, then, can be considered to have made up her mind about something, but has done so ‘irrationally,’ or ‘badly’.

The challenge for the moderate rationalist is to respond to the ‘garbage in, garbage out’ objection without begging the question. One might, after all, respond
to the objection by saying that the 'right reasons' are just those reasons that aren't the result of self-deception, i.e. ruling out self-deception from the outset. However, this move is not only ad hoc but also means that such an account would not "have much of a point", because we can reasonably expect theories of self-knowledge to be compatible with there being "room for error" (Coliva 2009, 372). Such a strategy thus appears to beg important questions and would render the account circular.

More positively formulated, the 'garbage in, garbage out' objection is not a knock-down argument but rather provides a challenge, which is for the rationalist to come up with an account of which reasons would render the procedure trustworthy without becoming circular. In the next section, I discuss how the rationalist might respond to the objection by filtering out circumstances in which only 'bad' reasons are salient, the result of which is that deliberation does not have the desired result.

Before turning to the question of why an appeal to normal circumstances is problematic, it might be helpful to briefly consider the more exegetical issue: where does Moran stand in all of this? As a starting point, the following passage is of particular interest (where 'authority' can be understood as being short for 'having the capacity to constitute one's attitudes by answering a world-directed question'):

> Of course, this authority can be partial or hedged in various ways. When I know this to be the case, for instance when I know that I am akratic with respect to the question before me, that compromises the extent to which I can think of my behavior as intentional action, or think of my state of mind as involving a belief rather than an obsessional thought or a compulsion. Nor does a person speak with first-person authority about such conditions. (Moran 2001, 127–28 emphases added)

This is interesting. The passage suggests that if I know that I am akratic or compulsive in some domain – which involves knowledge of my habits and/or character traits – then I am not entitled to assume that my judgements constitute my attitudes, i.e. I am not entitled to make the 'linking assumption', and so answering a world-directed question in such a case is not going to give me self-knowledge. Perhaps Moran would think that the Angry Spouse, likewise, is not entitled to assume that

---

12 Crispin Wright (1984b) suggests something along these lines when he claims (in the context of so-called 'C-conditions') that the subject should not be self-deceived.
her judgements constitute her beliefs. What's interesting about Moran's passage is that it allows, in principle, for a 'holist' version of rationalism, where the question of whether making up one's mind allows one to have self-knowledge cannot be answered by ruling out a few exceptional circumstances, such as knowing that one is compulsive or akratic, or indeed the emotional circumstances like those of the Angry Spouse, but can depend on whether one is tipsy, hungry, tired, depressed, nervous, in love, and so on.

However, Moran does not discuss the circumstances in which one's 'authority can be hedged' other than akrasia and compulsion. This is compatible, of course, with his account being non-atomist. I take it to be more likely, however, to be indicative of the fact that Moran thinks the list of excusing circumstances is rather short and that we don't need to concern ourselves with it too much (recall also that Moran thinks that whenever one is in the process of thinking one's way to a conclusion about some subject matter, one is entitled to make the linking assumption (Moran 2001, 405)).

However, my interests at this point are not primarily exegetical. Given the rationalist's silence regarding the question of what qualifies as the right material (reasons) for the procedure or what sort of deliberation would make the procedure successful, I am primarily interested in the question of how rationalism could be understood by exploring answers to these questions rather than reconstructing how existing rationalist views are to be understood.

Having described the different options regarding rationalism, we can now turn to the more critical question: are there indeed just a few obvious or trivial circumstances in which following the transparency procedure is not going to deliver self-knowledge? Can we therefore simply add 'normal conditions of deliberation' to our definition of what is required for the transparency condition to be successful and the linking assumption to be legitimate by default?

5. The 'filtering strategy' as a response to the 'garbage in, garbage out' objection

A natural way for the rationalist to respond to the 'garbage in, garbage out' objection is to filter out the garbage. For instance, the rationalist could filter out the 'angry circumstances' of making up one's mind. In this section I argue that the filtering
strategy faces two problems. The first problem (5.1) is that if we begin by filtering out certain circumstances, it’s not clear when and where we should stop. In other words, the exceptions seem to become the rule. The second problem (5.2) is that many exceptions or supposed ‘bad’ circumstances aren’t necessarily bad circumstances at all.

5.1 Why it filters out too much
The first problem with the filtering strategy is that the case of the Angry Spouse generalizes. In other words, the filtering strategy will involve ruling out quite a lot of reasons that one ought not to appeal to when answering world-directed questions, at least if doing so is meant to deliver self-knowledge. For we can easily imagine a Depressed Spouse, an Overexcited Spouse, an Afraid Spouse, an Ashamed Spouse, a Jealous Spouse, and so on. Consider this example from Annalisa Coliva:

[Take a jealous wife who sincerely asserts with her friends that she believes that her husband is faithful to her—and has good reasons to do so—but, then, once at home, is often inquisitive, searches his belongings, etc. ... So, she is self-deceived, since she sincerely avows a belief and behaves in ways that run contrary to it. (Coliva 2009, 372)]

If the Jealous Spouse wants to know her own attitudes and does so by answering the question “Do I believe my husband is faithful to me?” by answering the corresponding world-directed question “Is my husband faithful to me?”, then, in her jealous moments, all the Jealous Spouse can consider are ‘jealous reasons’ and so the judgement she arrives at will be a jealous judgement. The ultimate outcome will be self-deception, not self-knowledge. The same can be said of, for example, overexcited reasons: if you’ve just gone bungee-jumping, or if you’ve just walked across a shaky bridge (as in the famous Dutton and Aron (1974) experiment), then this is probably not the best moment for you to try and acquire self-knowledge ‘transparently’. This is true mutatis mutandis for other emotional states of mind.

13 This isn’t the (only) conclusion Coliva herself draws, though. Coliva instead thinks, following a proposal by Akeel Bilgrami (2012), that she has two different beliefs. This is the view that Schwitzgebel (2010) refers to as the ‘contradictory belief view’. I think the contradictory belief view can make pretty good sense of the Jealous Wife (as believing both P and not-P), but it’s not plausible when it comes to the Angry Wife, because she followed the transparency procedure in the heat of the moment, i.e. the Angry Wife doesn’t appear to really believe P at all.
The rationalist might stick with the filtering strategy and claim that depression, fear, excitement, shame and jealousy don’t belong to the right circumstances either. Excited judgements, shameful judgements, fearful judgements, and so on aren’t the right sort of judgements: these judgements don’t constitute one’s attitudes, and so they don’t provide the subject with self-knowledge. This, though, will amount to ruling out judgements made on the basis of ‘emotional reasons’ in general, which is a rather radical move.

Also, filtering out emotional reasons won’t do, for it seems we should also rule out judgements made in particular ‘noetic’ states of mind, such as being, for example, overconfident, impulsive, closed-minded, careless, and so on (cf. Dokic 2012). These are not emotional states of mind, but they may well sometimes be problematic states of mind for making up your mind in such a way that it provides you with self-knowledge. If you’re very insecure, for instance, and don’t really trust your own reasons, then it’s reasonably likely that your judgements won’t constitute your attitudes. Consider, for instance, the following excerpt from Simone de Beauvoir’s diary, as discussed by Miranda Fricker in Epistemic Injustice:

Day after day, and all day long I measured myself against Sartre, and in our discussions I was simply not in his class. One morning in the Luxembourg Gardens, near the Medici fountain, I outlined for him the pluralist morality which I had fashioned to justify the people I liked but did not wish to resemble: he ripped it to shreds. I was attached to it, because it allowed me to take my heart as the arbiter of good and evil; I struggled with him for three hours. In the end, I had to admit I was beaten; besides, I had realized, in the course of our discussion, that many of my opinions were based only on prejudice, bad faith or thoughtlessness, that my reasoning was shaky and my ideas confused. ‘I’m no longer sure what I think, or even if I think at all,’ I noted, completely thrown. (De Beauvoir 1959, 344; cited by M. Fricker 2009, 50–51; for discussion, see also B. Roessler 2015; and M. Fricker 2003)

If Simone de Beauvoir had made up her mind under these circumstances, i.e. if she were to have answered a mind-directed question such as “Do I want to be a philosopher?” by answering a world-directed question, e.g. “Are there good reasons for me to be a philosopher?”, then would she have come to know what she wanted or believed? This is a very hard question to answer. I don’t think we should conclude that
de Beauvoir was necessarily self-deceived because she came to think “philosophy is not for me”, and “soon moved to identify herself not as a philosopher but as a writer instead” (M. Fricker 2003, 218). The point is rather that it is possible that following the transparency procedure in certain states of mind, such as being very insecure or overconfident, could lead to self-deception rather than self-knowledge. And for the transparency procedure to have ‘self-knowledge’ as output, the rationalist will have to filter such reasons out, too.

The overall worry should be clear by now, namely that the filtering strategy makes the arsenal of reasons the subject can draw upon when answering world-directed questions in a way that would provide her with self-knowledge rather limited, because so many reasons need to be filtered out. The exception seems to have become the rule. The second problem with the filtering strategy is that the reasons that remain — call them one’s ‘considered’ reasons — are neither necessary nor sufficient for a judgement that would constitute and provide knowledge of one’s attitudes.

5.2 Why ‘considered reasons’ are neither necessary not sufficient for self-knowledge

Up till now, I’ve suggested that the Angry Spouse (and the other Emotional Spouses) failed to acquire self-knowledge by following the transparency procedure for the sole reason that she was angry. But is this necessarily true?

Notice that when the Angry Spouse judges that P, it’s not the case that she is self-deceived simply because her answer to the relevant world-directed question was based on ‘hot’ or ‘emotional’ reasons. Emotional reasons, and judgements that result from reflecting on them, can be perfectly good reasons and judgements: there’s nothing the matter with them in principle. Indeed, some have argued that one’s emotions hold the key to acquiring self-knowledge (Mackenzie 2002). I am less optimistic, but I am optimistic enough to at least be sceptical of filtering out all emotional reasons just because emotional reasons can, on the face of it, sometimes be self-deceiving.

To see why someone’s judgements made on the basis of ‘hot’, emotional reasons might be conducive of self-knowledge and hence should not be filtered out of the transparency procedure. Consider a variation on the Angry Spouse. Meet the

---

14 And maybe not just in certain states of mind, but in certain sociopolitical circumstances, too (cf. literature on adaptive preferences, e.g. Elster 1983; Rüülik 2014; Stoljar 2015). See also Chapter 7.
Deferential Wife. The Deferential Wife is in many ways like the Angry Spouse: she too follows the transparency procedure and does so on the basis of angry reasons. However, the Deferential Wife is also importantly different, because, unlike the Angry Spouse, the Deferential Wife is utterly devoted to her husband:

She buys the clothes he prefers, invites the guests he wants to entertain, and makes love whenever he is in the mood. She willingly moves to a new city in order for him to have a more attractive job, counting her own friendships and geographical preferences insignificant by comparison ... She does not simply defer to her husband in certain spheres as a trade-off for his deference in other spheres. On the contrary, she tends not to form her own interests, values, and ideals, and when she does, she counts them as less important than her husband's. She readily responds to appeals from Women’s Liberation that she agrees that women are mentally and physically equal, if not superior, to men. She just believes that the proper role for a woman is to serve her family. As a matter of fact, much of her happiness derives from her belief that she fulfills this role very well. No one is trampling on her right, she says; for she is quite glad, and proud, to serve her husband as she does. (Hill 1991, 5; Westlund 2003, 485)

Now suppose that one day, when, for example, her husband comments negatively on the dinner she’s made, the Deferential Wife has finally had enough and gets very angry. She comes to judge, by considering the reasons available to her at that time – her angry reasons – that getting a divorce is the only right thing to do.

It’s not obvious, in the Deferential Wife’s case, that her judgement that it is true (that a divorce is the right thing to do) does not constitute her actual belief just because it’s based on ‘angry reasons’. Anger can be a good guide for what one really believes or how one should act. Anger may be instrumentally useful, and even necessary, for the Deferential Wife to secure a basic level of self-respect. In the example of the Deferential Wife, though not that of the Angry Spouse, it seems plausible to think that it’s about time she gave her own hurt feelings and her emotional reasons their due. This is true even if the Deferential Wife falls back into her old submissive routines shortly afterwards, or perhaps even feels ashamed for having judged that a

---

15 As Hill points out, the Deferential Wife is confused about herself: she believes she has a duty to defer to her husband but “she cannot fully understand that she has a right not to defer to him” (Hill 1991, 10), the latter of which “betrays a certain kind of self-respect” (Hill 1991, 6).
divorce is a good idea. If that were to happen, it would not be strange to think that it's unfortunate that she feels ashamed and falls back into old habits of servility rather than thinking she went back to being her actual self. The fact that the Deferential Wife might later come to regain trust in her non-emotional or 'considered' reasons is arguably precisely what makes her case tragic. She had self-knowledge and then lost it.

There are many other examples in this context that strengthen the idea that the filtering strategy filters out reasons that would allow one to acquire self-knowledge. Nomy Arpaly, for instance, imagines a Victorian Lady wondering what her sexual desires are, if she's got any in the first place:

If a Victorian lady experiences her sexual desires as alien, intrusive, “not truly her own,” our natural reaction is to tell her that she is wrong, that these desires are in fact her own, and that only the false, asexual self-image that she acquired with her upbringing makes her experience them as threatening to her integrity as a person. We would think that she denies her real self. (Arpaly 2002, 16; 123)

Suppose Victorian Lady tries to answer the question of what her sexual desires are by answering the corresponding question of what her sexual desires ought to be. If she were to do so in her cooler, non-aroused moments, then the chance will only be greater that this will lead to self-deception. Something similar must be said of Lynn (Arpaly 2002, 16), who “discovers that she is a lesbian and is deeply disturbed by that discovery” and whose “homosexual desires conflict with her values and her sense of her identity”. Lynn does not “want her desires to motivate her into action under any circumstances—the very thought scares her more than anything else” (Arpaly 2002, 16). In terms of self-knowledge, Victorian Lady and Lynn are arguably better off not trying to figure out what their attitudes are by following the transparency procedure on the basis of their ‘considered’ reasons.

The point here is not that one's emotional reasons are always representative of one's attitudes. The point, rather, is an anti-atomist one: focusing on the procedures or methods of acquiring self-knowledge, such as following the transparency procedure, just does not seem to be the right way to go. Whether or not emotional reasons lead to a judgement that would constitute and thereby provide knowledge

---

16 Or, if one is unhappy with this formulation of the world-directed question, we can imagine that she answers the world-directed question of whether such-and-such is desirable or desirable to her.
of one's attitudes is not something we can say is in general true, for whether or not certain circumstances or states of mind are 'bad' states of mind for successfully following the transparency procedure depends on the specifics of the case, i.e. the specific individual who is making up her mind. It depends, for instance, on what role the reasons considered by Victorian Lady, Lynn or the Angry Spouse play relative to their values, self-conception and other attitudes and the broader social circumstances in which they find themselves. The question of 'which' reasons are the 'right' reasons to make up one's mind with such that one acquires self-knowledge is something that depends on the person one is and the life one leads.  

When trying to find out what we feel, want or believe, we cannot always trust our transparent outlook on the world, including our carefully considered reasons. In this context, Victoria McGeer's (e.g. 1996, 2008) account of self-knowledge is particularly instructive. McGeer suggests (in her discussion of Bulstrode, a character from George Eliot's Middlemarch) that one's reasons may sometimes be "hijacked by psychological forces" that are "completely invisible" from the first-person perspective. McGeer writes that if that's true, then Moran's Kantian ideal of “handing over the question of one's beliefs or intentional action to the authority of reason" cannot be an entirely happy one. At the very least, we must give sober consideration to how vulnerable we are to such corruptions of reason and how best they can be guarded against. (McGeer 2007a, 98)

17 Sometimes deliberating about world-directed issues and coming to a conclusion in general can lead to self-deception. In other words, 'transparency' itself can be problematic. Jonathan Lear (2004) provides a good example of 'Mr. A', a patient suffering from borderline personality disorder who constantly interprets events in his life under the concept of betrayal. The problem for Mr. A, as Lear explains, is that the reasons he gives for supporting his beliefs and feelings fulfil a constraining function rather than one that facilitates (rational) freedom. The more Mr. A transparently avows, the more locked in he becomes to his interpretative schemas: "For Mr. A, self-conscious reflection is a manifestation of his unfreedom: for as he reflects on his reasons for feeling betrayed, he digs himself ever deeper into a crabbed and constraining world. Rational, self-conscious reflection is being deployed as a defense, one which helps sustain the betrayed world. For him, avowal is the culmination of his imprisonment" (Lear 2004, 453). The avowal of betrayal effectively blocks, at least at that moment, assessment and avowal of other (and more adaptive) feelings, desires and beliefs that he also (and perhaps also currently) has, e.g. his love for his girlfriend, his desire for personal intimacy or his belief that most people generally mean well, etc., but which contrast with the general theme of betrayal. The case of Mr. A indicates how deliberative avowal as such may have an undermining effect on self-knowledge and obstruct self-understanding. For discussion, see Strijbos and Jongepier (forthcoming).
[D]eveloping deliberative autonomy or “spontaneity” in Moran’s purist sense can be a sign of real psychic disease, indicating a capacity to manipulate oneself through the power of one's own reason into a condition of deep self-deception. (2007a, 92)

McGeer reminds us of the fact that in everyday life, merely making up one's mind about something or avowing or committing oneself to some proposition in a deliberative spirit is often not enough to determine one's mental states. She argues that we need self-regulation to guard us against the more subversive inclinations of our rational faculty. Self-regulative agency implies that one is continually ready to take a step back from one's first person, rationalizing inclinations in order to reflect on them from a second- or third-person point of view on self, thereby making a more comprehensive assessment of one's own situation and the relevant factors (contextual, psychological, pharmacological, etc.) that shape one's reasoning and one's capacity to stay true to one's commitments (see also de Bruin, Jongepier, and Strijbos 2015).

The upshot of all of this is that answering world-directed questions on the basis of the ‘wrong reasons’, ‘bad deliberation’ or following the transparency procedure irrationally is not necessarily going to give one knowledge of one's attitudes. Someone's considered reasons aren’t sufficient for the transparency procedure to deliver self-knowledge, because one's considered reasons may be self-deceptive. It should also be clear why someone's considered reasons are not necessary for the transparency procedure to deliver self-knowledge: one's emotional reasons can be material that is just as ‘good’ as the reasons that are available to the subject in calmer or cooler states of mind – indeed, sometimes they are much more trustworthy. Judgements made on the basis of reasons that are emotional or hesitant, and so on aren't necessarily self-deceptive, and one's non-emotional, cool or considered judgements aren’t necessarily self-knowledge conducive. Taking all of the issues just discussed into consideration, the strategy of filtering what goes into the transparency procedure does not seem to be very promising.

The more general conclusion is that the moderate rationalist version of rationalism is problematic. The moderate rationalist holds that following the transparency procedure will, in normal circumstances, be sufficient for a subject to acquire knowledge of her attitudes. However, there are good reasons to think that the appeal to normal circumstances, by filtering out bad reasons for transparent
deliberation, is problematic, because there aren’t just a few obvious or trivial circumstances in which following the transparency procedure does not deliver self-knowledge. Also, the reasons that are available to someone in emotional, hesitant or overexcited circumstances, and the judgements that follow from them, are not necessarily self-deceptive, and one’s considered judgements are not necessarily self-knowledge conducive.

It seems we must conclude that much of the work that the transparency procedure was supposed to do is in the end is done by the circumstances rather than the procedure itself. If the circumstances end up doing more work than the procedure, though, then the idea that we can say that the transparency procedure provides self-knowledge, all else being equal, becomes empty.

The rationalist can do one of two things. One is to ‘go radical’ after all and deny that there are circumstances where making up one’s mind does not constitute one’s attitudes, or, differently formulated, to simply claim that all circumstances are the right circumstances to acquire self-knowledge by making up one’s mind. In this chapter, I’ve only articulated some prima facie worries and challenges for the radical rationalist view, and so a fuller discussion will be necessary, which is the task of the next chapter. The alternative option would be to see whether rationalism can be understood along holist lines. I consider this option below.

6. Holist rationalism: the self as part of the world?

What would a non-atomist or holist version of rationalism look like? Very roughly, we can imagine that a holist version of rationalism is something like the reverse of radical rationalism. A holist rationalist might claim that whether or not instances of a person having made up her mind — in short, her judgements — constitute her attitudes depends on, for example, her previous and future actions, her overall state or the mood she is in, the (social) circumstances in which she makes up her mind, or what induced her to engage in transparent question-settling, and whether she is disposed to take steps to live up to the commitments underlying her self-ascriptions (cf. de Bruin, Jongepier, and Strijbos 2015). Her reasons must not just satisfy the Anscombean Constraint but must also cohere with the subject’s other reasons or other attitudes (at the same time or at other times), her plans or (life) projects, character traits, personality, strengths, weaknesses, values, cares and self-
conception. In short, in order to know whether someone acquires self-knowledge by making up her mind, we need to know more than just the fact that she has done so — we need to know when, where, why and how she made up her mind.

Construing rationalism along holist lines requires tweaking or reformulating the world-directed question; in other words, requires redefining what counts as the possible material of or 'input' to the transparency procedure — what counts as a 'reason' that one can draw on when answering world-directed questions or making up one's mind. The 'input' should be broadened to include not just the reasons that are available to me here and now, in my present mood and situation. If the rationalist were to 'go holist' then we would have to include values, feelings and emotions and the subject's plans and self-conception, and so on as belonging to the pool of 'reasons' that she could reflect on when answering the question of whether \( P \) is desirable or true, etc. Put differently, a holist version of rationalism is one in which the world-directed question is given a 'self-referential' twist. A formulation of the world-directed question that includes a self-referential component, such as, "Is \( X \) desirable to me?" would give the right outcome, i.e. would provide one with self-knowledge (see e.g. Kloosterboer 2015).

This would amount to making 'the self' part of 'the world', i.e. part of the world-directed question. Giving the transparency procedure a self-referential twist seems like the right way to go, at least if we want the procedure to actually provide one with knowledge of one's attitudes. When I answer the question of whether \( P \) is true, desirable or to be feared, then what I'm supposed to do is not reflect on \( P \) simpliciter, but rather reflect on how I relate to \( P \), that is, how my self relates to the world. When answering the question of whether a career change is a good idea, I should not consider this question about whether a career change is a good idea in general but whether it's a good idea for me. What I'm doing then is answering a question not about the world as such, but 'the-world-as-I-see-it', or 'the-world-as-I-take-it-to-be'.

Seeing the world-directed question as having a self-referential component might also go some way to dissolving an important 'puzzle' that those who have written on transparency have been much concerned with: the puzzle of how it's possible to answer a question about one subject matter — one's mind — by answering a question about a completely different subject matter, namely the world (Cassam 2011, 2014, Byrne 2005, 2011a; O'Brien 2007; Golob 2015; J. Roessler 2013). What would give one the right to assume that answering the question of whether \( P \) (something
about the world) provides one with an answer to the question of what one's belief is (something that concerns my mind)? Those who accept there is a puzzle here suggest that we need some sort of story to explain how answering the question of what one believes by answering the question about another subject matter – the world – can be made intelligible. Alex Byrne, for instance, writes that the idea of transparency is that "one can know that one believes (or knows) that it's raining by checking on the rain" and goes on, "But surely meteorology sheds little light on psychology!" (Byrne 2005).

But if we give the transparency procedure a self-referential twist, as it seems we must, there is a catch. Recognizing the fact that the self is something to be taken into consideration when answering a world-directed question, or, differently put, that the world-directed question has a self-referential component, means that the transparency procedure is not really very ‘transparent’ after all. This is how Moran describes transparency:

> from the first-person point of view, the question “Do I believe that P?” is transparent to a corresponding question “Is P true?,” a question which involves no essential reference to oneself at all. (Moran 2001, xvi; 106 emphases added)

But on the self-referential view, the question of whether P is true/desirable/to be feared, and so on does involve an “essential reference to myself”, because when answering the question of whether P is true, I am asking whether accepting that P would, for example, cohere with my other reasons, attitudes and values. In short, in answering the question of whether P is true, I am trying to answer the question of whether P being true would be compatible with the sort of person that I am. The transparency procedure thus has a crucial, ineliminable “reference to oneself” and in that respect isn’t transparent.

This also has consequences for how and whether the self-referential formulation handles the ‘puzzle’ mentioned above. On the self-referential version of transparency, the puzzle dissolves, but only because we’re answering a mind-directed question by answering another mind-directed question. The puzzle dissolves because we’re no longer dealing with two different subject matters (mind and world). What we have ended up with is that the mind-directed question “Do I want X?” is answered
by answering the question "Is X desirable to me?". If the transparency procedure ought to be construed along self-referential lines, then the procedure can hardly be said to be a useful procedure for acquiring self-knowledge. It seems that what we're doing is answering one mind-directed question by answering another mind-directed question, which seems to make the procedure trivial.

There's a general problem with reformulating the world-directed question in general, which is that any reformulation that is going to work will work only because it appears that the self-knowledge we want to account for has to be assumed from the beginning. What we're doing is reformulating the transparency procedure such that it has the right outcome (self-knowledge), but we weren't supposed to know what the 'right outcome' was, because that's what the transparency procedure was meant to tell us. The general strategy of trying to find the right formulation of the world-directed question appears to have to assume what our attitudes are and involves adjusting the world-directed questions to them accordingly.

In short, if we take rationalism to offer a theory of how we know our own attitudes, then modest rationalism is implausible, and a holist version of the transparency procedure may not have much of a point. But I don't think this is the end of the story. Maybe transparency isn't meant to be a procedure for acquiring self-knowledge at all – rather, it's a normative theory that is meant to tell us what it means (not) to be alienated with respect to one's own attitudes. This version of rationalism will be addressed and evaluated in Chapter 7. The other option is to stick with the claim that rationalism is meant to offer an account of self-knowledge, and to 'go radical' and avoid making an appeal to normal circumstances from the start.

7. Conclusion

In this chapter I have addressed the question of whether the circumstances of making up one's mind are relevant to the question of doing so provides the subject with self-knowledge. To this end, I have distinguished between two different atomist versions of rationalism, a 'radical' and a 'moderate' version. On the former view, making up

18 Or imagine you wanted to answer the mind-directed question of whether you believe a career change is a good idea and someone suggests that you do so by answering the corresponding self-referential world-directed question "Does a change of careers seem like a good idea to you?". This is not going to help.
your mind is always sufficient for acquiring self-knowledge, whereas on the latter view, it’s only sufficient in ‘normal circumstances’.

I have suggested that radical rationalism faces what I’ve referred to as the ‘garbage in, garbage out’ objection, according to which deliberating on the basis of bad or suspect reasons has self-deception rather than self-knowledge as its outcome. I then concentrated on a possible rationalist response to this objection, which seems available to the moderate rationalist, which is to filter out the garbage, i.e. bad circumstances of deliberation.

I argued that the ‘filtering strategy’ requires specifying what the right circumstances or states of mind are in which making up one’s mind is a way of constituting one’s attitudes and allows the subject to acquire knowledge of them. I argued that the filtering strategy is not a satisfying response to the ‘garbage in, garbage out’ objection, because there are no ‘right’ circumstances or states of mind that the rationalist can appeal to, at least without ruling out self-deception from the outset by saying that the ‘right’ circumstances are just those in which making up one’s mind is self-knowledge conducive, which would make rationalism a rather uninformative, even question-begging, view.

In the final section, I briefly considered what changes we would need to make to rationalism in order for it to be construed along holist lines. I have tried to show that it’s not evident that rationalism can even be construed along holist lines, again without begging the question. A natural next move would be for the rationalist to ‘go radical’ and avoid appealing to the right or normal circumstances of deliberation. This is the view I will consider in the next chapter.
6

The Limits of Activism

1. Introduction

In the previous chapter, I discussed the moderate atomist version of rationalism. I described this as a view according to which it is possible for a subject to judge that P without believing that P, and according to which judging that P is only sufficient for self-knowledge in the 'right' circumstances, namely those circumstances in which one is entitled to assume that one's judgement actually constitutes one's belief. I then argued that this theory is problematic because there is no way of defining what the right or normal circumstances should be such that they lead to self-knowledge rather than self-ignorance or self-deception that does not beg the question. However, I also mentioned that the rationalist may instead refrain from appealing to right or normal circumstances altogether and opt not for a moderate but for a radical version of rationalism. One could avoid appealing to normal circumstances if one thought that judging that P entails, or is sufficient for, believing that P.

The aim of this chapter is to critically evaluate this version of rationalism. I will explore, specifically, Matthew Boyle's account (2009a, 2009b, 2011a, 2011b), according to whom (1) judging that P entails believing that P and (2) belief and knowledge of belief are one and the same state. I will contrast and compare Boyle's so-called 'Activist' account of what beliefs are, and how we should understand the relation between judging that and believing that P, with alternative 'Dispositionalist' accounts thereof.¹ For, as Bertrand Russell wrote, "Psychology, theory of knowledge and metaphysics revolve about belief, and on the view we take of belief our philosophical outlook largely depends" (Russell 1921, Lecture xii). The more

¹ I borrow the label 'Activism' from Cassam (2014, 112-119-153), though his use of the term is somewhat different than mine.
general aim of this chapter, therefore, is to show that when we want to answer or indeed properly understand the question of what is required for a subject to know her own attitudes, we first of all need to be clear about the question of what we take ‘attitudes’ to be, in order to evaluate different proposals.

The structure of the chapter is as follows. In the next two sections, I begin by describing Boyle’s account of self-knowledge (Section 2) and his account of belief (Section 3). I then describe two versions of the ‘Dispositionalist’ view that Boyle argues against (Section 4) and discuss Boyle’s overall worries and what I take to be his principal objection, which is that Dispositionalists cannot respect the intuitively correct constraint “I believe Q because I believe P” (Section 5). I respond to Boyle’s principal objection and claim that the objection fails to convince (Section 6). This, then, leads to the question central to Section 7: who has the better ‘metaphysical cards’, the Activist or the Dispositionalist? I propose to answer this question by considering a number of examples in which a subject seems to judge that P without believing that P, and so an apparent gap between the two emerges. I suggest that there are two ways in which the Activist may respond to such cases, one of which involves saying that these subjects did not genuinely (really, truly) judge that P, which I argue is implausible (7.2). The other is a more ‘flat-footed’ response, which would involve denying that the examples show a gap between judgement and belief at all. I argue that this response is problematic because it forces the Activist to decouple action-related features of belief. However, in Section 9, I critically reflect on the common strategy of assessing metaphysical theories of belief on the basis of considering our intuitions about paradigm cases and claim that the metaphysical dispute between Activism on the one hand and Dispositionalism on the other is not so easily settled after all. It’s not clear which intuitions are being ‘pumped’, or indeed what intuitions are, where they come from and why they matter. I conclude that when addressing the question of what is required in order for someone to know her own attitudes, we not only need to be clear about the metaphysical question of what we take attitudes to be but also need to reflect on meta-theoretical questions such as how we should go about adjudicating between different theories of judgement and belief in the first place.

---

2 I borrow this term from Wright (1989a, 177).
2. Boyle’s epistemic view: reflectivism

Boyle defends what he calls a ‘reflective’ approach to transparency. His central claim is that we should not think of transparency as involving a subject making an inference from ‘P’ to ‘I believe that P’ but instead should think of the subject as taking a different sort of step, from believing P to reflectively judging (i.e. consciously thinking to himself): I believe P. The step, in other words, will not be an inferential transition between contents, but a coming to explicit acknowledgment of a condition of which one is already tacitly aware. (Boyle 2011b, 226 emphases in original)

On the reflectivist view, transparency involves “shifting one’s attention from the world with which one is engaged to one’s engagement with it—an engagement of which one was already tacitly cognizant even when one’s attention was ‘directed outward’” (2011b, 228). So believing that P and knowing that you believe that P are not distinct states, but different aspects of one and the same state: “in the normal and basic case, believing P and knowing oneself to believe P are not two cognitive states; they are two aspects of one cognitive state—the state, as we might put it, of knowingly believing P” (2011b, 228). The transparency procedure is not an explanation of how we get self-knowledge, because there is no ‘procedure' by which we acquire self-knowledge. Self-knowledge is something we somehow get for free: it's just built in to believing that P – if you really believe that P, then you know that you believe that P.

Boyle’s view is thus best described (somewhat paradoxically, perhaps) as a metaphysical account of self-knowledge, or that he is doing epistemology by doing metaphysics.3 He seeks to explain self-knowledge in terms of “the nature of belief itself” (2011b, 228). It is the metaphysical nature of the very states themselves, on Boyle’s account, which “implies that their subject has tacit knowledge of them” (2011b, 235). All that's needed for self-knowledge of one's (judgement-sensitive) mental states is to reflect and articulate what one already knew at a tacit level. In other words, second-order beliefs are built in to the notion of what mental states are: one cannot be in a mental state without knowing that one is (if only tacitly). Hence,

---

3 This is Boyle’s own description. He writes, for instance, that reflectivists offer an account “that is primarily metaphysical rather than epistemological” (Boyle 2011b, 235).
Boyle's account can be understood as a version of constitutivism regarding self-knowledge (cf. Shoemaker 1994; Coliva 2009; Bilgrami 2012) according to which self-knowledge is not a cognitive achievement but comes down to a conceptual truth that becomes apparent by reflecting on what rationality is.

Boyle's account of self-knowledge raises many interesting questions, such as what 'knowingly believing' should be, that is, why and how first-order attitudes come attached to second-order beliefs that the subject tacitly knows about. These questions concern Boyle's theory of self-knowledge. In what follows, however, I want to focus on a different question, namely on the metaphysical account of belief that underlies Boyle's account, in particular on how Boyle understands the relation between judgements and beliefs.

The guiding assumption of the discussion that follows, an assumption that I take Boyle to have drawn attention to, is that if we want to assess the respective strengths of different theories of self-knowledge, the question of which theory is to be preferred cannot be answered without considering the underlying metaphysical view of what beliefs and attitudes are. The metaphysics of mind, however, is not something that gets much explicit attention in the debate. Moran, for instance, has remained largely silent on the underlying metaphysical questions, and hence, what's particularly praiseworthy about Boyle is that at least he has put his cards on the table. So I shall bracket, for the sake of the argument, the question of whether if you believe that P, you know that you do. I don't think this constitutivist assumption is very plausible; indeed it strikes me as wrong. There seem to be good reasons for thinking that I can have an attitude but don't believe that I do or to believe that I have some attitude without actually having that attitude. However, it might well be the case that the fact that the constitutivist assumption strikes me as wrong is simply because I have a different metaphysical view of what attitudes are. This, at least, is what I would imagine Boyle would say. Therefore it is crucial to be clear about the different metaphysical options about belief and other attitudes that are available. If Boyle's metaphysical view of what beliefs are is true, then maybe the reflectivist idea becomes more plausible, i.e. that believing something and knowing that one believes it aren't two distinct states.

4 The terminology here is admittedly somewhat confusing because Boyle's epistemic account described above is typically referred to as a metaphysical account. To clarify, we might say that we can either focus on the attitude as had by a subject or once the attitude is 'in place'—and evaluate Boyle's reflectivist proposal—or on the prior question of what is required for a subject to have an attitude, which is what I will be centrally concerned with.

5 Boyle is of course not the only exception. Other views, such as Schwitzgebel's, will be discussed and contrasted with Boyle's account below.
3. Boyle's metaphysical view: Activism

A more or less standard view in philosophy of mind is to think of judgements as conscious or occurrent mental activities, as something like the inner analogue of assertions, and of beliefs as standing or dispositional states. Consider Shoemaker:

[B]elieving something – having the standing belief that so and so is the case – is not an act. Judging, thought of as a mental occurrence rather than a standing state, is an act. (Shoemaker 2009, 36; cited in Boyle 2009a, 120)

Or Cassam:

It will save time and help to prevent various kind of misunderstanding if I make a few things clear at the outset ... The attitudes I'm talking about are “standing” rather than “occurrent”. Standing attitudes remain in existence when you are asleep; they aren’t mental events like judging or deciding. It’s controversial whether a belief can ever be occurrent but when I talk about belief I’m talking about beliefs understood as standing states. Ditto for desires, hopes, and so on. (Cassam 2014, 138)

Because of the metaphysical difference between judgements and beliefs, there is room for a potential gap to emerge: judging that P without having the belief that P. It is this gap, I argued in the previous section, that requires the rationalist to recognize that one might follow the transparency procedure or make up one’s mind by coming to judge that P, but that does not constitute the person’s attitudes, and so require the rationalist to invoke ceteris paribus conditions.

Boyle, however, has defended an alternative account of the relation between judgements and beliefs that he suggests underlies rationalist accounts like Moran’s. On the ‘Activist’ version of rationalism that Boyle defends, there can be no gap between judging and believing that P, and so having attitudinal self-knowledge requires no extra conditions (apart from being conceptually competent and being capable of practical reasoning). The appeal to the ‘right’ circumstances is only

---

6 Recall from the previous chapter that I take ‘following the transparency procedure’ to be equivalent to (1) ‘making up one’s mind’, (2) ‘answering the world-directed question’ and (3) ‘coming to a judgement by reflecting on the reasons in favour of the subject matter’.
necessary if we think that it's possible for there to be a gap between one's judgements and one's beliefs in the first place. If there can be no gap between one's judgements (one's conclusions on the reflection on the reasons in favour of P) and one's beliefs, then the arguments from the previous chapter do not work, because any of one's judgements constitute one's beliefs, since judging that P entails believing that P, whatever reasons one's judgement was based upon. Boyle thus argues against the orthodox view that judgements are occurrent states and beliefs are standing states. Boyle’s account is somewhat complex, so I will spend some time trying to explain his view. His central claim is that believing something is an exercise of one’s agency, not a standing state. To believe something is not a passive affair, but an active one. Hence, Boyle talks about a person’s capacity to ‘actively believe’.

Boyle’s metaphysical account of what beliefs are can be explained by returning briefly to some Anscombean themes discussed earlier (Chapter 5). Recall that on Anscombe’s view, the special sort of knowledge the subject has is a sort of knowledge she has ‘in action’. The special sort of knowledge I have of my intentions is something I have while watering the flowers or while making coffee. The relevant Anscombean why-question is “Why are you doing A?” and not “Why did you do A?”. Answering the latter question, after all, arguably involves making inferences from memory, whereas the former does not, or so the Anscombean story runs. But it’s only while I am doing something—acting—that I have the relevant sort of self-knowledge, i.e. that I can know without ‘self-observation’ that I am doing something, and why. Boyle explicitly stresses this point when explaining his view:

Something that is striking in Anscombe – something that sets her apart from many subsequent action theorists – is her resolute focus on action in progress. Anscombe’s “why?”-question is first and foremost “Why are you doing A?” ... Her assumption, in effect, is that to understand the nature of intentional action, we must describe the specific character of the subject’s relation to it as it unfolds. (Boyle 2009a, 138–39)

If we want to apply the Anscombean model not just to intentions but also beliefs, hopes, desires, and so on, this raises a challenge, because the relevant ‘Anscombean’ version of the why-question applied to beliefs would seem to be “Why are you believing that P?” which, intuitively, seems rather different from the more passive question “Why did you believe that P” or “Why do you believe that P?”. It makes sense
that I am now drinking coffee, or watering the plants, and that I have a peculiar sort of self-knowledge because the action unfolds or is in progress. But this does not straightforwardly apply to attitudes like beliefs, for, ordinarily, we don't think of beliefs as things that 'unfold' at all.

I think we have two options here. The first is that it needs to be recognized that the application of the Anscombian view to attitudes other than intention does not have a straightforward application, despite what many who defend some account of 'transparent' self-knowledge have suggested (Hieronymi 2008; Hamilton 2000, 2008). An explanation needs to be provided of why an account that focuses on a person's present doings can be applied to a person's attitudes that, at least prima facie, do not involve actions 'in progress'. Alternatively, one might challenge the orthodox metaphysical view of belief in philosophy and claim that we do need to think of beliefs in the active sense, as things that unfold, just as Anscombe suggests is true of intentions, and thus to opt for a more straightforward application of Anscombe's account.

As I read Boyle, this is his project. He recognizes that the Anscombian view of thinking about (the knowledge we have of our) intentional actions is "controversial", but adds that his purpose is "not to defend Anscombe's view in its own right, but to draw a comparison between her understanding of our relation to our own actions and what we have seen about our relation to our own beliefs" (Boyle 2009a, 140), and thus Boyle sees the "relation to our own beliefs [as] structurally comparable to our relation to our intentional actions" (Boyle 2009a, 144).

How does Boyle propose to achieve his purpose? According to the traditional view, believing something is not itself an activity, because 'belief' is a standing state. Boyle disagrees, and suggests that we should think of 'belief' along the following lines:

As in the case of action, so too in the case of belief, a rational subject normally stands in an active affirmative relation to her own present beliefs. Her persisting belief that P is grounded in her assent to P as meeting the measure that a proposition must meet to merit belief, namely truth. This assent is not an act that precedes her belief and produces it; the very existence of her belief that P is constituted by her persisting assent to P. Her believing P, we might say, just is her enduring act of holding P true, and hence to-be-believed. This act is not occurrent — it need not involve any bustle or commotion ... [The] relationship
between her believing and her sense of what is reasonable is brought to the forefront of her attention when she occurrently considers whether she accepts P and what grounds she has for doing so, but it is present—actually, not merely potentially—even when she does not occurrently reflect. Her holding the belief she does is itself an enduring act of her power to assent to whatever proposition she deems reasonable. (Boyle 2009a, 144)

Boyle’s suggestion is thus that belief is a persisting act, and that in an important sense it’s correct to say that I am believing it is cold outside. A belief is “an enduring, non-occurrent act of assenting to a proposition” (Boyle 2009a, 143), and so it ‘persists’ or ‘endures’, and thus believing something appears to have some sort of temporal dimension. However, Boyle later seems to have changed his view somewhat, claiming that “the primary form of agency we exercise over our beliefs in deliberation is not an agency exercised over time” (2011a, 3). Boyle’s suggestion seems to be that thinking that it has a temporal dimension is to make some sort of mistake relating to category. Believing that Amsterdam is the capital of the Netherlands is not something that’s either short lived or long lived. It seems to be a-temporal. I do not discuss these exegetical issues further, because they don’t matter much to understanding Boyle’s overall account.

His central claim, I take it, is that judgements and beliefs are very much alike: judging that P entails believing that P. Boyle’s view thus takes the idea of ‘maker’s knowledge’ quite literally: I come to know that I believe that P by making it the case that I believe that P, and I do this by simply judging that P. Boyle’s account is distinctively Kantian in the sense that the capacity for judgement is absolutely central to believing anything and to knowing one’s beliefs, as well as the idea that there is a fundamental difference between knowing our own judgements and beliefs about what we should believe or do on the one hand and knowing our own sensations and appetites, etc. on the other.7 Interestingly, on Boyle’s Kantian metaphysical picture

---

7 Boyle states: “A way of putting the thesis ... is to say that our immediate, authoritative knowledge of our own judgments is a necessary byproduct of our ability to reason about what is the case and what to do. It seems clear, though, that our knowledge of our own sensations and appetites is not in this sense maker’s knowledge. However we explain our privileged knowledge of our own sensations and appetites, we should acknowledge that these are states that come to pass with us, not states we arrive at through deliberation. And this sounds strikingly like what Kant says: that whereas our apperceptive knowledge of our own judgments is a knowledge of ‘what we do [thun],’ our knowledge of our sensations and appetites through inner sense is a knowledge of what we ‘undergo [leiden]’” (Boyle 2009b, 158).
of what beliefs are, it seems we don’t need to appeal to normal circumstances of making up our minds, because there could be no cases in which judging that P would not entail believing that P. The metaphysical nature of belief rules out that there could be a gap between one’s beliefs and one’s judgements.

According to Activism, the relation between judgements on the one hand and beliefs on the other is tighter than tight. However, the claim that the relation between judgement and belief is close is not what is distinctive of Activism, because those who think of beliefs as standing states can also think that this is so. For instance, by claiming that judging that P typically leads to, or causes, the belief that P. What is distinctive about Activism, rather, is a specific conception of why the relation is close. According to the view Boyle rejects, judgements and beliefs can, in principle, come apart, even though typically they come together. For an Activist, though, judgements and beliefs necessarily come together. The relation is constitutively ‘tight’.

But what, exactly, is the positive account that Boyle puts in place of the view that beliefs are standing states? Towards the end of his more recent paper, Boyle writes that for a rational subject to believe something “is for him to have his power to be persuaded by reasons actualized in a present and persisting act—where an act in this sense is not a species of event or process, but an act of an altogether different type” (Boyle 2011a, 22). In explaining the activity of making up one’s mind, Boyle refers to Aristotle’s notion of *energeia*, translated as “activity” or “actuality”. Aristotle’s notion of *energeia*, Boyle writes,

is an actualization of a capacity “in which the end is present”: one whose existence does not consist in the unfolding of a process proceeding towards a certain result, but rather in a mode of active *being*, every moment of whose existence constitutes a moment of the completion of this activity. (Boyle 2011a, 20)

holding a belief might itself be an “energetic” act of rational self-determination ... The relevant agency is at work not primarily in the installation or modification of beliefs, but in the kind of believing characteristic of rational creatures, as such. This believing is self-determined, not in virtue of some precedent process or event, but by being the special kind of self-affirmed condition that it is. (Boyle 2011a, 23)
It's not entirely clear what we should make of this. But whatever the Aristotelean appeal comes down to, it seems clear what the view is that Boyle rejects, namely the view according to which beliefs aren't acts. This idea must be false, Boyle says, because it contradicts our nature as rational animals who have the capacity to judge-and-thereby-believe or who have the capacity to, as he calls it, 'actively believe'.

In the next section, I discuss the view that Boyle argues against, which is the view according to which the existence of a potential gap between judgements and beliefs is possible, even if it's rare.

4. Radical versus moderate Dispositionalism

We can crudely distinguish between two traditional, 'non-Activist' views of how we should think about the relation between judgement and belief: radical Dispositionalism and moderate Dispositionalism. Radical Dispositionalists think

There's a somewhat controversial issue that I want to flag here regarding the question of whether there are such things as occurrent propositional attitudes, or that one might 'consciously' believe something, and that it's possible to believe something for a very brief moment. Even though many philosophers seem to think that there are such things as occurrent or momentary beliefs, this view is not uncontroversial. According to Tim Crane, for instance, though "there is such a thing as being conscious of one's belief, that does not mean that there is such a thing as consciously believing. 'Occurrent belief' is a myth" (Crane 2001, 108). Cassam shares Crane's view, noting, "When I judge that P I do not occurrently believe that P because there is no such thing as occurrently believing" (Cassam 2011, 16). Though I am tempted to think that occurrent beliefs don't exist (I agree with Dennett 1981, 303) that "when such judgments are called occurrent or episodic beliefs, this is a serious misnomer", I doubt that discussions regarding occurrent beliefs amount to more than a terminological dispute, given that it seems to come down to the question of whether someone's mental state is a very 'long' judgement or a 'short' belief. The question, though, is whether Activism is only meant to apply to such occurrent states if they did exist. For instance, in a recent paper, Antonia Peacocke (2016) has defended a constitutivist view similar to Boyle's, but one that explicitly appeals to the notion of occurrent beliefs. She claims, "Actively recognizing the truth of p—that is, judging that p—must involve at the very least having a momentary belief that p" (A. Peacocke 2016 emphases mine). Crucially, Peacocke only thinks her claim holds for such momentary beliefs: "I do not ... endorse the claim that judgement at some time t is sufficient for belief at any other time t', or for any interval of time T" (ibid.). This gives the impression that the sort of view Boyle and Peacocke defend only applies to these momentary states, which means Activism would have a very limited application, and this in turn suggests that the epistemic account of self-knowledge is limited in this way, too. However, while this might be the case for Peacocke, I don't think it is true of Boyle's account, for Boyle does not seem to think that there can be such things as occurrent beliefs in the first place, given that he thinks that beliefs cannot be had at a time or over time (at least in his later paper). And so it seems plausible that Boyle's constitutivist view has a more ambitious scope, namely our "judgement-sensitive' attitudes such as belief, desire and intention" (Boyle 2011b, 223).

Schwitzgebel (2010, 2013) offers a more encompassing list of views. Making finer-grained distinctions becomes especially important once we consider complex cases of whether someone believes that P and shifts to believing not-P, or should be described as having contradictory beliefs or as having in-between beliefs, etc. For present purposes, however, three views – Activism and two versions of Dispositionalism – should do.
that judgements aren’t necessary for belief; moderate Dispositionalists think that they are necessary but not sufficient. Both are incompatible with the Activist view, according to which judgements are necessary and sufficient for belief. I discuss the two varieties of Dispositionalism below.

A radical Dispositionalist thinks that propositional attitudes like beliefs or desires can be explained exhaustively in Dispositionalist terms: to believe that $P$ just is for one to be disposed to behave in the relevant way(s), where the disposition to judge or assert that $P$ is not taken to be particularly special. For radical Dispositionalists, judging that $P$ is something that might, but need not, happen. Someone can be said to believe that $P$ without judging that $P$. One might take Donald Davidson’s and Daniel Dennett’s views as radically Dispositionalist in this sense. On the radical Dispositionalist view, beliefs and other propositional attitudes are considered, primarily, as explanatory entities: things by which another’s actions, or line of thinking, is made intelligible or predictable. Beliefs and desires are identified “by the sorts of actions they are prone to cause, given the right conditions” (Davidson 2001, 216). Dennett (1987) thinks that having a propositional attitude simply means being treated as having it:

It is not that we attribute (or should attribute) beliefs and desires only to things in which we find internal representations, but rather that when we discover some object for which the intentional strategy works, we endeavor to interpret some of its internal states or processes as internal representations. What makes some internal feature of a thing a representation could only be its role in regulating the behavior of an intentional system. (Dennett 1987, 32)

On this approach, someone who behaves according to $P$’s being true, believes that $P$ is true. On the radical Dispositionalist approach, it’s not only possible to believe that $P$ without judging that $P$; it’s also possible to believe that $P$ while judging that not-$P$. Christina Borgoni, for instance, thinks that there can be what she calls ‘resistant beliefs’: “A resistant belief is a recalcitrant cognition that persists in an individual’s psychology despite the individual’s epistemic reasons against the belief” (Borgoni 2015, 212; see also Cassam 2014, e.g. 107-111). This would happen when a person judges his or her belief to be false but is unable to stop believing it. Borgoni gives the example of Emelia, who was “raised in a sexist community, where from infancy she heard her family and friends claiming that women were unfit for politics”. This resulted in the fact that Emilia eventually forms the belief that men and women are...
not equally competent in politics. Later in life, however, Emilia comes to judge that men and women are equally competent in politics, but notices that “she is sexist in most of her unguarded, instinctive and automatic behavior when it comes to assessing female performance in politics” (Borgoni 2015, 212). But despite Emilia’s deliberative efforts and willingness to change her behavioural responses, she does not succeed. According to Borgoni, Emilia has a resistant belief: she believes that not-P while failing to judge accordingly.

One might wonder why one would not simply say that Emilia believes, later in life, that men and women are equally competent in politics rather than ‘merely judges’ that this is so? This depends on what one thinks beliefs are, and the point here is precisely to emphasize the fact that this is not how radical Dispositionalists would understand such a case, for radical Dispositionalists think that judgements aren’t (always) necessary for belief. (I come back to the sort of example Borgoni discusses below.)

By contrast, both moderate Dispositionalists and Activists think that taking P to be true, or judging P, is necessary in order to believe that P. One of the principal arguments against the radical Dispositional view is given by Moran. In an early paper, Moran (1994) takes issue specifically with Dennett’s and Davidson’s accounts and argues that their accounts miss out on a fundamental dimension of belief and approach the question of what beliefs someone has purely from an outsider’s perspective, overlooking the first-person perspective of the believer (Moran 1994; see also Brandom’s critique of Dennett’s view in Brandom 1998, ss-62).

10 According to Moran, radical Dispositionalism only explains one aspect of belief— namely, its ‘theoretical’, third-personal dimension. The limits of this approach become evident, Moran suggests, once we adopt the perspective not of the interpreter but of the interprettee. When you consider your beliefs from a first-person rather than a third-person perspective, you do not ask the theoretical question “Do I believe that P?”, where you consider your own beliefs merely as dispositions to behave in certain ways and things that have explanatory merits, but rather the deliberative question “Is P true?”. This second question isn’t settled by facts about you (how you behave, etc.) but facts about P. The agent herself cannot answer the question of how she is disposed to act, or which attitudes she in fact has, before having settled the question of whether P is true, whether Q is desirable or whether R is worth pursuing, and so on, or so Moran argues (1994, 170). His main point is that it is not irrelevant for the agent herself whether her belief is true or false (Moran 1994, 169). If the subject thinks that her own belief that horse A is most likely to win the race is false, then she cannot continue to believe that horse A is most likely to win. The asymmetry here with the third person is obvious. I can attribute to A the belief that horse A will win, while believing (or even knowing!) this belief to be false, without any apparent contradiction or conflict. For the interpreter it is possible to attribute a belief even if it is false. This is why Moore’s paradox of saying “P, but I don’t believe it” only applies in the first-person case. Moran’s idea, then, is this: if, in the first-person case, one brackets the question of the truth of one’s belief, one would not be a believer. More positively, taking-as-true is a condition of having beliefs at all. Moran writes that the first-person stance is “not eliminable” from a theory of what it means to believe something. We cannot, he says “take the purely third-person project of explaining and predicting the behaviour of others to define and exhaust the meaning of psychological terms”, as Dennett and Davidson appear to do (Moran 1994, 170).
from a different angle, Schwitzgebel refers to what I have referred to as radical Dispositionalism as the ‘anti-judgement view’ and says that this view omits what the subject explicitly endorses, how she is disposed to judge the overall state of affairs all things considered, what side she would take in an argument, how she is disposed to reason about the case in reflective moments, her best conscious assessment of the evidence. (Schwitzgebel 2010, 542)

Both moderate Dispositionalists and Activists disagree with a view like Davidson’s, Dennett’s or Borgoni’s, because they think that to believe or want something is not just to be disposed to do X but must also involve a person’s taking something to be true; must involve her reasons to do or believe that P. The latter is something that cannot be explained simply from the interpreter’s stance, but crucially involves considering the stance of the interprettee, i.e. the believer herself.11

The difference between moderate Dispositionalism and (Boyle’s) Activism is that moderate Dispositionalists do not think that judging that P is sufficient for believing that P. Nicholas Silins, for instance, claims that “[j]udging that p is insufficient for believing that p, I take it, because believing that p requires having various dispositions, where judging that p is insufficient for having those dispositions” (Silins 2012). On the moderate Dispositionalist view, then, one does not necessarily know what one’s beliefs are if one knows one’s judgements. In the end, the issue is between radical atomism and moderate atomism; between those who think that making up one’s mind by judging that P entails or implies believing that P and that judging that P necessarily leads to self-knowledge and those who think that making up one’s mind by judging that P entails or implies believing that P, and thus that judging that P potentially leads to self-knowledge.

11 Curiously, Moran and others seem to have failed to take note of Dennett’s distinction between ‘opinions’ and ‘beliefs,’ where opinions are described by Dennett as being remarkably similar to what Moran refers to as beliefs. Dennett explicitly says that his intentional stance theory only applies to beliefs, not opinions. In How to Change your Mind, Dennett writes, for instance, that “making up your mind is coming to have an opinion” (Dennett 1981, 307). Consider also this passage: “all cases of making up or changing one’s mind is that changes of mind are a species of judgment, and while such judgments arise from beliefs and are ultimately to be explained by one’s beliefs, such judgments themselves are not beliefs – when such judgments are called occurrent or episodic beliefs, this is a serious misnomer – but acts, and these acts initiate states that are also not states of belief, but of something rather like commitment, rather like ownership. I trust it sounds at least faintly paradoxical to claim that when I change my mind or make up my mind, the result is not a new belief at all, but this is just what I want to maintain” (Dennett 1981, 303). Maybe, then, Dennett is a radical Dispositionalist about beliefs, but not opinions.
I should note that my description of moderate Dispositionalism is deliberately broad. It includes the view defended by, for example, Schwitzgebel, but it can also include Moran's view, depending on whether or not Moran in the end thinks judgements entail beliefs or rather thinks that it is only in the right circumstances that judging that P constitutes the belief that P, and so a 'gap' between the two is possible. Needless to say, there will be deep differences between various moderate Dispositionalist positions. We might perhaps think of moderate Dispositionalism as a spectrum, ranging from those who are optimistic and think that judging that P normally leads to—but does not entail—the belief that P to those who think that believing that P requires a lot more than being disposed to judge that P, such as Schwitzgebel:

On one approach to belief, the sincere endorsement of [certain propositions] is sufficient, or nearly sufficient, to qualify as believing them, if you really do wholeheartedly judge these propositions to be true when you reflect upon them. On an alternative approach—the one I will defend—it is not enough to sincerely embrace such propositions in reflective moments. To qualify as someone who genuinely believes such things, you must live that way. (Schwitzgebel unpublished manuscript)

Moderate Dispositionalists think of the relation between judging that P and believing that P as one where believing itself is not, pace Boyle, an 'activity'. Instead, as Boyle suggests, on the moderate Dispositionalist view, believing that P appears to be the product of our judgements. Boyle goes on to describe the view as one that sees the relation between judgement and belief (or some other attitude) as a process of some sort, and describes the view thus:

12 Notice that moderate Dispositionalism can be construed either atomistically or holistically, depending on how we understand the way in which judgements 'produce' beliefs. The moderate Dispositionalist might, for instance, appeal to the idea that judgements normally or typically produce beliefs, or do so when all else is equal, etc. Christopher Peacocke seems to adopt a view along these lines: "I said that for some ways of coming to make knowledgeable self-ascriptions, the nature of belief and judgement is part of the explanation of their correctness. It is not the full explanation, and my exposition was peppered with occurrences of the qualifying phrase 'when all is working properly'. Someone can make a judgement, and for good reasons, but it (may) not have the effects that judgement normally do—in particular, it may not result in a stored belief which has the proper influence on other judgements and on action. A combination of prejudice and self-deception, amongst many other possibilities, can produce this state of affairs" (C. Peacocke 1998). Peacocke is one of few authors to make explicit the implicit appeal to these ceteris paribus clauses, which suggests that, in my terminology, the sort of view Peacocke is defending is not of a radical atomist sort. Peacocke does not appear to want to claim that judgements constitute beliefs in any circumstances.
If we exercise agential control over our own beliefs, [Dispositionalists] maintain, this must consist in our performing occurrent acts of judgment that give rise to new beliefs, or cause extant beliefs to be modified. (Boyle 2011a, 4)

Deliberation whether P is a process that culminates, if things go well, in a judgement on the truth of P. Judgement is an occurrent act by which a subject installs a new belief in herself, or modifies one she already holds. Belief itself is not an act but a state. (Boyle 2011a, 5)

Boyle’s description seems applicable to representationalist or cognitivist approaches to the mental or those who think of beliefs and other propositional attitudes as things ‘stored’ in the mind or stored in Belief or Desire ‘boxes’, caused by the subject’s judgements (when the new belief gets ‘installed’) and which in its turn causes certain dispositions to behave in certain ways. I am sceptical, however, about the idea that Dispositionalists must think of attitudes along these lines (see e.g. McGeer and Schwitzgebel 2006; McGeer 2007b; Schwitzgebel 2013; Slors 2012). But Boyle’s more general suggestion that the moderate Dispositionalist appears to think of beliefs as products rather than acts seems plausible. This leaves the question of what Boyle thinks is wrong with this view. This question is important because if Boyle’s arguments against moderate Dispositionalism are successful, then we may have reason to accept the view according to which judging that P implies believing that P, which in turn paves the way towards Boyle’s preferred account of self-knowledge described at the beginning of this chapter.

To briefly recap, in this section I have described the metaphysical view that Boyle’s Activist account is meant to challenge and offer an alternative to. I have described the radical Dispositionalist view as the view according to which a subject believes something if she can be suitably interpreted as believing that P, whether or not she judges that P, whereas a moderate Dispositionalist further claims that the question of whether a person believes something requires her to have capacity to judge, i.e. to answer world-directed questions such as whether P is true, X is desirable or Y is to be feared. Activists, lastly, think that the latter is all there is to believing something.

13 Since so few theorists are explicit about what they think judgements are and whether and how they differ from belief, and what their relation is, it’s hard to give any references here. Boyle describes what he calls the Process Theory, and with those who adopt this view he includes Shah and Velleman (2005), Christopher Peacocke (1998) and Cassam (2010, 2011).
The reason for discussing these different metaphysical views is that doing so allows us to be able to evaluate Boyle's metaphysical proposal by comparing it to alternative conceptions of what beliefs and other attitudes are. This allows us to evaluate Boyle's approach to attitudinal self-knowledge, given that Boyle himself rightly stresses that the question of self-knowledge is intimately connected to the question of how we should understand the objects we have knowledge of.

The next step is to consider why Boyle thinks that the view according to which judgements lead to or cause beliefs, and so a view on which a gap is conceivable, is false, which is the question I address in the following section.

5. The Activist argument against moderate Dispositionalism

5.1 Prima facie doubts against moderate Dispositionalism

Why should we accept Activism over moderate Dispositionalism? Before turning to what I take to be Boyle’s principal argument, I first want to briefly review some of the other criticisms he has voiced.

Boyle writes, for instance, that the moderate Dispositionalist, or what he calls the ‘process theorist’, holds that “judging is an act whereas believing itself is not”, so “is forced to represent our actually believing that P as at most a product or result of our agency”. As a result, this approach, he thinks, “leaves our agency standing in a too extrinsic relation to the condition of belief itself” (Boyle 2011a, 6). The view “distorts our understanding of the basic sense in which we are capable of doxastic self-determination and gives rise to difficulties about the very rationality of this activity” (Boyle 2011a, 7).

I agree with Boyle that, if one is an Activist, this is a problem. But one wonders why this should be a problem for the moderate Dispositionalist (hereafter dispositionalist). For only if you think that beliefs are acts, in the strong sense that Boyle does, does it seem to follow that Dispositionalism distorts our understanding of the basic sense in which we are capable of doxastic self-determination. And so it isn’t obvious that Activists, or at least Boyle, have a real argument against those who adopt moderate Dispositionalism rather than simply presenting a different metaphysical paradigm and rejecting the one that lies at the basis of Dispositionalism. It seems we need something more robust in order to embrace Boyle’s metaphysical and, ultimately, his epistemic account.
Boyle also discusses a paper by Shah and Velleman (2005), which he takes to reveal a problematic approach to what ‘believing’ involves. Shah and Velleman discuss the distinction between what one already believes and what one now believes, and argue that the transparency procedure only applies in ‘making up your mind’ cases, i.e. cases in which you don’t already have a belief about P, but in which you still have to decide what to think. Boyle points out that Shah and Velleman’s account is motivated by a “basic feature” of the Dispositionalist view, namely its assumption that we actually exercise our capacity for doxastic self-determination only on those occasions when we now reflectively make a judgement about whether P, whereas when we merely hold a belief without reflection, we are not presently exercising this capacity. On closer examination, however, I think this assumption should seem suspect. (Boyle 2011a, 10).

Shah and Velleman do not recognize “the right sort of connection between our capacity for doxastic self-determination and our presently believing what we do” (2011a, 11). But, again, I don’t think we need to abandon Dispositionalism on this basis. What we need is not that the Activist finds it suspect, but that Shah and Velleman, too, would have to recognize that the assumption is suspect. But Boyle gives no reasons to think that those who think judgements cause beliefs would have to think that the assumption is suspect, other than giving a description of his preferred Activist metaphysics, which, again, is presumably one that Shah and Velleman would reject.14 And so we seem to have made no progress.

Boyle concedes he mostly offers mere “prima facie doubts” (2011a, 11). I am not sure that what Boyle offers are prima facie doubts about Dispositionalism rather than prima facie confidence in Activism. But, in any case, Boyle comes with a “more direct objection” (2011a, 11) that I describe below (5.2) and to which I then respond (Section 6). I argue that Boyle’s objection fails to convince and that this means that we are left with the question of whether the Activist metaphysics ought to be preferred over (a version of) moderate Dispositionalism, which I go on to address in Section 7.15

14 In all fairness, Boyle does go on to level his principal objection against Shah and Velleman, which I turn to below. But Boyle’s rhetoric may be taken to suggest that more is going on, argumentatively, which I doubt is the case.

15 I should say that the discussion of Boyle’s objection, as well as my reply to it, keeps very close to Boyle’s text and is in that respect a rather internal discussion. Some readers may therefore want to proceed directly to Section 7.
5.2 The principal objection: meeting ‘Boyle’s Constraint’

Boyle’s central objection boils down to the idea that Dispositionalists are unable to respect the following constraint (C): “I believe Q because I believe P”. Boyle does not give concrete examples, but we might imagine the constraint to be applicable to ‘making up your mind’ beliefs and to beliefs you already have, given that Boyle stresses that both beliefs are ‘active’ in his sense. So let’s consider an example of each, in order to get a better understanding of Boyle’s objection:

(1) I believe that the streets are wet (Q) because I believe that it’s raining (P)

(2) I believe that the UK is headed for a recession (Q) because I believe that Brexit is going to happen sometime in the future (P).

In these sorts of cases, I believe that Q because I believe P. The questions then are, how would a dispositionalist approach these two cases and, more particularly, in what sense would s/he fail to respect Boyle’s Constraint (C)?

Boyle’s reconstruction of the Dispositionalist view is as follows: “I believe that P, and that if P then Q, and on this basis I judge that Q at time t. A cause must precede its effect, so if my judging Q is the cause of my believing Q, then I come to believe Q only after t” (2011a, 12). It may be helpful to try to unpack this a bit and say that, on the Dispositionalist interpretation of what’s going on in cases like (1) and (2), at some point in time, (t1), I judged that P, e.g. that Brexit is going to happen. This judgement gave rise to, or caused, at t1+ the belief that P. Assuming that if I believe that P, and further believe that “if P then Q”, then I come to judge, at t2, that the UK is headed for a recession. This judgement causes, at t2+, the corresponding belief that Q. Assuming a cause must precede its effect, I come to believe that Q only after t2.

---

16 Cf. “one’s capacity for doxastic self-determination plays a role even in one’s knowledge of one’s extant beliefs” (Boyle 2011a, 10).

17 I am not sure whether my example is a plausible example of a case of ‘making up one’s mind’, but the reader might imagine his/her own preferred example.
(t1) judgement that $P \rightarrow (t_1+) \text{ belief that } P$

if $P$ then $Q$

(t2) judgement that $Q \rightarrow (t_2+) \text{ belief that } Q$

If I understand Boyle correctly, his worry is that the Dispositionalist must say that my belief that $Q$ at $t_2+$ can only be explained by a belief that was caused earlier, namely at $t_1+$. According to Boyle, “The relevant psychological causes are, it seems, all in the past”, and so it’s hard to see how the Dispositionalist can deliver the “intuitively correct explanatory claim expressed in (C)” (Boyle 2011a, 12). What we want to say in these examples, or so I understand Boyle, is that I believe that $Q$ because I believe (present tense) that $P$. Boyle goes on to say that the Dispositionalist, who thinks of the relation between judgement and belief in terms of a process, must say that my belief that $Q$ can only be explained by my having believed that $P$ at $t_1+$, and so fails to respect (C): that a subject is in a position to know he believes that $Q$ because he believes (present tense) that $P$.

What should we make of this?

6. Response to Boyle’s Constraint

Boyle’s discussion seems to me to be confused. This is because Boyle fails to distinguish or equivocates between a psychological and an epistemic understanding of the notion “because” that’s central to his constraint (C). If so, then the “intuitively correct explanatory claim expressed in (C)” (Boyle 2011a, 12) is unclear, because it is ambiguous.

In Chapter 4, I suggested that when we are talking about the causes of one's belief, how one 'arrives at' one's belief or how one's belief is formed, as well as talking of (temporal) structures, we are considering the psychological features (histories, pathways) of a person's state. Epistemic considerations, on the other hand, have to do with the state's 'credentials', and concern the question of how appealing to one's belief

---

18 Boyle also discusses a non-causal variant of PT “on which my judging that $Q$ does not cause my believing that $Q$ but is itself an event of forming or acquiring the belief that $Q$” (Boyle 2011a, 14). The causal variant, however, seems to me much more common, so I will stick to that version here, especially because ultimately Boyle’s objection to the non-causal version, with which I agree, is “that the meaning of the proposal is unclear”.

that P can make it appropriate for one to believe that Q, or what one's grounds or bases are for believing what one does (Pryor 2005; Cassam 2015a).

With this distinction in mind, we can turn to the question of how we must understand Boyle's Constraint (C) "I believe Q because I believe P". Should we take the "because" to be an 'epistemic because' or a 'psychological because'?¹⁹ I think Boyle leaves the answer to this question unanswered (perhaps deliberately so - I turn to this option in a moment). What I want to do, therefore, is consider a specific passage from Boyle in which psychological and epistemic features seem to run together. The epistemic claims are in bold, whereas the psychological claims are underscored (notice some concepts/phrases are both bold and underscored):

On the causal variant of [the process theory], the objection is straightforward. Suppose I believe that P, and that if P then Q, and on this basis I judge that Q at time t. A cause must precede its effect, so if my judging Q is the cause of my believing Q, then I come to believe Q only after t. What explains my then believing Q? The relevant psychological causes are, it seems, all in the past: the proximate cause is my judging Q at t, and the more remote causes are the beliefs I held at t that gave rise to this judgement. Given these assumptions, it is difficult to see what basis there can be for the intuitively correct explanatory claim expressed in (C). In what sense can I be said to believe Q because I believe P? My belief that P may indeed persist, and it may be true that if it were changed, this would bring about a change in my belief that Q. But it seems that the only actual explanatory connection whose existence is entailed by the fact that I have reasoned "P, so Q" ... is a relation between my believing that Q and my having believed that P at t. How then could reasoning in this way put the subject in a position to know that he believes Q because he believes P? (Boyle 2011a, 12)

Three phrases, which seem to be the crucial phrases, are both bold and underlined, i.e. are ambiguous between epistemic and psychological readings ("What explains my then believing Q?", the "explanatory claim expressed in (C)" and "explanatory connection"). More generally, it seems that notions like 'explanation', 'because' or

¹⁹ These two questions are arguably still too coarse-grained, especially since the notion of 'psychological', even if we go with Pryor's description of it, is still ambiguous because 'psychology' can refer to a myriad of things. If so, then matters may be worse regarding how we're supposed to understand Boyle's Constraint (C), though I will have to ignore these complexities here.
‘why’ are ambiguous in this way. This would mean there are potentially two different ways of understanding (C). Which is the one that Dispositionalism is supposed to fail to respect?

Suppose we take (C) to be an epistemic constraint, one that would make appealing to one’s belief that P appropriate for one to believe that Q. On the epistemic understanding of (C), the constraint is not “about what psychological processes you’ve undergone” nor is it about whether you did or did not “arrive at P by deriving or inferring it from other beliefs” (cf. Pryor 2005). When Boyle asks “why” someone believes that P (2011a, 8) or “what explains my then believing Q” (2011a, 12), on an epistemic reading, he is not after a psychological explanation, but rather an explanation of the basis of one’s belief, or one’s justification for believing what one does.

But if (C) is an epistemic claim, which seems to be the most plausible option, then I see no direct reason why Dispositionalists do not or could not ‘respect’ (C). Boyle describes, after all, what he calls the process theory, and what I have referred to as Dispositionalism, as a theory about how we arrive at our beliefs, i.e. how they are formed. It’s a theory that’s mostly concerned with the psychology, not the epistemology, of attitudes. Dispositionalism is the view according to which judgements normally or typically lead to or cause beliefs. On Boyle’s reconstruction (which I’ve been assuming is true), “Deliberation whether P is a process that culminates, if things go well, in a judgement on the truth of P” where such a judgement is “an occurrent act by which a subject installs a new belief in herself, or modifies one she already holds. Belief itself is not an act but a state” (Boyle 2011a, 5). 20

Clearly, then, Dispositionalism is a psychological account of attitudes. Understood as such, I see no reason why Dispositionalists could not be neutral on (C) taken as an epistemic constraint. Why should the fact that my belief that P was caused earlier on, as the Dispositionalist claims (or so I’m assuming, as Boyle does), commit the Dispositionalist to deny the epistemic claim “that I believe Q because I believe P”, i.e. to deny that my belief that Q fails to respect (C)? Epistemic-based relations are a-temporal, so it should be irrelevant how the belief that Q comes about or how one arrives at one’s belief. The Dispositionalist can say “I believe that Q because I believe that P”, while claiming that, as a psychological matter of fact, “I believe that Q because I believed that P”, i.e. because there was an earlier moment at which some new belief was ‘installed’, or something similar.

---

20 Again, I don’t think (all) Dispositionalists have to accept this description (‘installing’, etc.), but I will ignore this here.
The above assumes that (C) is an epistemic constraint, not a psychological one. This seems to me to be the most natural reading of (C). But of course it's possible that (C) is a psychological constraint after all, i.e. a constraint on how we (ought to) arrive at our beliefs. But taken as a psychological criterion, I don't see why the Dispositionalist is in trouble, either. It's not clear why it would be incoherent for the Dispositionalist to hold, simultaneously, the following claims: (1) judgements cause beliefs and (2) causes precede their effects, and so agree with (3) the belief that Q came after the judgement that Q (as well as after the belief that P) but to insist that (4) I still believe that P, so long as I still have the relevant dispositions (am still disposed to judge that P, and so on). And so it seems to me that the Dispositionalist can deny Boyle's suggestion that the Dispositionalist would be forced to say that I believe that Q because of "my having believed that P".

And even if we were to assume for the sake of the argument that the Dispositionalist would have to say "I believe Q because I believed that P" (or because I judged that Q, or something similar), the Dispositionalist might say that this explanation of one's belief that Q is only implausible or absurd if, without realizing, we took the notions of 'explanation' or 'because' to have an epistemic meaning. But this would mean that we would, again, be equivocating between psychological and epistemic understandings of the notions of 'explanation' and 'because', etc., which I have just suggested we had best avoid doing.

There might be an even more radical response available to the Dispositionalist. S/he might say that, taken as a psychological criterion, it's not really clear what it means to 'respect' (C). Suppose I tell you that I believe that the UK is headed for a recession and you ask me 'why' I believe this, or ask me to 'explain' my belief. If this is a psychological why-question, then what am I supposed to say, exactly? My thought here is that if we want to give psychological explanations of people's mental states, it's not clear that what the subject herself will say in response to the question of why she believes a particular thing (again, taken as a psychological question) is going to be a good strategy. Maybe the right answer is one that appeals to my virtuous or vicious epistemic character (maybe I believe that Q 'because' of closed-mindedness (cf. Zagzebski 1996; Cassam 2016, 2017)). Or maybe I was primed, nudged or manipulated to have the belief that Q, in which case, the confabulatory reasons I give to answer the why-question probably won't square with the available psychological explanation. Maybe the story is even more complex, and requires considering theoretical psychology, cognitive science, and so on. If we have to believe
Daniel Kahneman (2013) and others (Tversky and Kahneman 1974; Thagard and Nisbett 1983; Ariely 2010) who have suggested that we're (predictably) irrational, then it's not obvious that when I answer the psychological why-question by saying “because P”, it is accurate, taken as a psychological description of how I arrived at the belief that Q. This more radical response would bring us back to the idea that, most plausibly, Boyle’s Constraint probably isn’t a psychological constraint but an epistemic one.

There’s one final possibility, which is that Boyle’s Constraint (C) is both an epistemic and a psychological constraint. But how would that work? This would mean that as a psychological matter of fact, my belief that Q really did come from my belief that P. If that is the case, the psychological question of how I, as a matter of fact, came to believe that Q, and the epistemic question of what my grounds are for believing that Q have the same answer: what grounds my belief that Q is also how I arrived at my belief that Q, namely my belief that P.

But why would we think epistemic and psychological explanations would ‘conflate’ in this way? For (C) to be both an epistemic and a psychological constraint, we must take what the subject says about the causes or aetiology of her belief that Q as giving us the truth about the epistemic grounds of her belief. I find this hard to believe. More fundamentally, though, given Boyle’s overall project, it seems much more plausible to take his constraint to be an epistemic constraint and the relevant why-question to be a question about the subject’s credentials, grounds or bases for believing that Q:

The relevant why-question does not inquire into the explanation of his coming, at some past time, to hold the belief in question, except insofar as the subject’s knowledge of how he came to hold the belief speaks to the reasonableness of his continuing to hold it now. Our interest is not in his psychological history, but in the present basis of his conviction. (Boyle 2011a, 10–11)

But if it’s true that (C) is neither a psychological constraint nor a psychological-cum-epistemic constraint but a distinctively epistemic constraint, as I think the above passage suggests, then Boyle’s argument against the Dispositionalist fails to convince.

Where does all of this leave us? If the Activist’s principal objection against Dispositionalism does not convince, then what is left in support of Activism
is simply whether its metaphysical framework is more plausible than the Dispositionalist metaphysical framework. In order to assess the respective merits of Activism and Dispositionalism, I consider, in the next section, how the Activist deals with the examples of where there is an apparent gap between a person’s judgements and her beliefs. I suggest that unless Activism were to build in certain ceteris paribus clauses (in which case it ends up being a version of moderate atomist rationalism criticized in the previous chapter), the Activist way of dealing with these cases has counter-intuitive results, and this would speak in favour of adopting Dispositionalism instead. However, I also point out that arguably it leads to these counter-intuitive results only given certain implicit assumptions about what we should expect from a metaphysical theory of mind in the first place, and that we should not expect the Activism-Dispositionalism dispute to be settled in any easy way by considering ‘hard cases’.

7. The apparent gap between judgement and belief

In this section, I want to consider some examples (7.1) which illustrate that there is a gap between judgements and beliefs where, according to the Activist view, there couldn’t be one. I then discuss two ways in which the Activist might deal with these examples, which I refer to as the ‘no proper judgement’ response (7.1) and the ‘what gap?’ response (Section 8), and assess the implications of each.

7.1 The Implicit Racist, the Unethical Professor and the Family Man
Stuart Hampshire gives a good illustration of a situation in which one might ‘distance oneself’ from one’s passing (occurrent) thoughts:

The Fearful Flyer
In the airplane, the thought that it is going to crash comes to me, and perhaps I find that I cannot help thinking that it will crash. But I may dismiss the persisting thought as imagination, and not belief, since I am far from ready to commit myself to the thought that the plane will crash. I do not think, in a strong sense of “think”, which is equivalent to “believe,” that the plane will crash, and more than I do those things, in a strong sense of “do,” when I find myself unintentionally doing. (Hampshire 2015, 101–2; for discussion see Cassam 2011, 7; a similar example is discussed in Shah and Velleman 2005, 507)
Hampshire's example plausibly suggests there can be a gap between (passing) thoughts and one's beliefs. But judging that P, one might argue, is very different from having a mere passing thought that P. I might entertain the thought that pink elephants exist, without judging, i.e. 'actively' considering it to be true, that they exist. So it's not clear that the fearful flyer judges that the plane will crash (as Hampshire notes, "I am far from ready to commit myself to the thought that the plane will crash"). What we need, therefore, is a case in which one genuinely judges that P, while, intuitively, not believing that P.

In the previous chapter, I've already discussed a case that was meant to illustrate that one might judge that P without constituting one's attitudes accordingly, namely the Angry Spouse. To recall, the Angry Spouse made up her mind 'angrily'. She followed the transparency procedure, i.e. came to a conclusion regarding P (whether a divorce was a good idea) by considering the only (Anscombean) reasons that were available to her in the mood she was in, namely angry and self-deceptive reasons. If judging that P would have constituted her belief that P, she would have constituted the belief that a divorce was desirable. Given that this belief was strongly incompatible with most of her other attitudes and values, and was, as Schwitzgebel would say, strongly incompatible with how she lived, I suggested we must conclude that the Angry Spouse is an example of where judging that P was not constitutive of the belief that P.

I've argued that ruling out the sort of circumstances the Angry Spouse was in is not a plausible strategy and hence that rationalists cannot just ignore her or hide her in implicit ceteris paribus clauses. But let's assume my argument to that effect does not hold, and hence that we need an example of someone who made up her mind, or followed the transparency procedure, in less abnormal circumstances, that is, in circumstances where the person seems to be "in her right mind". Here, considering the phenomenon of implicit bias might help. Consider Schwitzgebel's example of Juliet:

Juliet the Implicit Racist

[Juliet] is prepared to argue coherently, sincerely, and vehemently for equality of intelligence and has argued the point repeatedly in the past. Her egalitarianism in this matter coheres with her overarching liberal stance, according to which the sexes too possess equal intelligence and racial and sexual discrimination are odious. And yet Juliet is systematically racist in most of her spontaneous reactions, her unguarded behavior, and her judgements about particular
cases. When she gazes out on class the first day of each term, she can’t help but think that some students look brighter than others—and to her, the black students never look bright. When a black student makes an insightful comment or submits an excellent essay, she feels more surprise than she would were a white or Asian student to do so, even though her black students make insightful comments and submit excellent essays at the same rate as do the others. This bias affects her grading and the way she guides class discussion. She is similarly biased against black non-students. When Juliet is on the hiring committee for a new office manager, it won’t seem to her that the black applicants are the most intellectually capable, even if they are; or if she does become convinced of the intelligence of a black applicant, it will have taken more evidence than if the applicant had been white. When she converses with a custodian or cashier, she expects less wit if the person is black. And so on. (Schwitzgebel 2010, 532)

Christopher Peacocke gives a similar example of what appears to be a case of judging that P without believing that P:

The Unethical Professor
Someone may judge that undergraduate degrees from countries other than their own are of an equal standard to her own, and excellent reasons may be operative in her assertions to that effect. All the same, it may be quite clear, in decisions she makes on hiring, or in making recommendations, that she does not really have this belief at all. (C. Peacocke 1998, 90)\textsuperscript{21}

Finally, consider the following example, again from Schwitzgebel:

The Family Man
I say I value family over work. When I stop to consider it, it seems to me vastly more important to be a good father than to craft a few more essays like this one. Yet I’m off to work early, I come home late. I take family vacations and my mind is wandering in the philosopher’s ether. I’m more elated by my rising prestige than by my son’s successes in school. My wife rightly scolds me: Do I really believe that family is more important? (Schwitzgebel 2012)

\textsuperscript{21} Saying that she does not believe that P doesn’t necessarily mean that she believes the converse (that not-P).
I take it to be a key strength of the Dispositionalist view that it is able to handle complex cases like those discussed above in a way that recognizes that it's possible to make up your mind and come to judge that P while failing to believe that P, and hence that there is what Schwitzgebel (2010) refers to as a 'gulf' between judgements and beliefs. As Cassam remarks, it seems that in the above examples “the belief that P fails to stick despite the acknowledged presence of good reasons in favour of P, reasons which lead one to judge that P” (Cassam 2011, 15). This, in turn, suggests that knowing what you believe is not the same as knowing what you judge: even if you know what you judge, you might fail to know what you believe, because judging that P does not entail believing that P. So knowing what you believe requires further (interpretative or inferential) work to be done. Your judging that P can be evidence of your belief that P — very good evidence, but fallible evidence all the same.

But notice that the intuition that there is such a gulf only emerges on a broadly dispositionalist metaphysical view of belief. We should therefore expect the Activist to think about these examples very differently and to be able to respond to the above suggestion that only the Dispositionalist is able to make proper sense of the sort of cases discussed above. I discuss two responses below.

7.2 The ‘no proper judgement’ response
When considering the apparent gap between judgement and belief that appears to occur in the examples of Juliet the Implicit Racist, The Unethical Professor and the Family Man, the Activist can choose between either of two strategies. One option is what I’ll call the ‘no proper judgement’ response, which is to say that none of these subjects genuinely judged that P, which would explain why they did not end up genuinely believing that P, either. The other is the ‘what gap?’ response, i.e. the denial that in these examples there is a gap between judgement and belief, which I discuss in the subsequent section. There is, of course, a third strategy, which would be to say that the subjects really did judge that P but failed to believe that P, because judgements do not always lead to, constitute or cause beliefs. But this would be tantamount to giving up Activism, and will mean that the arguments against the moderate atomist version from the previous chapter will re-emerge, so I will be assuming this is not an option.

The ‘no proper judgement’ response involves the claim that the subjects merely seemed to judge that P; that is, they merely seemed to take P to be true. This means that Juliet has not truly endorsed the proposition that all races are intellectually equal;
the Family Man has not truly ‘staked himself’\(^2\) when it comes to valuing family over work; and the Unethical Professor has not really asserted that undergraduate degrees from countries other than her own are of an equal standard to her own.

I suspect this is not a particularly appealing road for the Activist to go down. The ‘no proper judgement’ response is problematic, first of all, because we seem to have no good reason to think that Juliet and the others are not judging that P. Nothing seems out of order – we have no reason to think that Juliet or any of the others are, for example, emotional or hesitant or are not fully rational or self-conscious or something similar. And so this would mean that any account of ‘judgement proper’ is going to have to be ad hoc. One begins to suspect that, at the end of the day, it just so happens that only rationalists will be able to tell us what judgements really are. This should make us suspicious.

Also, if Juliet and the other characters haven’t made a judgement, this would make the notion of judgement a technical one, which would be fine if it weren’t for the fact that it seems that rationalists want it not to be a technical one. On closer inspection, judging that P is no longer equivalent to what we would normally take it to be, e.g. taking P to be true and endorsing P, etc. (unless the subjects were not taking P to be true either, but then it would be a mystery what taking something to be true would be). If that’s the case, then that means that the question of whether someone has (really) judged that P is not something that can be answered from the subject’s first-person perspective. The question of whether or not someone has ‘really’ judged that P is something she herself might not have the answer to. This would be bad news for the rationalist, I take it, given that the first-person perspective is so central to the rationalist view – and rightly so.

All in all, I think the ‘no proper judgement’ response is a no-go. This leaves the ‘what gap?’ response, i.e. the denial of the fact that in these examples there really is a gap.

\(^{22}\) Cf. Moran: “At some point, I must cease attempting to infer from some occurrence to my belief; and instead stake myself, and relate to my mental life not as something of symptomatic value, but as my current commitment to how things are out there.” (2001, 150 emphasis in original).
8. The ‘what gap?’ response

If judgements entail beliefs, and if Juliet has really judged that P, which I think we must accept if we want to avoid begging the question, then the Activist must conclude that Juliet simply cannot fail to believe that P. The Activist says that judging that P entails believing that P, and so Activism is committed to construing certain cases, for example Juliet’s case, in this way. She does not ‘sort of’ believe this, or ‘in-between’ believe it, but believes it determinately. This amounts to saying that, assuming the subject is minimally self-conscious and is able to deliberate on the basis of reasons, no appeal to circumstances is needed. Judging that P entails believing that P even when one is angry, depressed, in love, nervous, hungry, hypnotized, nudged, and so on. On the assumption that judging is something one does knowingly, it would seem to follow that one comes to know one’s attitudes in these circumstances.

My main worry boils down to the fact that this has the implication that the ‘what gap?’ response wiggles loose the action-related features of a belief, and, more generally, that Activism ultimately cannot avoid decoupling action-related dimensions from what is involved in having attitudes. This will need some unpacking.

The basic intuition appealed to by the Dispositionalist is that whether or not someone has a certain attitude and can be said to know that she does is something that depends on how the subject is inclined to act in the future and how she has acted in the past. Believing, wanting or hoping for something is not just something that happens ‘in the head’. As Krista Lawlor puts it, someone’s “having a special prerogative to speak about his or her own states of mind” derives from “reliable correlations between what one is inclined to say and what one is inclined to do, between the attitudes one sincerely asserts and the attitudes one acts upon” (Lawlor 2003, 549, 560). An Activist, on the other hand, will have to insist that such reliable connections regarding the subject’s future (or past) conduct are not a necessary aspect of believing, wanting, hoping or intending. One can believe something, and know that one does, without ‘acting like it’.

---

23 Lawlor’s argument centres around a number of psychological studies (esp. Seligman, Fazio, and Zanna 1980; T. D. Wilson, Hodges, and LaFleur 1995), but I think her general argumentative strategy, and responses to her argument by Ferrero and Bortolotti, which I turn to below, can be discussed in abstraction from these studies.
The latter is a conclusion that has been explicitly accepted recently by Luca Ferrero (2003) and Lisa Bortolotti (2009), who defend what seems to be an Activist conception of what beliefs and other attitudes are. Ferrero writes, “The distinctive first-person authority of [someone’s] self-ascriptions concerns ... whether she takes responsibility for them, not whether the self-ascribed attitude is both correct and a reliable guide to future conduct” (2003, 570). So someone has self-knowledge of her beliefs even if it isn’t a reliable guide to her future conduct. Bortolotti claims that subjects can have knowledge of their attitudes “no matter how representative of their future behaviour those attitudes would [turn] out to be” (Bortolotti 2009, 639). And Ferrero claims, as indeed I think the Activist must, that the sort of diachronic, action-related considerations that are central to Lawlor’s critique are “only a regulative ideal of deliberation, not a necessary condition of it” (Ferrero 2003, 575; see also Bortolotti 2009).

We can, I think, imagine Boyle giving a similar sort of response to the examples of Juliet, the Unethical Professor and the Family Man. All of them have followed the relevant rationalist self-knowledge procedure, that is, all have judged that P and have come up with reasons for P, and they have done so successfully. Hence, they all came to believe that P, and acquired self-knowledge in the process.

This is quite a revisionist proposal if we stop and think about it. If reliability, diachronicity or stability is really merely a “regulative ideal”, then this means that the concept of belief (desire, etc.) does not in and of itself involve any claims about the subject’s future-directed actions. Activists may want to take issue with the crude behaviourist account of what counts as having ‘dispositions to act’ and point out that inner, deliberative actions, such as judgings, decodings, reasonings, calculatings, acceptings and attendings to something (cf. C. Peacocke 2009) are all actions too, albeit of a ‘mental’ (inner) rather than behavioural (outer) kind, and I would not disagree (cf. de Bruin, Jongepier, and Strijbos 2015; but see also G. Strawson 2003, who disagrees). It seems essential to the belief that ‘2+2=4’ that you are able to judge, reason or infer that adding another 2 equals 6, that the answer times two equals 8 and to think that ‘+’ means addition, and so on. These mental actions are part of the ‘future conduct’ that your self-ascription of the belief that 2+2=4 should be a reliable guide of. If you’re not prepared to ‘mentally behave’ in the relevant ways, then we have a good reason to think that you don’t really have the belief that 2+2=4, and so we must conclude that you couldn’t know that this is what you believe. The question is whether beliefs other than such simple beliefs like
2 + 2 = 4 can be treated in the same way, i.e. whether it’s plausible think that in order to know one’s beliefs it’s sufficient to know one’s ‘mental’ future dispositions. To say that this is all there is to what it means to have beliefs generally (whatever their content), let alone attitudes generally, is hard to accept.

I should stress that the Activist is certainly not committed to saying that the subject’s not acting in accordance with the belief that P is irrelevant all things considered, or indeed that someone like, for example, Juliet should not be held responsible for ‘not practising what she preaches’. The point is, rather, the fact that acting in accordance with the belief that P is, on the Activist view, irrelevant to the metaphysical question of whether someone believes that P. This has a curious implication, namely that the Activist will have to treat cases as similar that we would normally regard to be very different, in terms of their beliefs. Suppose the Family Man is married to the Family Wife. Both say and judge that they value family over work. But, whereas the Family Man goes off to work early and comes home late, and his mind is wandering in the philosopher’s ether, the Family Wife’s mind is at home, wandering about in thoughts about her son. Apart from the fact that the Activist, seeing the wider action-related dimensions of a self-ascription as a “regulative ideal”, must say that the Family Man genuinely believes (rather than ‘sort of’ believes, or wishfully believes or something similar) that family is more important than work, the Activist is forced to treat the Family Woman the same in terms of what she believes.

If this is hard to accept, then the Dispositionalist view, according to which we need a more nuanced account of what it is to believe something, one that would allow a metaphysical difference between the two, seems more plausible. As Jane Heal puts it, in such cases we might say that the self-ascription “retains a kind of shadowy credence” and is a case in which we are “inclined to say that the person ‘sort of’ believes or intends as she insists she does” (Heal 2002, 1–2). But the ‘sort of’ must be recognized as having important metaphysical ramifications, which, it seems, only the Dispositionalist can account for. Someone’s acting contrary to the judgement that P, we must conclude if we were to adopt the Activist view, is irrelevant to the metaphysical question of what someone believes. The Activist thus appears to decouple the action-related features of an attitude and consider it non-essential. This, I think, provides strong reasons against the Activist metaphysics, and, in turn, against the Activist epistemology of what is required in order to know our attitudes.
Another worry with the Activist metaphysics of belief is that it makes believing something, and knowing what one believes, too easy. On the Activist view, as Schwitzgebel writes, it is “easier than it should be to regard ourselves as free of racist, sexist, elitist, and other objectionable attitudes” (2010, 546). A second line of argument against Activism is that the Activist view has counter-intuitive results. The Dispositionalist might suggest, for instance, that what we should do is imagine that Juliet decides not to invite Jim, the only non-white candidate, for a job interview, and we ask her to explain her decision. Now suppose Juliet explains her decision by saying “I believe Jim was an unpromising candidate”. On the Activist view, Juliet is right – that is indeed what she believes. This, the Dispositionalist might say, flies in the face of our intuitions and/or immediate practical concerns. Maybe Juliet’s response is not as problematic as a theoretical position in metaphysics, but it’s rather counter-intuitive outside philosophy. If we further imagine that Juliet is one of our colleagues, we would certainly take offence at her ‘Activist’ response and if she told us that given that she’s got her judgements right, she’s got her beliefs right. The Dispositionalist might say this answer is highly unsatisfying. If we imagine that Activism is a position that actual people would actually adopt, this makes it evident that Activism must be false, or at least that moderate Dispositionalism is more plausible.

This raises a rather fundamental question, though. Is appealing to intuitions or practical concerns the best way to settle on one theory or another? The argument that Activism is false because it has counter-intuitive results appears to be premised on implicit assumptions about what we should expect from metaphysics of mind in the first place. In other words, the conclusion that Activism should be abandoned, and that we had better do so, appears to depend on certain meta-metaphysical and/or meta-philosophical assumptions. But it’s not unlikely that Activists think this is the wrong way to assess metaphysical accounts of mind. In the next and final section, I therefore turn to some of these more abstract issues which, I think, at the end of the day, may (co-)determine which metaphysics of mind should be accepted, which, in turn, largely determines which theory of self-knowledge ought to be preferred.
9. Meta-theoretical questions about metaphysics of mind: ‘intuitions’ and ‘practical concerns’

Earlier, I argued that the Activist’s argument against Dispositionalism (that it fails to respect Boyle’s Constraint) fails to be convincing (Section 6). The result of this was that, if indeed there are no good arguments against Dispositionalism, we needed to tackle the more fundamental question of whether Activism or Dispositionalism was the right view. This was the aim of the previous section, where I tried to show that the ‘what gap?’ response comes with two significant ‘costs’, one that concerns the fact that ‘believing’ something is decoupled from one’s future conduct, and another which means that Activism has counter-intuitive implications. However, the latter cost only emerged by appealing to intuitions and practical concerns and, more specifically, after considering a variety of everyday examples, and imagining that Activism would be a metaphysical view adopted by one of our colleagues in real life.

The Dispositionalist view appears, in this respect, to be a ‘pragmatic’ view of what beliefs are. Take Schwitzgebel, who has made explicit that his concerns are importantly pragmatic:

It’s not a passion for factually correct metaphysics that animates me here. Actually, I think there’s no such thing as factually correct metaphysics. There are just better and worse ways of conceptualizing the world, given our values and the empirical facts. I hope the interest of this essay doesn’t hinge on that controversial metaphilosophy; but you may not fully understand the argument now coming if you don’t see that it is intended as a pragmatic argument. ... The practical question is this: Do we want to highlight this empirical fact about ourselves – what I’d call the gulf between occurrent judgement and dispositional belief – or do we want to marginalize it as anomalous? (Schwitzgebel 2010, 546)24

24 In a recent manuscript, Schwitzgebel explains his notion of ‘pragmatic’ thus: “The approach I favor is pragmatist in two ways. First, it fits with the pragmatist tradition of Bain (1868/1973), Peirce (1877, 1878), James (1896/1912, 1907/2004), and Dewey (1920/1957, 1938) in emphasizing behavioral patterns as the core of belief. Second, it is metaposophically pragmatist in relying on pragmatic criteria to choose among competing metaphysical approaches” (unpublished manuscript).
A somewhat similar pragmatic appeal to the role of our intentions or ‘responses’ is also present in Lawlor’s work. Her strategy is to “[c]onsider our reactions” to certain psychological experiments, and she assesses the rationalist view relative to these reactions (Lawlor 2003, 558). She writes, “I believe that the findings under discussion ... cast light on our commonsense commitments regarding when self-reports are authoritative and when not” and that “[w]e react to the social psychological cases the way we do, precisely because we have this commitment” (Lawlor 2003, 560–61).

Arguably, this reveals an important methodological position – a general view of how we might assess philosophical theories. But the pragmatic view may not be a view that Activists would share. An Activist might think that we should not be led by our everyday intuitions when considering the question of what beliefs essentially are. By implication, the worries voiced in the previous section will fail to convince Activists, because for them, these worries beg important methodological questions.

There’s also another possibility, namely that the Activist also thinks that considering our intuitions matters, though I doubt the Activist would be happy to call them intuitions and would prefer to refer to them as our ‘practical concerns’. The Activist isn’t interested in offering a clinical or “factually correct” metaphysics, i.e. a sort of metaphysics that is decoupled from the sort of lives of (real) human beings. Instead, the Activist could plausibly say that s/he is precisely interested in (real) human beings, and real lives, but is concerned not with our immediate reaction to cases but rather with our practical self-understanding – the sort of understanding we have of ourselves (and others) in our capacity as beings who have the capacity to judge (see also the appendix to this chapter). What does it mean to understand oneself as an agent, as someone who has reasons to judge that P or act in this way or another? When faced with examples like Juliet or any of the other examples, these are the sort of questions the Activist will be interested in. Juliet has the capacity to judge, and this is fundamental to our reactions (concerns) to the case.

The Activist might say that it is Dispositionalism, not Activism, that fails to respect our intuitions (understood here as ‘practical concerns’). If a “pragmatic metaphysics requires that there not always be a single best way of classifying things, independent of our projects and interests” (Schwitzgebel unpublished manuscript), then we can see Activism, too, as offering a ‘pragmatist’ metaphysical account. For when it comes to intuitions/practical concerns, there’s actually quite a lot to say in defence of Activism. Imagine Juliet told Dylan the Dispositionalist that she believes
all human beings are equal or something similar. Suppose Dylan were to respond by saying “No you don’t”. This, the Activist might say, would also fly in the face of some of our ‘intuitions’ and important practical concerns. Dylan would be failing to treat her as a being who has the ability to form judgements and would thereby be disrespecting her.

A good, though rather sad, illustration of the role of intuitions about what it means to believe something and to know that one does, is the ‘Black Pete’ debate in the Netherlands. In the Netherlands, there is an annual folk tradition, involving St. Nicholas, aimed at young children, in which a tall, old white man is accompanied by his helpers: blacked-up men with bright red lips, curly wigs and golden earrings, referred to as ‘Black Petes’. A large part of the Dutch population does not want to change the tradition, but a growing number of people (and the UN) have urged the Netherlands to change the tradition because it is unethical – it reinforces negative stereotypes and impedes the struggle against racism. The annual Black Pete discussion has taken on rather grim proportions, though, and has resulted in a number of protests involving many arrests. People in the pro-Pete camp insist they have no racist beliefs, whereas those in the anti-camp claim that the people who are pro-Pete do have racist beliefs. (This issue isn’t nearly as simply or as clean-cut as I describe it, but this will have to do for now.)

The discussion bears striking parallels to the Activist/Dispositionalist dispute about what beliefs are and what is required to know one’s own beliefs. When pro-Petes deny that they have racist beliefs, they do so by appealing to something like the Activist conception of belief. They sincerely judge, after all, that racism is bad and ought to be dealt with; what more is needed to believe something? The anti-Pete camp, on the other hand, seems to appeal to a broadly Dispositionalist notion of belief: you can say whatever you like, and give reasons against racism as much as you want, but at the end of the day it’s a further question whether or not this amounts to having racist beliefs or not. The Black Pete tradition “accentuates the kind of implicit racism that

---

25 As Catarina Dutilh Novaes observes in a blog on NewAPPS: “it seems pretty obvious that the association between being black, fun-oriented, not very bright and being a servant is not likely to foster associations between being a black person and being a highly capable professional, for example. What effects will this have when these children grow up and become those in a position of power, e.g. in charge of hiring employees? There is the very palpable possibility that these associations established at an early age will influence negatively a person’s perception of black people for the rest of their lives – if not explicitly, at least implicitly. This is indeed the likely negative effect of the tradition that worries me most: it establishes problematic associations at a very young age, which thus become deeply entrenched in people’s minds. In other words, the Black Pete tradition accentuates the kind of implicit racism that we are all prone to, and as such is highly problematic – and racist” (Novaes 2013).
we are all prone to, and as such is highly problematic – and racist” (Novaes 2013).

A possible reconstruction of what is going on (and, I think, what is going wrong) is that individuals in the pro-camp are denied self-knowledge about their professed anti-racist beliefs, which in turn induces in them a feeling of being wronged. This, in turn, by the look of things, causes the pro-Petes to take up a rather aggressive stance with respect to those who deny them self-knowledge.

Now, when confronted with such scenarios, we might be tempted to ask “Who’s got it right?”. Arguably, Activists would have to say that the pro-Petes have got their metaphysics right. But Dispositionalists seem to be right to point out that this seems wrong, or in any case incomplete. Then again, Dispositionalists don’t give a satisfactory answer either, given that they must, it appears, side with the anti-Pete camp and deny anti-Petes’ knowledge of what they believe. In other words, Dispositionalists have a hard time securing the principle of first-person authority – that we are not prepared or entitled to simply correct, ignore or overrule people’s sincere self-ascriptions.

But maybe both views are wrong. I agree with Schwitzgebel that there is a lot to say in defence of a pragmatist account of belief and other attitudes, but it’s not obvious that this points (only) towards Dispositionalism, let alone towards a specific version of Dispositionalism, such as the sort that takes our “day-to-day choices and habits as more central than our judgements”. Defending a metaphysical theory “on practical grounds” (cf. Schwitzgebel) does not have an unequivocal meaning.

A third option is to deny that we can formulate a (more or less) determinate answer to the question of ‘who’s right’ – the Dispositionalist or the Activist. Considering our intuitions and practical concerns about certain cases is a complex matter, if only because the very notion of ‘intuition’ can mean a myriad of things, and because theorists don’t necessarily agree in this respect. As a result, it is not clear what conclusion(s) we can or should draw on the basis of certain cases. An account that could make room for both Activist and Dispositionalist concerns and that would deny that either of these views has the ultimate answer would arguably be a genuine liberal or pragmatic view. Maybe such a view is worth exploring – or maybe it isn’t. As Schwitzgebel says, “The human mind is a complex, fuzzy-bordered thing, right at the center of our values.” I absolutely agree. But I think this means that we should see both Dispositionalism and Activism as trying to articulate what these values are.

Where does this leave us? The Activist’s claim is that Juliet determinately believes all races are equal and has knowledge of her belief. This is implausible or absurd
only on a broadly Dispositionalist metaphysical view. Or, differently formulated, it's only implausible or absurd on the assumption that the metaphysics of belief/mind is something that is determined at least in part by a specific conception of our everyday intuitions and practical concerns, namely, as something that is principally concerned with people's habits and (unreflective) behaviour.

But now if we switch perspectives and try to see things from the Activist's perspective, then the Dispositionalist's epistemic claim that someone like Juliet does not know that she believes all races are equal is, in turn, absurd. After all, the Activist thinks that beliefs, not their effect, are exercises of rational agency, so how could Juliet fail to have the relevant belief? It's arguably even more absurd assuming, as we may imagine the Activist to do, that the metaphysics (of belief) is something that is not determined by our immediate and superficial intuitions about a number of cases. The Activist might not be happy with the notion of 'intuitions' at all, and, given the Activist's Kantian roots, s/he may instead think that what matters is Juliet's practical self-understanding: how she understands herself as a being who has the capacity to judge. It's such practical self-understanding, and its (moral) implications, that matters most, not our intuitions (unless we understand 'intuitions' in terms of the notion of practical self-understanding).

If it's true that Activists and Dispositionalists have different – incompatible – views, then the discussion about whether we should prefer one over the other is starting to look a lot like a stalemate, and this appears to be bad news. However, I think the discussion is at the same time fruitful in terms of what it teaches us about what we should expect from theories of self-knowledge, as well as how different theories should be evaluated. First of all, Boyle has shown that we cannot just assume or leave implicit our answers to the metaphysical question of what beliefs or other attitudes are when we are trying to answer the epistemic question of what is required in order to know them. The conclusion to draw is that the question of whether, for example, the Activist epistemic account of self-knowledge is true depends on one's prior (implicit or explicit) metaphysical account of what attitudes are. The latter account, may or may not depend on our 'intuitions' and/or 'practical concerns'. The content, relevance and the implications of what we consider out 'intuitions' and/or 'practical concerns' to be is likely to depend on one's meta-metaphysical views, such as whether "questions of metaphysics really have answers" (Manley 2009, 1) and whether these answers are "substantive or just a matter of how we use words" (ibid.) or what the best procedure is for arriving at these answers, for example an
appeal common sense, conceptual analysis, transcendental arguments. If an appeal to common sense is made, one might subsequently ask whose common sense we are trying to get at, that is, how we should determine what common sense even is. These are not impossible questions to answer, but they are important questions to consider when determining what mental states are.26

So, if it has achieved nothing else, I hope the discussion has shown where the attention should be in the philosophical debate on self-knowledge. It should not be on the different methods or procedures of self-knowledge but on (1) the metaphysical question of what attitudes/beliefs are and (2) the more fundamental meta-theoretical questions of what one takes the metaphysics relevant to self-knowledge to be and how different metaphysical views can and should be assessed, if this is possible at all.

Though there are some exceptions (Boyle and Schwitzgebel, for instance), metaphysics is not usually the topic of discussion in the self-knowledge debate, let alone the more abstract questions mentioned above. Hence, even if I have not been able to give sufficient arguments against Activism or for Dispositionalism, it would seem that the above conclusion should still be worth something, because it shows where the action should be.

10. Conclusion

I ended the previous chapter with a dilemma: either the moderate rationalist must ‘go holist’ and concede that appealing to so-called ‘normal circumstances’ of deliberation is not an option, or s/he must retreat to radical atomism and deny that circumstances can stand in the way of someone’s judgements constituting her attitudes. The main goal of this chapter was to consider a view that comes closest to denial of the relevance of circumstances, namely the view defended by Matthew Boyle, who claims that judging that P entails believing that P. In particular, the aim

---

26 One’s answers to meta-metaphysical questions (of how we should settle the question of whether we should prefer one metaphysics over another) may after all be influenced by one’s meta-philosophical commitments. If that is possible, then answering the metaphysical question of what attitudes are may in the end even involve addressing certain meta-philosophical questions, e.g. what philosophy is, whether it can be defined, what philosophy is for, how it should be done, and so on (Joll 2016). It’s certainly not meant to be obvious that one’s (implicit) answers to meta-philosophical questions are relevant to what one’s preferred theory of self-knowledge is, but it’s not ruled out that one’s preferred answers to such questions can play a role.
was to examine how an Activist understands the relation between (knowledge of) one's judgements on the one hand and (knowledge of) one's beliefs on the other and how this view contrasts with and compares to the view according to which the existence of a potential gap between judgements and belief is possible.

This exploration took the shape of assessing the relative merits of Activism and what I've referred to as moderate Dispositionalism, according to which even though believing that $P$ requires being disposed to judge that $P$, judging that $P$ does not necessarily amount to believing that $P$. I considered Boyle's main argument against moderate Dispositionalism, namely that it cannot satisfy the constraint that "I believe that $Q$ because I believe that $P$", and concluded that it fails to be convincing, because the constraint was ambiguous between a psychological and an epistemic reading of it. But on either reading it isn't obvious that the moderate Dispositionalist fails to be able to meet it, or that s/he ought to meet it at all.

Given that the moderate Dispositionalist view of the relation between judgements and beliefs did not appear to be false after all, this brought us to the question of what can be said in defence of Activism relative to moderate Dispositionalism. I attempted to answer this question by considering a number of examples in which an apparent gap exists between what someone judges and what she believes and asking what the Activist might say about them. I claimed that the 'no proper judgement' response was a no-go area, which left the flat-footed 'what gap?' response. By drawing on work by Lawlor and Schwitzgebel, I suggested that the 'what gap?' response leads to counter-intuitive results. However, I ended by pointing out that the question of whether our responses to or intuitions about the examples matter, and whether one even has these intuitions, is something that depends on much more fundamental issues, such as those having to do with one's general take on what 'metaphysics' (relevant to self-knowledge) is, as well as the more fundamental meta-philosophical question of how one should go about answering the metaphysical question.

One important conclusion to draw from all of this is that what I've called 'radical atomism' arguably cannot be dismissed as easily, without argument, as I suggested in the introduction, for its dismissal is dependent on certain meta-theoretical assumptions. Another important conclusion is that these more fundamental discussions are fundamental to understanding the question of how we know our own minds, as well as to assessing the theories that have been developed that have aimed to answer this question.

What's next? In this chapter and the previous one, I have been concerned with rationalism in so far as it offers a theory of how we come to know our own attitudes,
and have argued that it falls short in this regard, at least the moderate version thereof. In the next chapter, I want to consider (yet another) take on rationalism, which is to understand it principally as a normative project that seeks to explain not self-knowledge in general, but non-alienated self-knowledge in particular. This leads me to consider the question of what (not) being alienated from your attitudes involves, and to assess rationalism in this light.
APPENDIX:
A Dispositionalist Account of Self-Knowledge and an Activist Account of First-Person Authority

I have assumed that Activists and Dispositionalists are interested in the same question: what attitudes are and how we acquire knowledge of them. This is not just an assumption I have made; it is an important assumption of those whose work I have discussed. Ferrero and Bortolotti, for instance, respond directly to Lawlor's argument against the rationalist/Activist view, which suggests they agree on the explanandum/question. But it's possible that, despite appearances, Dispositionalists and Activists may have been talking at cross-purposes, and taking them as being interested in a different explanandum might be helpful, for it allows for a possible way of unifying Activism and Dispositionalism.

The idea would be as follows: Dispositionalists are discussing self-knowledge, whereas Activists are talking about first-person authority. First-person authority, as discussed in Chapter 2, is the idea that the subject is the 'authority' on what state she thinks (or says) she is in. Having first-person authority involves other people not being prepared, indeed not being entitled, to override, question or correct someone's sincere self-ascriptions. If this is how we understand the notion of first-person authority (cf. Bar-On 2004), then it seems the Activist has the best way of making sense of it: being granted first-person authority just is being considered to be a person with the capacity to judge. Therefore, to simply override, correct or ignore someone's sincere self-ascription might, on this line of thinking, amount to a specific form of disrespect, namely wronging someone in her capacity as a 'judger' and failing to recognize her as having the capacity to make up her mind.

It's true that, typically, first-person authority is explained in terms of self-knowledge (or even equated with self-knowledge): people's self-ascriptions are said to be authoritative because self-ascribers know their own attitudes, and so the two usually come together (see also Jongepier and Strijbos 2015). But this is not the only way of understanding the relation. One might instead say that one can have first-person authority without having self-knowledge. Even when someone's self-ascription is false, and even if we know it is false, we might still not be entitled to deny first-person authority to her. In other words, first-person authority and self-knowledge do not necessarily come together; the former may require a non-epistemic explanation. On this line of thinking one might, coherently, accept a
Dispositionalist account of what it means to believe (want, intend) something and insist that believing that P and knowing one believes that P requires more than judging that P, while adopting an Activist account of first-person authority that is grounded in people's capacity to judge. When someone says “I'm tired”, “I am going to go to bed early” or “I believe I treat my (fe)male students as intellectually equal”, such claims are authoritative – deserve our respect – whether or not they are true. Whether, on such a ‘unified’ account, either Activism or Dispositionalism is ‘more fundamental’ depends on whether self-knowledge or the capacity to judge is more fundamental.

Clearly, such a proposal is not without costs. The Dispositionalist will have to accept that his/her theory does not ipso facto amount to an explanation of first-person authority. This is a cost because many theorists of self-knowledge have explained and justified their interests in self-knowledge in terms of it – maybe rightly so. We would need some other reason why having self-knowledge matters (on this topic, see esp. Cassam 2014, chap. 15; Schwitzgebel, n.d.). The costs for the Activist are also clear: s/he hasn't given us an account of what s/he said she is giving an account of, namely how we know our own attitudes. But it's still possible for the Activist to argue that s/he has explained knowledge of our own conscious or ‘present beliefs’ (see also footnote 7). However, the question of whether there are such things as conscious beliefs is not uncontroversial, and also, if that is the case, the rationalist view seems to be that ‘by deliberating about some world-directed question and coming to a judgement, you thereby come to know your judgements’. This seems a rather trivial thesis to me.

Needless to say, this is just a very rough sketch of how elements of Activism and Dispositionalism may be accommodated, and much more needs to be said about it. The development of a more detailed account will, however, have to wait for another occasion.
7

Limits of Rationalism:
Failures of Autonomy

1. Introduction

In Chapter 5, I discussed the respective examples of the Angry Spouse and the Deferential Wife. The two were similar in so far as both followed the transparency procedure, i.e. made up their minds, and did so during a fit of anger, but they were also importantly dissimilar in so far as I described the Angry Wife as being autonomous, whereas the Deferential Wife, who believes that the proper role for a woman is to serve her family, was not. The question I want to address is how we should think about the capacity for making up one's mind in autonomy-undermining circumstances.

In my discussion of rationalism in the last couple of chapters, I have been concerned with rationalism as a theory of self-knowledge and have taken 'transparency' as the procedure for acquiring such knowledge. Thus understood, my conclusion was that the moderate version of rationalism is unable to give us a satisfying, non-arbitrary distinction between normal and abnormal circumstances of deliberation and that radical rationalism is only plausible on the (not uncontroversial) assumption that attitudes are not states, but activities that 'endure' or 'unfold'.

However, as Moran puts it in the introduction to Authority and Estrangement, his goal is to place “the more familiar epistemological questions in the context of wider self-other asymmetries which, when they receive attention at all, are normally discussed outside the context of the issues concerning self-knowledge” (Moran 2001, 1). Arguably, though, Moran’s critics, myself included, have ignored these “wider self-other asymmetries”. For Moran explicitly claims that the non-deliberative, i.e. ‘theoretical’, stance is also a route to self-knowledge. This means that making up your mind is not principally a claim about ‘how we acquire self-
knowledge'; rather, it’s a claim about how we acquire a specific sort of self-knowledge, namely ‘autonomous’ or ‘non-alienated’ self-knowledge.1 His goal is not simply to say something about self-knowledge but to say something about the nature of the first-person perspective and what it means for that perspective to be autonomous or non-alienated. Moran describes ‘transparency’ not as a procedure but as a “normative demand” or “a normative requirement” (Moran 2001, xvi–xvii).

Moran’s critics have hence overlooked the rationalist’s non-epistemic claims concerning these normative issues and focus almost exclusively on Moran’s epistemic claims, in particular whether and how making up your mind is a way of acquiring self-knowledge and whether such self-knowledge is immediate, as Moran claims it is (see Chapter 4).2 The rationalist might thus think that these critics have overlooked precisely what is essential about the rationalist view, namely the practical rather than the purely epistemic or psychological story it tells. This means that what appears to be central to the rationalist view may not have been systematically evaluated so far.

I should say, though, that rationalists, like Moran, are themselves partly to blame for this situation. Although Moran explicitly positions his theory of self-knowledge in the context of normative concepts such as freedom, autonomy and alienation, he is largely silent on the question of what deliberation is, and is mostly concerned with arguing against introspectionist and interpretationist views rather than against other theories of freedom, autonomy and alienation (except for Moran 2002).

The goal of this chapter is therefore to reconstruct and evaluate rationalism as a distinctly practical/normative project. I do so by centring my discussion around concrete examples of deliberation and by situating Moran’s account of rational agency in the theoretical landscape of theories of authenticity and autonomy. I address the question of what is needed for rationalism to be defended as a plausible practical/normative theory, and aim to show that if we take rationalism to be a normative theory of non-alienated self-knowledge, then, at present, it is underdeveloped. To show this, I compare the ‘rationalist’ take on alienation and autonomy to two general approaches towards autonomy, a ‘procedural’ and a ‘relational’ approach. If we understand rationalism along proceduralist lines, then it is susceptible to the same line of critique that proceduralist views are susceptible to, such as those put forward by relational and feminist theorists of autonomy. Or so I argue.

1 For the moment, I’ll use these notions (non-alienation vs autonomy) interchangeably – I return to this later.
2 A notable exception is Alec Hinshelwood (2013), who discusses Moran’s account in relation to the work of Harry Frankfurt and Michael Bratman.
The goal of this chapter is not so much to defend a specific account of autonomy (or rationalism) but rather to study the relation between self-knowledge and autonomy, which has not received much attention, at least not within the confines of the debate on self-knowledge (but see B. Roessler 2015; Christman 2005). Relational and feminist theories of personal autonomy are of particular interest because these focus specifically on the social and/or moral circumstances of agency and deliberation. Insights from discussions about autonomy provide a more sophisticated and realistic account of what it means to be an autonomous agent and what (some of) the preconditions are of ‘judgement’ that is relevant to rationalism, but, I want to suggest, considering these discussions in other areas of philosophy also teaches us something about (the limits of) atomism about self-knowledge.

The chapter is structured as follows. I start by explaining why Moran’s project is a normative one in Section 1, and suggest that some of its normative claims are often overlooked in the self-knowledge literature. This in turn raises the question of how the theory fares as a normative project, which the remainder of this chapter is concerned with. In Section 2, I introduce examples of making up one’s mind in contexts of oppression and brainwashing and address the question of what we should say about such cases on the basis of a rationalist account of self-knowledge. Can you make up your own mind if you have no mind of your own? In Section 3, I consider a possible way for the rationalist to say that oppressed subjects do not have non-alienated self-knowledge, and consider in more detail possible ways of understanding the elusive phrase ‘having a mind of one’s own’. In Section 4, I locate the rationalist account of ‘rational autonomy’ in the theoretical landscape of theories of personal autonomy more generally, and point out that on one interpretation, rationalism can be understood as a proceduralist account of autonomy. In Section 5, I turn to relational/feminist approaches to autonomy, which provide an alternative way of looking at the examples, and the conclusions we should draw from them. In the final two sections, I explore the implications of the relational framework for thinking about self-knowledge and observe that the considerations regarding what is necessary for an agent to be autonomous cannot be translated directly to the question of what is necessary for an agent to have self-knowledge (Section 7.1). Instead, I suggest (7.2) that we should not (just) distinguish, as Moran does, between different routes to self-knowledge, but that it may be helpful to talk about different types of self-knowledge – introspective, non-alienated and autonomous self-knowledge – and allow for a sort of pluralism that at present is absent in the self-knowledge debate.
In the final section, I take a step back and abstract away from the details of rationalism in order to reflect on what we might learn about atomism and holism in light of the discussion on what is required for reasons and attitudes to be ‘one’s own’ and what is required for a subject to know them.

2. The deliberative account of self-knowledge as a normative project

Let me begin with a brief recap. According to what Moran calls the ‘transparency condition’ of first-person statements of, for example, one’s beliefs, one should treat the question of one’s belief about \( P \) as equivalent to the question of the truth of \( P \). This means that one defers answering the self-directed question “Do I believe that \( P \)?” and instead answers the world-directed question “Is it the case that \( P \)?” This second, world-directed question is, Moran claims, a deliberative question, issued from the first-person. Answering world-directed questions in a transparent fashion requires that one take an active stance towards the subject matter by thinking about it and judging whether or not \( P \) is true, desirable, and so on. The basic idea is that by deliberating one comes to a judgement, and by making such a judgement, one constitutes one’s beliefs. To obey the transparency condition, then, one must (1) raise a world-directed question, (2) deliberate, (3) come to a judgement and thereby constitute one’s beliefs. One subsequently acquires knowledge of one’s beliefs on the assumption that one is entitled to assume that the inward-directed and the outward-directed questions are linked in the right way.

Critics of Moran’s view rightly claim that “his concerns are plainly epistemological” (Cassam 2015a) in so far as Moran’s account is after all a theory of self-knowledge. As such, it seems only fair that his rationalist view is assessed, and criticized, as a theory of self-knowledge. However, it’s also plain that Moran’s concerns are not exclusively epistemological. Many of Moran’s critics take transparency to be a ‘method’ or ‘procedure’ – just as I have done in the previous chapters. But Moran himself never talks about transparency in this way in Authority.

---

3 Cf. “I address myself to the question of my state of mind in a deliberative spirit, deciding and declaring myself on the matter, and not confront the question as a purely psychological one about the beliefs of someone who happens also to be me” (Moran 2001, 63).
and Estrangement, nor in more recent work. Instead, he talks about the ‘transparency condition’. Whereas a procedure or method is something one can or should follow, it may be argued that a condition is something to be obeyed – it is rather more like a command or something one ‘ought to do’. The latter is compatible with the fact that we don’t do it regularly, or that it does not give an accurate psychological description of how we normally come to acquire self-knowledge. Moran is explicit about the fact that he is interested in “trying to do justice to a certain tension in our thinking about the possibilities of self-knowledge” and “the distinctiveness of the first-person perspective more generally” (Moran 2001, xxx). Conforming to transparency, Moran says, is a “normative demand” or “a normative requirement” (Moran 2001, xvi–xvii).

But what are the normative claims, exactly? Transparency matters to ‘rational freedom’ (e.g. 2001, 107, 138) and/or ‘rational autonomy’ (2001, 31, 117, 139), and the ‘practical’ self-relation that Moran aims to describe has a “deeper relation to freedom or rationality than any other one (e.g., various modes of perception)” (Moran 2001, xxvi). He places his theory of self-knowledge in the context of Kantian and Sartrean discussions about agency:

The basic point can be expressed in a loosely Kantian style, although the idea is hardly unique to Kant. The stance from which a person speaks with any special authority about his belief or his action is not a stance of causal explanation but the stance of rational agency. In belief as in intentional action, the stance of the rational agent is the stance where reasons that justify are at issue, and hence the stance from which one declares the authority of reason over one’s belief and action. (Moran 2001, 127)

The view Moran defends, he says, is one that “aligns, or even identifies, reflective consciousness with rational freedom” (Moran 2001, 138). He writes,

the Sartrean and Kantian language ... is meant to convey [the] idea of different stances toward what is in some sense the same state of affairs, the same person. When I avow a belief, I am not treating it as just an empirical psychological fact

---

4 As Boyle writes, “the importance of Moran’s work is not primarily that it identifies a connection between doxastic self-knowledge and a special way of arriving at beliefs—by ‘making up one’s mind’—but that it shows how normal doxastic self-knowledge reflects something about what believing is” (Boyle 2011b, 235). On my reading of Moran, though, Moran’s project is about what being an autonomous agent is.
about me; and to speak of a transcendental stance toward it is meant to register
the fact that it is explicit in the avowal that it commits me to the facts beyond my
psychological state; and as a commitment it is not something I am assailed by,
but rather is mine to maintain or revoke. (Moran 2001, 8g)

Now, Moran makes what appears to me to be two distinct claims. One is about
being unable to make up one's mind, which, for the sake of simplicity, I'll assume
is the same as being unable to answer world-directed questions or to judge that P,
that is, being unable to adopt a 'transcendental stance' and unable to 'conform to
transparency'. Formulated positively, the claim is that making up one's mind is
crucial to human agency. However, Moran seems to (want to) make a different,
less general, claim, which is connected to his use of the concept of alienation. This
second claim concerns being unable to make up one's mind (judge, answer world­
directed questions, etc.) in specific cases. If one cannot answer the question of what
one believes by directly reflecting on its content, if one cannot 'avow' one's attitude,
then one is alienated from that attitude.

These two claims are importantly different. After all, cats do not have the ability,
or so I assume, to 'make up their minds' or adopt a transcendental stance, but it
seems farfetched too think they are somehow alienated. For human beings, the
capacity to make up our minds or to come to judgements about certain matters
seems an inescapable aspect of our agency (even if it's not a capacity we exercise
very often). It's possible to argue that this capacity to judge is itself somehow
'valuable' (rather than inescapable), and one might argue that in that regard cats and
cows somehow miss out on something. But this does not seem to be the sort of
claim Moran wants to make. The claim is rather that human beings have the relevant

5 This passage of Moran's, in particular the phrase that there are facts 'beyond' one's psychology, echoes the
following passage by Christine Korsgaard: "When you deliberate, it is as if there were something over and above
all your desires, something that is you, and that chooses which one to act on. The idea that you choose among your
conflicting desires, rather than just waiting to see which one wins, suggests that you have reasons for or against
acting on them. And it is these reasons, rather than the desires themselves, which are expressive of your will.
The strength of a desire may be counted by you as a reason for acting on it; but this is different from its simply
winning. This means that there is some principle or way of choosing that you regard as expressive of yourself, and
that provides reasons that regulate your choices among your desires. To identify with such a principle or way of
choosing is to be 'a law to yourself,' and to be unified as such. This does not require that your agency be located in
a separately existing entity or involve a deep metaphysical fact. Instead, it is a practical necessity imposed upon you
by the nature of the deliberative standpoint" (Korsgaard 1989, cf. 2009, 36, see also 1996, 100).

6 This is, roughly, what I've described in Chapter 4 as the 'transcendental' take on transparency, which I claimed is
implausible for reasons I turn to below.
capacity and that not exercising that capacity, with regard to one's attitudes, leads to alienation. For instance, discussing someone undergoing therapy, Moran writes,

he person might be told of her feeling of betrayal, and she may not doubt this. But without her capacity to endorse or withhold endorsement from that attitude, and without the exercise of that capacity making a difference to what she feels, this information may as well be about some other person, or about the voices in her head. From within a purely attributional awareness of herself, she is no more in a position to speak for her feelings than she was before, for she admits no authority over them. (Moran 2001, 93 Moran's emphases)

She is alienated because her feeling betrayed is, Moran says, “detached from her sense of the reasons” (Moran 2001, 93). If someone can only take a ‘theoretical’ or ‘empirical’ stance on herself and is unable to learn of her attitude by reflection on the content or object of that attitude, i.e. if she is does not conform to the transparency condition, then this leads to alienation. It is only by putting to use our capacity for practical reason, if we avow our own attitudes, that they are truly ‘our own’ in Moran’s sense. Conforming to the transparency condition is a way of avoiding alienation, or, more positively, a way of having authorship over one’s own attitudes. Moran mentions akrasia and self-deception as examples of where there is “the clash between these two perspectives on oneself” and where there is “a split between an attitude I have reason to attribute to myself, and what attitude my reflection on my situation brings me to endorse or identify with” (Moran 2001, 67).

Crucially, Moran explicitly claims that the theoretical or alienated route to self-knowledge is also a route to self-knowledge. This means that making up your mind is not, principally, a claim about ‘how we acquire self-knowledge’; rather, it’s a claim about how we acquire a specific sort of self-knowledge, namely non-alienated self-knowledge. Making up one's mind (avowing) versus theorizing about one's mind (attribution) are, he says, “seen as different routes to knowledge of the same thing” (Moran 2001, xviii; 89). The language of alienation brings out the fact that although alienated self-knowledge is also self-knowledge, it’s not the right or proper way of gaining self-knowledge. Moran’s project, therefore, is importantly normative, as is evident from Moran’s claims that there’s something “wrong” with the person who can only rely on behavioural evidence to report on his mental states (Moran 2001, 68) and that transparency matters to the psychological or psychological health of the person (Moran 2001, e.g. 107-108, 136-137).
The language of alienation and ownership that is central to Moran's work does not normally play a role in the self-knowledge debate. Instead, we find discussions about alienation, identification and ownership most notably in discussions about (personal) autonomy and/or authenticity (Christman and Anderson 2005; Oshana 2006; J. S. Taylor 2005). There are many (important) distinctions to be drawn, e.g. between personal and moral autonomy, local and global autonomy or indeed between autonomy and authenticity. Some of these distinctions will become relevant later on. For the moment, though, we can take the notion of autonomy to follow Christman and Anderson's definition and can

finds its core meaning in the idea of being one's own person, directed by considerations, desires, conditions, and characteristics that are not simply imposed externally on one, but are part of what can somehow be considered one's authentic self. (Christman and Anderson 2005, 3)

A common view is to think of authenticity as a requirement for autonomy (or to equate these notions) and to define authenticity in terms of non-alienation (see e.g. Christman 2009; B. Roessler 2015; Noggle 2008). As Robert Noggle writes,

The adjective “authentic” is commonly applied to elements of the person's psychology that are part of or produced by this true or real self. Thus, to say that an impulse is not authentic is to say that it does not lie within that part of a person's psychology that must be in charge if she is to be genuinely autonomous (or that must be the source of her actions if they are to count as autonomous). (Noggle 2008, 88)

7 The personal/moral distinction is controversial, and (sometimes) leads to greater confusion. Though let me say here that personal autonomy is often understood as something that is centrally concerned with knowing which maxims, principles, intentions, projects, roles, desires, and so on are ‘truly my own’, and hence is often connected to notions such as ‘alienation’ or ‘externality’. As Jeremy Waldron writes, personal autonomy “evokes the image of a person in charge of his life, not just following his desires but choosing which of his desires to follow” (Waldron 2005, 307). Or as Marlyn Friedman describes it, personal autonomy is “largely a term of philosophic art, yet it encompasses an array of notions familiar to ordinary people, notions such as being ‘true to myself,’ doing it ‘my way,’ standing up for ‘what I believe,’ thinking ‘for myself,’ and, in gender-egalitarian reformulation, being one’s ‘own person.’” (Friedman 2003, 3). Moral autonomy, on the other hand, at least on a (broadly) Kantian conception, has to do with which maxims can be adopted universally. Personal autonomy is, in a sense, primarily about how the individual relates to herself and her life, whereas moral autonomy is concerned with acting in a way that can be recommended to others or offering reasons that others could likewise adopt as principles of action (cf. O’Neill 2000, 25).
One reading of Moran’s view, then, is to see his rationalist view of self-knowledge as an articulation of the authenticity component, i.e. as providing a criterion of non-alienation.

Though Moran usually only talks about “rational autonomy” in his work on self-knowledge, he has suggested that his account can be applied to debates about autonomy, authenticity, alienation and ownership more generally. Though a great number of philosophers have engaged with the question of what it means for reasons or actions to be one’s own, to which I turn later, in his central work, Moran does not engage much with this literature. When he does engage with contemporary discussions regarding authenticity, ownership and identification, which he only does once, Moran (2002) argues that his account of making up one’s mind answers the question of what makes an attitude one’s own. So one’s ‘own’ reasons are just the reasons that one draws on to answer world-directed questions (in earlier terminology, just those reasons that meet the Anscombean Constraint). In his paper on the work of Harry Frankfurt, for instance, Moran writes that

> our concern with autonomy, both inside and outside of the study of philosophy, is not restricted to a concern with what we can be held responsible for, or how we may be proper subjects of praise and blame. … [T]he value of autonomy is broader than the requirements of moral evaluation, and is intimately connected with seeing oneself as living a life at all, rather than simply as a thing with a particular career through time. (Moran 2002, 189)

He adds that the latter brings with it “the distinction between what is ‘one’s own’ and what is experienced as other or alien” and that “it is only with respect to something with a quite particular kind of unity that we can speak of either identification or alienation” (Moran 2002, 189).

What I propose to do, then, is to take Moran’s rationalism as a theory of non-alienated or autonomous self-knowledge and to ask whether it’s plausible in that respect. On the rationalist view thus understood, what you need to do to make an attitude ‘yours’—so that you are not alienated with respect to it— is to ‘conform to the transparency condition’. Moran is not explicit about whether ‘conforming to transparency’ is a necessary condition for non-alienated self-knowledge or whether it is also meant to be sufficient. Moran does not mention any other conditions that would have to be in place over and above (being able to) meeting the transparency condition, or whether there are any preconditions for being able to make up one’s
mind, but this might just be because the target he is arguing against is introspectionist accounts. My strategy will be to assume, for the sake of the argument, that obeying the transparency procedure is sufficient for non-alienated self-knowledge, and I will consider what the implications of this are.

In the subsequent section, I turn to three examples in order to examine the idea that the circumstances of deliberation are irrelevant and that transparent question-settling is sufficient for non-alienated self-knowledge.

3. Deliberation in oppressive circumstances: three cases

The first example of the Deferential Wife should be familiar from Chapter 5:

The Deferential Wife: She buys the clothes he prefers, invites the guests he wants to entertain, and makes love whenever he is in the mood. She willingly moves to a new city in order for him to have a more attractive job, counting her own friendships and geographical preferences insignificant by comparison ... She does not simply defer to her husband in certain spheres as a trade-off for his deference in other spheres. On the contrary, she tends not to form her own interests, values, and ideals, and when she does, she counts them as less important than her husband’s. She readily responds to appeals from Women’s Liberation that she agrees that women are mentally and physically equal, if not superior, to men. She just believes that the proper role for a woman is to serve her family. As a matter of fact, much of her happiness derives from her belief that she fulfils this role very well. No one is trampling on her right, she says; for she is quite glad, and proud, to serve her husband as she does. (Hill 1991, 5; Westlund 2003, 485)

As Thomas Hill stresses, a woman does not have to be servile when she works to make her husband happy; she might “freely and knowingly choose to do so from love” (Hill 1991, 6). But the Deferential Wife is different. As Hill points out, the Deferential Wife is confused about herself: she believes she has a duty to defer to her husband but “she cannot fully understand that she has a right not to defer to him” (Hill 1991, 10), the latter of which “betrays a certain kind of self-respect” (Hill 1991, 6).
Though the example of the Deferential Wife raises many questions, I want to address a specific question: does the Deferential Wife have non-alienated knowledge of her attitudes? To answer this question, we need to ask whether she is able to deliberate over world-directed questions and come to a judgement. This, it seems, is something she is able to do (after all, there merely has to be “logical room” for her to do so.) So we can imagine that the Deferential Wife is able to consider a world-directed question such as “Is moving to X a good idea?”. Given that she considers her own values and interests as less important than her husband’s, she ignores her own prudential reasons and answers the world-directed question by considering her husband’s reasons or by imagining what his reasons would be. She comes to the judgement that yes, moving to X is a good idea, and settles her attitudes accordingly. Thus, if we apply the rationalist account to the Deferential Wife’s case, it seems the conclusion is pretty straightforward: she constitutes her attitudes by answering a world-directed question and acquires non-alienated self-knowledge in the process.

Second, consider the case of Patty Hearst, as discussed in Kathleen Taylor’s book Brainwashing:

Patty Hearst: On 4 February 1974, Patricia Hearst, heiress and granddaughter of the powerful US media magnate William Randolph Hearst, was kidnapped by an organization calling itself the Symbionese Liberation Army (SLA). She was kept bound and blindfolded in a closet for several weeks, physically assaulted, forced to have sex with SLA members, and threatened with death. Meanwhile the SLA demanded a ransom from the Hearst Corporation, including not only requests for money but for a food give-away worth millions of dollars and the release of two SLA members jailed for murder.

On 14 April of the same year Patty Hearst caused a sensation by participating in the SLA robbery of a bank in San Francisco, after which she publicly denounced her family and expressed her commitment to the SLA. (K. Taylor 2006, 10–13)

In order for Patty to know in a non-alienated way whether she believes, for example, that the SLA is an admirable organization or that her family is despicable, she has to answer the corresponding world-directed question “Is the SLA an admirable organization?” (or “Is my family despicable?”). Patty, we can imagine, considers the reasons, deliberates, weighs pros and cons and finally issues a verdict (a judgement): yes, the SLA is an admirable cause. So we can suppose that Patty thereby makes
up her mind and constitutes her attitudes. And so the rationalist view would say that Patty has made up her own mind and must have acquired non-alienated self-knowledge. She knows what she believes and what she wants, and these desires and beliefs are genuinely her own.

Third, consider Marina Oshana's (2003, 2006) example of the Taliban Woman:

**The Taliban Woman:** Some persons may deliberately forge lives in which autonomy is absent, though they possess the capacity and the freedom to do otherwise. Consider the situation of a woman living under a Taliban regime such as that which controlled Afghanistan until 2001. Suppose that this woman has embraced the role of subservience and the abdication of independence that it demands, out of reverence, a sense of purpose, and an earnest belief in the sanctity of this role as espoused in certain passages of the Qu'ran. Having previously enjoyed a successful career as a physician, this woman has since chosen, under conditions free of whatever factors might disable self-awareness, and with a considered appreciation of the implications of her decision, a life of utter dependence. She can no longer practice medicine (indeed, she is no longer permitted access to information about the science of medicine). She is not permitted to support herself financially. She has no voice in the manner and duration of any schooling that her children, particularly her daughters, may receive. She must remain costumed in cumbersome garb—a burqa—when in public. She is forbidden to enter common places of worship. She knows that any transgression, any show of independence counts as heretical defiance and invites punishment both swift and harsh. But a life of subservience is consistent with the Taliban woman's spiritual and social values, provides her with a sense of worth, and satisfies her notion of well-being. (Oshana 2003, 104)

It appears that the conclusion we have to draw about whether or not the Taliban Woman has non-alienated self-knowledge is the same as in Patty's case. The Taliban Woman can raise a world-directed question, deliberate over it, come to a judgement and thereby constitute her beliefs. Everything points towards the fact that, on the basis of Moran's account, we should conclude that the Taliban Woman, too, has what it takes to know her own mind in a non-alienated way. But despite the ease with which we can make sense of the examples by applying the deliberative account to them, it seems the result is somewhat unsatisfying.
It overlooks what seems to be central to the examples, namely the coercive or oppressive circumstances in which the Deferential Wife, Patty and the Taliban Woman deliberated and answered world-directed questions. Consider, for instance, the following dictionary definition of brainwashing:

The systematic and often forcible elimination from a person's mind of more established ideas, especially political ones, so that another set of ideas may take their place; this process regarded as the kind of coercive conversion practised by certain totalitarian states on political dissidents. (K. Taylor 2006, 3)

To brainwash someone is to “change a mind radically so that its owner becomes a living puppet—a human robot—without the atrocity being visible from the outside. The aim is to create a mechanism in flesh and blood, with new beliefs and new thought processes inserted into a captive body” (Hunter 1956; in K. Taylor 2006). If Patty's agency is severely limited or infringed upon—if she has become a ‘living puppet’—then should we really say that in Patty's case the preconditions for what is required in order to have the capacity to judge and/or that she is able to ‘make up her own mind’ through transparently answering the questions of whether P is true, Q is desirable, and so on, are satisfied?

This would seem to be a rather simplistic way of understanding what’s going on in these cases. If conforming to transparency is sufficient for non-alienated self-knowledge, then the circumstances in which deliberation takes place appears to be irrelevant, unless these women have not truly made up their minds, which appears to be a rather patronizing thing to say. But it is hard to overlook the fact that the answers that these subjects in the examples give to the world-directed questions are mere echoes of something else: the Deferential Wife echoes her husband, Patty echoes the SLA and the Taliban Woman echoes the oppressive cultural–religious norms around her. Can they really be said to know their own minds if, in some sense, they appear to have no mind of their own?

This is not meant to be a rhetorical question. The point here is certainly not that it should be obvious that the two women don’t have non-alienated self-knowledge. The point is merely that it also does not appear to be obvious that they do, either. My point is that either the way in which Moran defines alienation, in terms of adopting a deliberative stance versus adopting a theoretical/empirical stance to one’s attitudes, is unsatisfying or their having non-alienated self-knowledge does not seem to be the
most important thing to say about such cases, and so Moran's account is too limited to deal with them.

It's surprising that the deliberative account, which is so intimately connected to notions of freedom and autonomy, seems committed to treating the above three examples in the same way that it treats the example of, for example, knowing whether one believes a third world war is coming (Moran 2001, 61ff) or knowing that one believes it's raining (Moran 2001, 70ff). This is not to say that rationalism is committed to the conclusion that these women have non-alienated self-knowledge. But the rationalist view, in its current form, does not seem to have the materials to handle cases like those above in a satisfactory manner, i.e. in a way that respects their subtleties and intricacies, because it fails to differentiate between self-knowledge as it occurs in non-oppressive circumstances of deliberation and self-knowledge as it occurs in oppressive circumstances of deliberation. In its current form, rationalism appears to treat knowledge of one's belief that it's raining in the same way that it treats knowledge of one's desire to wear a burqa or the desire to remain faithful to those who kidnapped you. This does not mean that rationalism is false, but it does mean that Moran does not, at present, have the resources to deal with cases like those considered and that more needs to be said about what it means to have autonomous or non-alienated self-knowledge.

In the next section, I consider whether rationalism, as presently understood, may be able to say that the women in the examples are alienated after all, and reflect more generally on the question of how we might understand the notion of having reasons, attitudes or a mind of one's own.

---

8 This is symptomatic not just of Moran's theory of self-knowledge but also, it seems, of most if not all contemporary theories of self-knowledge. I have chosen to focus on Moran's account, though, because his account at least engages with notions of alienation and the broader Sartrean/Kantian tradition of thought that identifies "reflective consciousness with rational freedom" (Moran 2001, 138).
4. Having a mind of one’s own

As we’ve seen, Moran contrasts the first-personal, deliberative standpoint of making up one’s mind with the theoretical, third-person standpoint of simply attributing states to oneself. He contrasts the stance of ‘avowal’ with the stance of theoretical ‘attribution’, and writes that “attributional self-knowledge ... is the expression of an essentially third-personal stance toward oneself” (Moran 2001, 106). The only way, it seems, for the rationalist to say that these women are somehow alienated from their attitudes or that there is something not quite right, epistemically, in the above cases, would be to say that their route to self-knowledge is ‘theoretical’, not practical or deliberative. This comes down to the claim that they are not really ‘avowing’ their attitudes and, for example, that Patty takes a third-person perspective on herself.

Maybe this option should not be dismissed too readily. After all, it appears, for example, that Patty was not answering the question “What should I believe/desire?” by answering the corresponding world-directed question. Instead, she was answering the question of what she believes by answering the corresponding question “What does the SLA believe/desire?”. And so she does not, the rationalist might argue, properly make up her own mind, because she does not treat herself as an agent – she is, by deferring to the SLA, adopting a third-person perspective on herself. Understanding the cases of Patty, the Taliban Woman and the Deferential Wife as examples of taking a theoretical or attributional stance, would allow the rationalist to resist the simplistic conclusion that they have non-alienated self-knowledge.

This strategy does not seem very plausible, though. It does not seem as if, for example, Patty’s route to self-knowledge is theoretical or attributional in Moran’s sense. Attributional self-knowledge involves ascribing an attitude to oneself on the basis of psychological or other sorts of evidence, and is contrasted with expressing or avowing one’s mind. But it’s not as if their self-ascriptions take place in the context of a psychological experiment or psychotherapy. The non-alienated first-person perspective is one that “involves no essential reference to oneself at all” (Moran 2001, xvi), and the third-person perspective on oneself is defined in terms of a perspective that doesn’t look through one’s attitudes but at them. Adopting a third-person perspective is giving “a description of the person who happens to be myself” (Moran 2001, 160 emphases added). But it’s implausible to think, for example, that Patty considers herself in this way. The women in the examples are adopting a first-
person perspective. They're looking ‘through’ their own attitudes directly onto the world, and are not taking a detached or distanced perspective onto themselves. We have no reason to think that the subjects in the examples are not really coming to a judgement regarding some world-directed subject matter or that they are avowing their own attitudes, unless ‘judgement’ and ‘avowal’ are technical terms. 9

Actually, the above line of reasoning presents a possible problem for the rationalist account. It seems that for the Taliban Woman or Patty to be able to get a better understanding of their own attitudes, it would be helpful if they were capable not only of adopting the deliberative stance and looking ‘through’ their attitudes (transparently) but also of learning to look precisely at their attitudes, that is, of considering their beliefs and desires as psychological states. It would be good, in some sense of that term, if Patty came to realize that what she believes and why she believes what she does was due to being brainwashed by the SLA, or if the Deferential Wife came to realize that some of her desires and values are the result of oppressive circumstances. Even though taking a critical distance towards the sources of one’s attitudes is certainly no guarantee of well-being or (full) autonomy, it does seem helpful in at least some cases, perhaps especially those involving oppressive socialization.

Such self-understanding, though, is not of a ‘transparent’ kind. It involves an understanding of the representational nature of one’s states and so understanding oneself at least in part as a ‘mere’ psychological creature. It also involves placing these attitudes in a wider sociocultural context or “horizon of significance” (cf. C. Taylor 1989) in which one finds oneself. The problem, clearly, is that adopting such a bird’s-eye view of oneself is incredibly difficult—not just for the women in the above examples, but for anyone. However, what is surprising is that the demandingness or difficulty of adopting such an opaque or non-transparent perspective on oneself would not be the central problem for the rationalist. According to Moran, adopting a distanced perspective or looking at oneself and one’s attitudes rather than through them leads to alienation: “avowing one’s belief is strictly incompatible with maintaining reference to oneself” (Moran 2001, 106). If, therefore, any of these women were to consider their attitudes as having oppressive sources, which they would not upon reflection endorse, then arguably this opaque stance of looking at one’s attitudes would lead to more estrangement on the rationalist account, not less.

9 The problem here is parallel to the discussion of avowal in Chapter 3, as well as to the discussion of the ‘no proper judgement’ response in Chapter 6.
The sense in which the Taliban Woman and the others arguably lack a mind of their own in some sense of that term does not seem to have much to do with the two perspectives that Moran describes – it seems to have to do, principally, with the fact that no other life and no other reasons were possible or imaginable to them. The issue is arguably not what makes reasons or attitudes ‘one’s own’ in the sense of identification and/or alienation, but rather what makes reasons or attitudes one’s own in the sense of not being subjected to the will of others. The relevant notion of ‘autonomy’ should not be defined in terms of features such as identification and/or alienation that are principally relevant to ‘authenticity’, including Moran’s proposed deliberative approach, but in terms of what the preconditions are of being a deliberative agent. The central questions are what these preconditions are and whether or not being a deliberative agent in Moran’s sense is compatible with being, as the women seem to be, subject to the will of others.

In this context, Michael Garnett (2013, 2014) presents an interesting distinction between two different ways of making sense of the idea of being autonomous, or what it means to have reasons (attitudes) of ‘one’s own’. One is what he calls the ‘deep self’-view, according to which “[a]n agent is self-ruled just in case she is ruled by her deep self”, that is, a person is autonomous when the ‘self’ is doing the ‘ruling’. In this view, Garnett includes not just the sort of views as defended by, for example, Harry Frankfurt (1971, 1988) but also Moran’s account and those of others who identify the real or deep self with one’s capacity for practical reason (Garnett 2013, 23). Garnett argues that “even if we do have deep selves, this view does not represent a good way of thinking about self-rule” (Garnett 2013, 23). Instead, Garnett suggests that we should take the ‘self’ out of ‘self-rule’ and define autonomy negatively, or socially, in terms of being resistant to the will of others. Thus understood, the women in the examples arguably don’t have a mind of their own, which in turn raises questions with respect to whether making up their minds would be self-knowledge conducive.

These different approaches to ‘autonomy’ and how rationalism can be understood in light of them, require more careful discussion – this is what I plan to do in the remainder of this chapter. Whether or not Garnett’s proposed view is the right view, it seems that the rationalist is left with two options. Either s/he claims that the women in the examples and similar cases did not make up their own minds, because the reasons they used to answer world-directed questions were not ‘their own’ in a yet to be specified sense. This means we’ve got some explaining to do,
for such claims are presently no part of the rationalist account. Alternatively, the rationalist could say that they did make up their minds and so have non-alienated self-knowledge and are autonomous. If that is the case, though, we must accept that ‘making up one’s mind’ and knowing one’s mind in the process is possible even in circumstances of oppression and/or brainwashing.

The second option has a rather dominant position in the debate regarding personal autonomy. In the subsequent section, therefore, I discuss the second position in more detail by locating Moran’s account in the theoretical landscape of discussions regarding personal autonomy.

5. Procedural accounts of autonomy

If ‘making up your mind’ is a sufficient condition for non-alienated self-knowledge, then rationalism would bear striking similarities to so-called ‘proceduralist’ views of autonomy, which define autonomy in terms of internal, reflective criteria. Proceduralism is arguably the conception of autonomy that lies at the heart of most modern varieties of liberalism and has been the dominant view in the literature on (personal) autonomy (e.g. Frankfurt 1988; Dworkin 1988). Natalie Stoljar describes the view thus:

Procedural conceptions characterize autonomous agents—agents whose preferences and desires are genuinely their own—as those who critically reflect in the appropriate way to evaluate their preferences, motives, and desires. (Stoljar 2015)

On the proceduralist view, personal autonomy depends “only on the structural and/or historical character of a person’s psychological states and dispositions, and on an agent’s judgments about them” (Oshana 1998, 83). To be autonomous means that one’s attitudes or values are ‘one’s own’ and that the answer to the question of what makes an attitude one’s own lies in one’s capacity for self-reflection or one’s deliberative capacities. Marilyn Friedman, for instance, writes that “when an agent chooses or acts in accord with wants or desires that she has self-reflectively endorsed, then she is autonomous” (Friedman 2003, 5). Ronald Dworkin endorses a similar view of autonomy, which he suggests consists in “the capacity to raise the question whether I will identify with or reject the reasons for which I now act” (Dworkin 1988, 15).
An important element of proceduralist accounts is that they are ‘content-neutral’ or not ‘substantive’. The content of someone’s attitude – what she believes or wants – is “irrelevant to the issue of whether the person is autonomous with respect to those aspects of her motivational structure and the actions that flow from them” (Mackenzie and Stoljar 2000, 13). Mackenzie and Stoljar go on to say that proceduralist views “implicitly assume that the content-neutral procedural conditions they identify are both necessary and sufficient for autonomy”, whereas so-called ‘substantive’ theories suggest that there are “further necessary conditions on autonomy that operate as constraints on the contents of the desires or preferences capable of being held by autonomous agents” (Mackenzie and Stoljar 2000, 19).

If we regard the rationalist view as an account that aims to answer the question of what non-alienation or authenticity consists in, we can say that it principally consists in a person’s capacity to make up her mind, i.e. the capacity to conform to transparency. Autonomy, thus understood, consists in having the “capacity to endorse or withhold endorsement” from one’s attitudes (Moran 2001, 93). The rationalist account could then be considered to be ‘proceduralist’ in the sense that it likewise proposes that deliberative capacities are central and does not add any substantive or content-based conditions to what’s required to conform to the transparency condition.

However, the rationalist account also differs in important ways from dominant proceduralist views of personal autonomy. Most procedural views of autonomy are hierarchical. In order to explain what makes an attitude truly one’s own, proceduralists often distinguish between ‘higher’ attitudes, with which one identifies, and ‘lower’ attitudes, from which one feels alienated – an idea that is often associated with Harry Frankfurt’s work (Frankfurt 1971, 1976). A famous example is Frankfurt’s case of the unwilling addict, who is alienated from her alcoholic desires and so these aren’t ‘her own’, but this desire is still hers in some more minimal sense, or so hierarchical proceduralists are inclined to think. The appeal to hierarchies is evident in Friedman’s account, for instance, when he says that

to realize autonomy a person must first somehow reflect on her wants, desires, and so on and take up an evaluative stance with respect to them. She can endorse or identify with them in some way or be wholeheartedly committed to them, or she can reject or repudiate them or be only halfheartedly committed to them. If she endorses or identifies with her wants and desires, she makes them more
truly hers, more genuinely a part of who she is, and thus, more a part of her very identity as a particular, distinctive self than are the wants and desires that she has not thus self-reflectively reaffirmed. When she chooses or acts in accord with wants or desires that she has self-reflectively endorsed, and her endorsement is somehow a part cause of her behavior, then, according to this familiar generic account, she is behaving autonomously. (Friedman 2003, 4–5)

Self-reflection is central to hierarchical proceduralist accounts (only) because self-reflection makes identification with one’s attitudes possible. Self-reflection allows one to ask whether an attitude ‘belongs’ to oneself or not.

Hierarchical accounts are known to face the problem of a regress of never-ending higher-order desires. Also, on proceduralist views, the sort of questions that are central include, for example, “Do I (really) have the desire for X?” or “Does the intention to phi really belong to me?”. To answer such questions, most hierarchical proceduralists cannot avoid (or explicitly adopt) a so-called ‘dual model’ of the self, according to which there is a ‘shallow’ and a ‘deep’ sense in which one’s attitude can be attributed to oneself, the latter of which has likewise been taken to be problematic (see esp. Garnett 2015).

The rationalist view, if we understand it as a proceduralist view of autonomy, is different, arguably, because it isn’t hierarchical in this way. One might see the rationalist account of personal autonomy as trying to avoid the same problems that (some of) the hierarchical versions of the procedural view are said to face by defining autonomy in terms of the ‘activity of reason’ alone – this is Moran’s explicit suggestion (Moran 2002, 214). However, I don’t want to linger on the

10 The sort of questions that are central to the dominant proceduralist theories of personal autonomy (“Is the desire for X really mine?”) are not practical or deliberative questions issued from the first-person but attributional, empirical or theoretical questions. On the rationalist view, a person’s autonomy would not be a central matter of a person having certain pro-attitudes (second-order desires, etc.) towards her own attitudes. In fact, questions of this sort are arguably themselves self-alienating, because the very question of whether this or that attitude is one with which I do or do not identify only appears to be intelligible on the assumption that my attitude doesn’t change depending on what I have reason to want or intend. Instead of answering the question of what my desires are by asking “Do I really want X?”, the rationalist proposal might be that I ought instead to answer the world-directed question “Is X desirable?”. It is in virtue of the fact that the unwilling addict’s deliberative efforts (her transparent question-setting) don’t change what she wants or feels that it is an attitude from which she is alienated. So on the rationalist account of personal autonomy, it may turn out, paradoxically, that most of the existing proceduralist views are founded on a fundamentally self-alienating standpoint, namely the standpoint from which one asks whether this or that attitude ‘belongs’ to one where such belonging is assumed to be answerable by asking attributional, not deliberative, questions. In this respect, then, the rationalist proceduralist account might be preferable to hierarchical views. I will not pursue this argument further, though, because I think there are reasons to think proceduralism in general is implausible unless an answer is provided to the question of what the preconditions are of the relevant reflective capacities in terms of which autonomy is defined. I discuss this below.
differences and possible (dis)advantages of the different proceduralist views, or on how the rationalist construal fares with respect to hierarchical ones. Instead, I want to return to the examples discussed earlier and observe that on the rationalist view construed along proceduralist lines, it follows that the Deferential Wife, Patty and the Taliban Woman do not just have non-alienated self-knowledge; they are also autonomous. If being autonomous just means having the capacity to make up one’s mind, and assuming this is something they’re able to do, then there’s no reason to think that they would not be autonomous in the relevant sense. This would mean that if rationalism is a species of proceduralism, autonomy is compatible with being brainwashed and oppressed. As Stoljar writes, “[O]n procedural accounts in general, there is no reason in principle why choosing subservience, or adopting oppressive norms, could not be autonomous” (Stoljar 2015). Dworkin, for instance, considers a person who “wants to conduct his or her life in accordance with the following: Do whatever my mother or my buddies or my leader or my priest tells me to do.” Such a person, Dworkin claims, “counts, in my view, as autonomous” (Dworkin 1988, 21).

As mentioned, the proceduralist view is more or less the standard view in discussions about autonomy, as well as in debates regarding liberalism, and not without reason. It has its benefits. The proceduralist conception of autonomy respects pluralism with regard to different conceptions of ‘the good life’. Since proceduralists claim that what makes a person autonomous is not the content of her thoughts or choices but rather the ‘form’ it takes, such pluralism is easily taken on board. Also, it is able to accommodate the principle of first-person authority, that is, that the self-ascriptions of persons like the Taliban Woman should be taken seriously (because they are true, on proceduralist views) even if their own desires do not encourage their own well-being (Friedman 2003, 146). This last idea is one that goes back at least to Mill, according to whom paternalistic interventions are prohibited even if they increase the person’s well-being.

However, in recent years, the proceduralist view has been challenged, in particular its individualist conception of the person. Many have argued that when asking what autonomy consists in, the capacity for self-reflection or, for that matter, the capacity to make up one’s mind, cannot be the (whole) answer to the question of what it means to be autonomous. If rationalism is a species of proceduralism, then rationalism will face the same problems. I describe the alternative view of autonomy in the subsequent section.
6. Relational accounts of autonomy

Relational and feminist theorists have been concerned with the 'atomistic' conception of autonomy, which conceives of persons as independent and self-sufficient. In particular, Rawlsian and Kantian notions of the self are "said to be 'atomistic', that is, abstracted from the social relations in which actual agents are embedded" (Stoljar 2015). On a proceduralist view of autonomy, one's capacity for autonomy is enhanced the less others constrain one's actions. And so the goal, it seems, would be to create a society "allowing people to be as little dependent on others as possible" (Christman and Anderson 2005, 128). This conception of what it means to govern oneself is criticized by many as being "excessively individualistic" (Benson 2005, 118). Feminist and communitarian theorists, in particular, have raised concerns of this kind. Feminist philosophers, for instance, have argued that the idea that one can always shake oneself loose from one's history and societal ties has led to gross injustices. Personal autonomy, according to liberal feminism, depends on certain enabling conditions that are often overlooked (e.g. Oshana 1998, 2006; Mackenzie 2008; Anderson and Honneth 2005; Mackenzie and Stoljar 2000; Benson 1991, 2005; Baehr 2013).

Though I think the view that (some) relational theorists argue against is a straw man, I think their positive claims are important. On my reading of the broader relational approach to autonomy, the main question is what are the preconditions of the sort of capacities that proceduralists claim are fundamental to autonomy? It is argued that proceduralist views of autonomy are atomistic in the sense that they do not consider what makes the capacities (second-order reflection, conforming to transparency, whatever it may be) possible in the first place. The capacity for reflection, or being able to deliberate over and answer world-directed questions, is not a solitary achievement, at least not in terms of the development of such a capacity. Spelling out what it means to be autonomous requires taking seriously

---

11 The label is not mine but is used by relational/feminist theorists.
12 Communitarians such as Alasdair MacIntyre claim that one's history, family, city, nation, inheritance, and so on constitute "the given of my life, my moral standing point. This is in part what gives my life its own moral particularity." He goes on, "This thought is likely to appear alien and even surprising from the standpoint of modern individualism. From the standpoint of individualism I am what I myself choose to be. I can always, if I wish to, put in question what are taken to be the merely contingent social features of my existence" (MacIntyre 2007, 220). MacIntyre thinks these features are not contingent at all: "the self has to find its moral identity in and through its membership in communities" (MacIntyre 2007, 220).
various social, cultural and linguistic contexts that are necessary in order to develop the relevant capacities.¹³

These insights have proved to be the impetus for understanding autonomy itself in relational terms (e.g. Oshana 1998, 2006; Mackenzie and Stoljar 2000; Mackenzie 2008; Anderson and Honneth 2005; Benson 1991, 2005). The concept of ‘relational autonomy’ is not an entirely unambiguous concept. Perhaps in recognition of this fact, Catriona Mackenzie and Natalie Stoljar define the concept in their introduction to an edited volume as “an umbrella term” (2000, 4). They go on to say that the term is premised on the conviction that

persons are socially embedded and that agents’ identities are formed within the context of social relationships and shaped by a complex of intersecting social determinants, such as race, class, gender, and ethnicity. Thus the focus of relational approaches is to analyze the implications of the intersubjective and social dimensions of selfhood and identity for conceptions of individual autonomy and moral and political agency. (Mackenzie and Stoljar 2000, 4)

There are weaker and stronger ways of understanding the ‘relational’ approach. On the above description, it is in principle possible for a proceduralist to take these concerns on board without abandoning the claim that autonomy can be defined in terms of certain reflective capacities. Even if a proceduralist failed to mention the social and cultural contexts that are required for an individual to develop the capacity for reflection at all, accommodating the above concerns is not something that is in principle incompatible with the proceduralist view.

Some relational theorists, however, want to make a stronger claim that goes beyond claiming that the development of capacities for self-reflection or transparent deliberation has social roots. The stronger claim is that social circumstances do not merely facilitate a person’s autonomy but form “an inherent part of what it means to be self-directed” (Oshana 2006, 49). The phrase ‘inherent part’ is, of course, still ambiguous. It might mean that autonomy itself is a relational concept, perhaps in the way that being a parent is a relational concept: you can’t be a parent ‘on your own’. This, to me, seems a rather strange conception of autonomy, which seems

¹³ I think this can be accommodated on most (Rawlsian, Kantian) accounts. So the present point regarding the relevance of relational theories of autonomy can be taken as the claim that unless proceduralism provides us with an account of the preconditions of self-reflection/transparency, proceduralism is incomplete rather than false.
irreducibly first-personal. But in any case, the stronger conception regarding relational autonomy is that it's not just the development of autonomy that requires taking seriously all kinds of social circumstances; this is needed even once the capacity is in place. It's something that remains "vulnerable to disruptions in one's relationship to others", and so "full autonomy – the real and effective capacity to develop and pursue one's own conception of a worthwhile life – is only achievable under socially supportive conditions" (Anderson and Honneth 2005, 130). I am not concerned in what follows with whether or not this means autonomy itself is truly relational (in the way that, for example, being a parent is) or with whether autonomy itself is not relational but that its preconditions are (I refer to both views as 'relational' in what follows). The important difference is between those who think social contexts are merely causally necessary for the development of a capacity, including (some) proceduralists and those who think that you can fail to be autonomous depending on whether the socially supportive conditions are present or not.

Crucially, if it's true that one's autonomy is vulnerable to disruptions in one's relationship to others, then it appears possible for someone to satisfy whatever the proceduralist condition for autonomy is (being able to form second-order desires or being able to make up one's mind) without being autonomous. In other words, relational theorists argue that procedural conditions may be necessary for autonomy, but they are not sufficient, and they criticize proceduralist views in this respect. From a relational theorist's perspective, the proceduralist has mistaken authenticity for autonomy: proceduralists take the answer to the question of what it means for someone not to be alienated from her own attitudes as at the same time providing an answer to the question of what autonomous agency consists of. On a relational but not a proceduralist account of autonomy,

it is possible for two individuals to satisfy all the psychological and historical conditions we have been discussing, but to differ with respect to their status as autonomous beings—and this difference is to be explained in terms of some variance in their social circumstances. (Oshana 2006, 49)

The relevant circumstances may include, for example, coercion, oppression, manipulation, hypnosis, racism, being silenced, and so on.
This brings us to the three examples discussed earlier. On the relational view, the Deferential Wife, Patty and the Taliban Woman are not autonomous, in spite of satisfying the internal proceduralist criteria such as being capable of self-reflection or transparent deliberation. Oshana, from whom I’ve borrowed the example of the Taliban Woman, writes that

"the Taliban woman is not autonomous. In a “local” or occurrent sense of the term, she has chosen autonomously. Nevertheless, she fails to be autonomous in a “global” sense for the obvious reason that the life that she chooses, and toward which she experiences no alienation, is a life in which she is systematically subject to the ultimate will of others." (Oshana 2003, 104)

Despite the fact that the Taliban Woman "lives in a manner consonant with her preferences, and succeeds in achieving what she believes is in her best interests [and] does what she wants, what she wants frustrates the exercise of autonomy" (Oshana 2003, 105). Apart from certain internal criteria, then, certain ‘external’ criteria have to be met in order for a person to be autonomous. Put differently, one might take Oshana's point to be that whatever constitutes authenticity does not always, or at least not necessarily, constitute autonomy: "the road to autonomy is not always the road to achieving one’s aim" (ibid.).

Onora O’Neill, coming from a different angle, writes that “there is no general reason to think that action which receives second-order endorsement [as the proceduralist claims] is autonomous”; indeed she states that notions of self-identification, endorsement or second-orderedness more generally are “just too commonplace to guarantee any distinctive, ethically significant coherence or any sort of independence; it is not a likely basis for autonomy” (O’Neill 2000, 36).

The capacities that proceduralists have focused on are also essential on relational views, but rather than taking them for granted, they ask what makes these capacities possible in the first place, what sustains them and what can undermine them. Most relational theorists thus focus not on agents' capacity for self-reflection

---

14 In my view, the local/global distinction is not usually very helpful and leads to (greater) confusion. But ‘local’ autonomy, as it is usually described, is what is required for choices, preferences or desires at particular times to count as autonomous, and is transient or autonomous with respect to a particular attitude or action, whereas global autonomy is meant to be a temporally extended condition.

15 To which she adds that “none of this shows that second-orderedness is unimportant. Individuals bereft of all second-order preferences could hardly form coherent plans” (O’Neill 2000, 36).
and deliberation as such but on the circumstances in which such deliberation takes place. This applies not just to the traditional proceduralist view but also to the rationalist view understood along proceduralist lines. Carla Bagnoli, for instance, writes, in her discussion of Moran’s view, that we should realize that “the individual self is formed through practices of social recognition” and that “the self’s authority on itself cannot be understood independently of how it relates to other selves: they are instituted simultaneously and reciprocally limit their authority” (Bagnoli 2007).16

However, Moran’s view understood as a normative project does not, I think, mean that it is necessarily flawed in the sense that it is committed to the view that making up one’s mind is sufficient for being autonomous. It does not appear to be incoherent to say that making up one’s mind is, for example, only expressive of autonomy provided that certain conditions, such as those specified by relational theorists, are also satisfied.17 It does mean, however, that rationalism is, in its current form, incomplete.

To end this section, I want to emphasize that there is a great diversity of relational views of autonomy and that I haven’t been able to do justice to this diversity here. What all of them share, however, is the claim that deliberating over some subject matter is not sufficient for autonomy. What we need to do is address the question of what the (pre)conditions of such deliberative capacities are and to consider whether or not these conditions are met in specific cases. The question of what the preconditions of deliberation/autonomy are is not easy to answer. Minimally, it is plausible to think that autonomy requires things like food and clothing, health care, education, and so on. But, presumably, autonomy also requires (varieties of)

16 Bagnoli’s proposal is “that we conceive the deliberative perspective as second-personal, rather than first-personal” because “[t]he practice of self-reflection is also fundamentally second-personal, the activity of a self addressing itself” (Bagnoli 2007, 49). If Bagnoli means that (the development of) the capacity for transparent deliberation that Moran discusses requires intersubjectivity/other people, then this seems right and something that can easily be taken on board by rationalists. If, however, the capacity itself is second-personal, then I am not entirely sure what she means, and I tend to agree with Moran when he responds to Bagnoli by saying, “It is one thing to claim that reasons must be public, the sort of thing that can be made intelligible to another person, shared or respected by another person, and this is an idea that has roots both in Kantian and post-Kantian philosophy, as well as in Wittgenstein. It is quite another thing, however, to claim that reflection itself involves a second-person stance, a dialogue with a genuine other. For a second-person stance means a stance toward a separate freedom, a ‘self-originating source of claims’ (Rawls), something calling upon me and demanding my respect. I do not stand in such a relation to my own attitudes or to myself” (Moran 2007, 75 emphases added).

17 Alternatively, the rationalist could claim that one only really makes up one’s mind provided certain (relational) conditions are met, but this is problematic for reasons discussed in the previous chapter regarding what I called the ‘no proper judgement’ response.
self-respect (cf. Anderson and Honneth 2005) and not being subject to the will of another (cf. Garnett 2014). Furthermore, it seems that being autonomous might mean different things to different people, depending on their needs and capacities. This, too, is something that relational, feminist and disability theorists have drawn attention to.

If we ought not to conflate authenticity and autonomy, this leaves the question of what the implications are for how we should think about self-knowledge. This requires us to consider the relation between self-knowledge and autonomy.

7. Implications for self-knowledge

7.1 The simple view
Suppose we assume, regarding relational theorists, that for a person to be autonomous certain internal criteria must be met and certain external criteria (absence of oppression, etc.) must be met as well. According to what we might call the ‘simple view’ of the relation between autonomy and self-knowledge, a lack of autonomy and self-knowledge come and go together: a lack of autonomy implies a lack of self-knowledge. If that is the case, Patty, the Taliban Woman and the Deferential Wife simply don’t know their own attitudes, because they’re not autonomous. When the Taliban Woman, for instance, sincerely self-ascribes the desire to wear a burqa, she is mistaken, and we should not take her at her word.

I think the simple view should not be accepted. If the Taliban Woman is not autonomous, then the simple view suggests that we might say that her self-ascribed desires do not reflect what she ‘really’ wants. If that’s true, then one might wonder why doesn’t she just go ahead and overrule her sincere self-ascriptions. She might say that she wants to wear a burqa, but ‘deep down’, she doesn’t. Isaiah Berlin (1969) famously warned against the political dangers of talking about true selves, and true and actual wishes and desires, and so on. On what he describes as the ‘positive’ conception of liberty, the ‘self’ is

identified with reason, with my “higher nature”, with the self which calculates and aims at what will satisfy it in the long run, with my “real”, or “ideal”, or “autonomous” self, or with my self “at its best”; which is then contrasted with irrational impulse, uncontrolled desires, my “lower” nature, the pursuit of
immediate pleasures, my “empirical” or “heteronomous” self, swept by every
gust of desire and passion, needing to be rigidly disciplined if it is ever to rise to
the full height of its “real” nature. (Berlin 1969, 131)

We should, Berlin argues, be wary of just going against people’s sincere judgements
about what they want or believe and attributing ‘false consciousness’ to them. Doing
so, Berlin argues, has been at the heart of many oppressive and fascist regimes. 18 If
we say that people are not true to their ‘real selves’, then we are

in a position to ignore the actual wishes of men or societies, to bully, oppress,
torture them in the name, and on behalf, of their ‘real’ selves, in the secure
knowledge that whatever is the true goal of man . . . [it] must be identical with . . .
the free choice of his “true”, albeit often submerged and inarticulate, self.
(Berlin 1969, 133) 19

However, one might also say that just because the ‘real self’ reasoning has been at
the heart of oppressive and fascist regimes, as Berlin points out, doesn’t mean ‘real
self’ talk isn’t useful or indeed true. It seems that self-knowledge is an interesting
topic outside philosophy, precisely because we do distinguish between what’s on
the surface and what lies beneath. It appears to be part of commonsense to think
that sometimes you might have a kind of self-knowledge, but it’s not the real deal
– it’s not knowledge of your real or actual desires, preferences or beliefs. The
psychological literature on self-knowledge is all about knowledge of your ‘true self’
(cf. Timothy D. Wilson 2002). Also, the concept of authenticity and the idea that one
might have or lack a real self plays a central role in for example psychiatric contexts
and in discussions regarding deep brain stimulation (see, for example, Nyholm and
O’Neill 2016). But if Berlin is right, then perhaps we had better not talk about real
selves in these and other contexts.

18 This ties in with the literature on the (Marxist) notion of ‘false consciousness’ (e.g. Cudd 1994; Heath 2005;
McGlynn 2016b).
19 Cf. “It is one thing to say that I may be coerced for my own good, which I am too blind to see: this may, on occasion,
be for my benefit; indeed it may enlarge the scope of my liberty. It is another to say that if it is my good, then I am
not being coerced, for I have willed it, whether I know this or not, and am free (or ‘truly’ free) even while my poor
earthly body and foolish mind bitterly reject it, and struggle with the greatest desperation against those who seek,
however benevolently, to impose it” (Berlin 1969, 133).
Much has been written in response to Berlin's claims, and I do not have the ambition to add anything of substance to that literature here. However, even if one is not convinced by Berlin's argument and thinks 'real self' talk is not so problematic after all, there are still other reasons to think the simple view is false. These reasons have to do with the notion of 'first-person authority'. First-person authority, as explained in Chapter 2, is the idea that the subject is the (only) 'authority' on what state she thinks (or says) she is in, which explains why we are not often prepared to question or correct other people’s sincere self-ascriptions.

The standard view in the literature is that first-person authority is explained in terms of self-knowledge: people's self-ascriptions are authoritative because self-ascribers know their own attitudes. It seems to follow that if people lack self-knowledge, they lack the authority to say what their attitudes are. But this doesn’t explain that even if we have good reason to think someone’s self-ascription is false, we still are not prepared to challenge or correct their self-ascription. It might, for instance, be apparent to Sally that Harry is disappointed or insecure about himself, but Harry himself thinks he’s angry with his father. But this doesn’t necessarily mean Harry therefore has no first-person authority. If Harry were to sincerely say that he’s angry with his father, Sally may not be immediately prepared to override his self-ascription and say “No you're not” or something along those lines. It’s not unlikely that Harry would take offence if this was her reaction, and, we might think, rightly so.20, 21

The present suggestion is that not taking seriously or overruling someone’s sincere self-ascriptions without an excuse is not just impolite or rude but involves a specific sort of disrespect. The idea is that a possible reason why we do not normally challenge or correct people’s self-ascriptions, even when they are evidently false, is because taking someone’s self-ascription seriously is part of what it means to respect her as a person.22 Much more needs to be said about this interpretation

---

20 Or suppose you say you don’t want to eat meat, and a neuroscientist tells you, on the basis of some brain scans, that in fact you react very positively to meat and that you’re a carnivore without realizing that you are—a suggestion made in a recent Dutch documentary (‘Een Kijkje in de Hersenen van Marijn Frank’). Again, it would seem that a complete denial of someone’s sincere self-ascription is undesirable.

21 There are (many) different reasons for not challenging people's self-ascriptions. Some philosophers think that we have such authority 'by default' and that “in the absence of good reason to the contrary, one must accord correctness to what a subject is willing to avow, and limit one’s ascriptions to her to those she is willing to avow” (Wright 1998, 41). Expressivists, as we’ve seen, think that first-person authority is due to the fact that self-ascriptions are not reports but expressions.

22 See also the appendix to the previous chapter and Strijbos and Jongepier (forthcoming) for a discussion of the principle of first-person authority in the context of psychotherapy.
of the principle of first-person authority and its relation to self-knowledge. For now, though, the idea is simply this: if we adopt the simple view and deny without hesitation that the Taliban Woman, Patty and the others lack self-knowledge, this may be in conflict with the principle of first-person authority, given that first-person authority is naturally explained in terms of self-knowledge. In other words, if they lack self-knowledge, then we would need some other explanation of why we are not entitled to overrule or challenge their self-ascriptions and why we are not entitled to just overrule them, even if we know they are false.  

What's wrong with the simple view, I want to suggest, is that it doesn't distinguish between different types of self-knowledge. It's acknowledged that it's possible to distinguish between different objects of self-knowledge, e.g. knowledge of your sensations versus knowledge of your attitudes. But this does not help to get a more nuanced perspective on the Taliban Woman and the other cases, because in all cases we're considering their attitudes. As we've also seen, Moran further suggests that

---

23 As suggested in the appendix to the previous chapter, maybe such a (non-epistemic) explanation can be provided, i.e. an explanation in which self-knowledge and first-person authority do not necessarily come (and go) together. On what I imagine might be a 'moral' approach to the notion of first-person authority, such authority would be described by understanding such authority as connected to, or being parasitic on, respect for human dignity. On the broadly Kantian line of thinking I have in mind, disrespecting someone's self-ascriptions may involve treating her merely as a thing, not as a person (cf. P. F. Strawson 1974; Langton 1992, 2009). If the Taliban Woman expresses the desire to live a life of subservience, then the fact that she lacks self-knowledge, if indeed she does, should not be a reason to think that she also lacks first-person authority, i.e. is no reason to go ahead and correct, challenge or overrule her self-ascription. It seems to me that rationalism is best suited to explain such authority, for the Taliban Woman still has the capacity for coming to judgements, and arguably this provides the ground for the basic sort of respect for persons that I suggest is part of the notion of first-person authority. This would mean that rationalism offers the right theory of first-person authority but maybe not the right theory of self-knowledge. It would also mean that someone might lack self-knowledge, but not first-person authority, in which case, maybe the simple view is somewhat less controversial. For instance, in his discussion of whether women working in the pornography industry or prostitution really know their own desires, Aidan McGlynn (2016a, for related discussions about pornography see e.g. Bauer 2015 and Langton 2009) asks: "Isn't there a danger of denying the agency of a number of grown women, many of whom will even explicitly say that they enjoy their work? These are good questions, and they don't have easy answers." McGlynn I think rightly highlights that "such questions are not unique to the topic of pornography". He considers "the wearing of makeup, high heels, and uncomfortable underwear, or the desire to get married to a man, have children, and let one's husband be the main bread-winner" and asks to what extent these are "genuinely free choices that should be respected, and to what extent are they reflections of the facts that in patriarchal societies women are often rewarded for certain attitudes and behaviour and punished for others, and that women form their identities, attitudes, and behavioural patterns in a society seeped in patriarchal ideology, which they internalize to some extent?" I think these are questions that deserve our (philosophical) attention, and I also believe theorists of self-knowledge should at least be able to say something about what makes these questions so difficult. After all, from a purely 'introspective' point of view, there seem to be no challenges to self-knowledge whatsoever. However, once we address the more difficult question of what makes a mental state (e.g. desire) one's own, things don't seem to be so easy. How could one have knowledge of a mental state that in some sense does not really belong to one?
there may be different ‘routes’ to self-knowledge (theoretical versus practical). But this doesn’t help either, because we have no good reason to think that the Taliban Woman and the others were unable to make up their minds or that they are actually adopting a theoretical or empirical stance towards themselves.

My suggestion is quite simple: instead of talking merely about different objects or routes to self-knowledge, we should also talk about different types of self-knowledge. We can minimally distinguish between introspective self-knowledge, non-alienated self-knowledge and autonomous self-knowledge. I describe these types below, and define atomist/holist approaches in terms of them.

### 7.2 A nuanced view: three types of self-knowledge

Introspective self-knowledge is, simply, knowledge of one’s current mental states, one’s present judgements, conscious thoughts and words “running willy-nilly” through one’s head (Frankfurt 1988, 59). Introspective self-knowledge is prima facie easy: the Angry Spouse from previous chapters has introspective self-knowledge; so do the Deferential Wife, Patty and the Taliban Woman. Even those who are hypnotized, drugged, tortured, and so on know what they are thinking here and now, as long as they are minimally conscious. The key question is how one thinks of the relation between occurrent thoughts and attitudes. Is knowing one’s occurrent states (thoughts or judgements) sufficient for the latter? Or is the former merely an indication of the latter? I turn to these questions—and the relation to atomist and holist approaches to self-knowledge—in the next section. Here, I want to distinguish introspective self-knowledge from two other types of self-knowledge.

Non-alienated self-knowledge is, to put it somewhat unhelpfully, ‘knowledge of mental states from which one is not alienated’. What non-alienated self-knowledge is depends on what one takes ‘alienation’ to be. Non-alienation might mean that one positively identifies with one’s true self, or that no “negative judgment about or a negative emotional reaction” to one’s attitude is involved (Christman 2009, 155–56). If that is the case, non-alienated self-knowledge involves the presence of second-order identification or the absence of negative judgements or feelings with respect to the self-ascribed attitude, as most proceduralists suggest. Alternatively, on a rationalist conception of the concept of alienation, non-alienated self-knowledge is to be defined in terms of the way in which one gains self-knowledge. One has non-alienated self-knowledge if and only if one answers the question concerning one’s attitude by considering the reasons in favour of believing/wanting/hoping for
P. In other words, if one conforms to the transparency procedure. These are (very) different ways of specifying what non-alienation consists in, but both should be contrasted with what being autonomous consists in.

We can add a third type of self-knowledge, which we can call ‘autonomous self-knowledge’. Again, what autonomous self-knowledge is will depend on what one takes ‘autonomy’ to be. Proceduralists think that autonomous self-knowledge just is non-alienated self-knowledge. But on the relational view that I’ve been discussing, given that autonomy is defined not (just) in terms of non-alienation but in terms of certain extra or external conditions, a different type of self-knowledge emerges. On a broadly relational conception of autonomy, one would have autonomous self-knowledge if, for example, the attitude in question is not, for instance, the result of being subjected to the will of another (cf. Oshana 2006, 62; Garnett 2014). On a relational view, whichever way we define it, autonomous self-knowledge and non-alienated self-knowledge are different types of self-knowledge, and it differs from the proceduralist view in this respect.

If we apply this threefold distinction to the examples discussed in this chapter, we should say that the Deferential Wife, Patty and Taliban Woman all have introspective self-knowledge. They also have non-alienated self-knowledge, construed in either of the above ways. The Taliban Woman, after all, lives a life that is “consistent with her spiritual and social values, provides her with a sense of worth, and satisfies her notion of well-being” and she “experiences no alienation” (Oshana 2006, 60–61). It seems this is no different for the other two women. They can either (a) identify with their attitudes and/or be true to their real selves or (b) ‘conform to transparency’. However, if the women in the examples are not autonomous, then they lack autonomous self-knowledge. This is not because there is a mental reality (real self, etc.) of which they are unaware or not faithful to but because the notion of having a mind of one’s own requires the absence of coercive or oppressive circumstances; more generally, it requires that one isn’t subject to the will of another, which in their case is arguably a condition that is unfulfilled.

What this leaves us with, I think, is a more nuanced approach. The nuanced conclusion is thus that the Deferential Wife and the other women appear to have self-knowledge ‘in some sense’ but lack self-knowledge in another sense. This conclusion, and distinguishing between different types of self-knowledge more generally, is maybe not particularly exciting or controversial. Expressions like “She does not really know what she wants” or “He does not really believe that” illustrate,
I think, that the notion of self-knowledge is not homogenous at all. Perhaps in this respect the threefold distinction is trivial.

However, it does not seem to be trivial if we consider the philosophical debate on self-knowledge. Philosophers have only recently begun to distinguish between different types of self-knowledge by recognizing that different objects of self-knowledge may involve different ways of knowing them. Moran is an exception in this respect, for he at least recognizes that there may be 'good' (deliberative) and 'bad' (alienated) cases of self-knowledge. I've suggested that on (one reading of) Moran's view, the way he defines the good from the bad types of self-knowledge is unsatisfactory. The rationalist only has one way of accounting for the way in which one might 'really' or 'not really' know one's own attitudes. For Moran, varieties of self-knowledge are defined in terms of two "different routes to knowledge of the same thing" (Moran 2001, 89). So if we want to say, for example, that the Taliban Woman knows her own mind in one sense but does not know her own mind in another sense, the only way in which Moran can explain this would be if the Taliban Woman were to know her attitude by interpreting evidence of various kinds, i.e. by adopting a theoretical stance. But it seems clear that this is not what's going on. And so, as we've seen, the Taliban Woman 'has what it takes' in terms of self-knowledge, at least on the rationalist account that's currently available. However, whether Moran's way of defining the good and the bad cases is ultimately the most plausible is a matter that is independent of the claim that there are good and bad cases. The topic of self-knowledge involves more than just asking how one acquires self-knowledge, for it also involves asking how one acquires which type of self-knowledge and what makes some types of self-knowledge more important than others. So despite having been critical of Moran's view and the different versions thereof, there is an important insight that we owe to the rationalist, which is that the very question of self-knowledge is intimately connected to the question of what makes certain mental states yours and that the question of self-knowledge should be studied with this in mind.

8. The metaphysical and epistemic sources of atomism

In this final section, which is also the last section of this thesis, I want to take a step back and abstract away from the details of rationalism (or expressivism) to consider
the metaphysical and epistemic sources of atomism, for the idea of ‘ownership’ regarding one’s mental states is not something that many philosophers in the self-knowledge debate have been concerned with (though this certainly isn’t true outside the debate). Maybe this does not mean anything, but I am inclined to think it does. I think the fact that the question of what it means for a desire, belief or emotion to be one’s own has been ignored is because issues regarding ownership, identification, authenticity and autonomy are taken to be irrelevant to the question of what attitudes are and therefore what is required for a subject to know them. This, I think, reveals a distinctive way—an atomist way—of understanding the question of self-knowledge.

Though it may not be immediately evident, the discussion regarding autonomy and self-knowledge has a lot to do with atomist and holist perspectives on self-knowledge, in particular with the metaphysical question of what one takes mental states to be and, in turn, what is required to know them. The central question is at which level should we locate the knowledge a person has of her own attitudes: the introspective, non-alienated or autonomous level?

My thought here is as follows: the atomist thinks that the question of how one knows one’s attitudes can be answered by answering the question of how someone acquires introspective self-knowledge, i.e. how someone acquires knowledge of occurrent states, such as one’s judgements or current thoughts or what’s in inner speech. Moderate atomists and holists, by contrast, will think that the question of attitudinal self-knowledge is not necessarily answered by answering the question of how someone acquires introspective self-knowledge. What you consciously think or judge might, but need not, tell you what your attitudes are. The moderate atomist will cling to procedures and methods to try and answer the question of how and why judgements provide one with knowledge of one’s attitudes. The holist, on the other hand, will think this a question that can only be answered by considering the life the subject leads, i.e. by giving the ‘self’ in ‘self-knowledge’ its due.

Many of the theorists of self-knowledge claim that what they want to explain is the knowledge that subjects have of their own attitudes, and they go on to focus specifically on beliefs—not just beliefs but also ‘conscious’ or ‘present’ beliefs, and not just present beliefs but also occurrent thoughts and/or judgements. What we thus get is something like the following cascade:
A pervasive assumption, made by theorists who defend very different accounts of self-
knowledge, is that we can provide an account of how we know our own attitudes by
providing an account of occurrent thoughts or judgements. This assumption would
be reasonable (only) if one thought that such things as thoughts and propositional
attitudes are somehow the same or sufficiently similar to warrant similar treatment.
This assumption, and the tendency to equivocate between attitudes, beliefs,
judgements and conscious thoughts, reveals a distinctly atomist conception of what
attitudes are. It is this metaphysical assumption that lies at the root of the atomist idea that if you speak your mind or make up your mind – an activity restricted to the
here and now – you (normally) acquire knowledge of your attitudes.

The 'metaphysical face' of atomism can be understood as involving a specific
way of thinking about the cascade above or about way of thinking about the relation
between the sort of things at the bottom of the cascade (occurrent thoughts and
judgements) and those located at the higher end (selves or persons, and, to zoom
out further, families, societies or cultures). We may imagine that an atomist thinks
there is no such link: in order to know what one's attitude is, one does not need to
have knowledge of the subject – the person or self – whose attitude it is.

The atomist approach to the metaphysical question is problematic for various
reasons, the main one being that it just isn't plausible to see a person's 'attitudes' and
'thoughts' as (sufficiently) the same, not even if we restrict ourselves to beliefs (as
discussed in Chapter 6). Even the sort of states that fall into the category of 'attitudes' are
so diverse that it's not obvious that we could even treat, say, beliefs as similar to desires, or
desires as similar to intentions. Beliefs and desires appear to be fundamentally different
kinds of states, and it's not evident that answering the question of how we know our
beliefs gives us an answer to the question of how we know our desires. If there is so
much variety within the class of attitudes, then it appears to be rather presumptuous to
think that providing an account of how we know our own thoughts/judgements gives us
an account of how we know our own attitudes.
A couple of decades ago, Richard Rorty complained, “The attempt to hitch pains and beliefs together seems ad hoc—they don’t seem to have anything in common except our refusal to call them ‘psychological’” (Rorty 1979, 22; cf. Martin 2000, 99). It’s certainly a good thing that philosophers now recognize the difference between ‘sensations’, such as pain, on the one hand, and attitudes, like belief, on the other (cf. Coliva 2016; Boyle 2009b). But the class of ‘attitudes’ is still too diverse and should not be lumped together metaphysically, nor should we assume that we can give a similar epistemic account of them. Thinking of attitudes as things that allow us to focus exclusively on the bottom of the cascade amounts to an atomist way of thinking about what is required to know such states.

A holist, on the other hand, thinks that the question of what attitudes are should be addressed by considering the person whose attitude it is and the life she’s lived and/or intends to live, i.e. requires giving due attention to what’s located at the higher end of the cascade: persons, families, societies and cultures. In this context, a holistic approach to self-knowledge can be compared to various narrative theories of personhood and the metaphysics of mind. According to narrativists, the question of whom a mental state belongs to is not a contingent feature of that mental state (see e.g. Wollheim 1986; Schechtman 1996; Slors 2001; Bruner 2004; MacIntyre 2007; Goldie 2012; Jongepier 2014). Marya Schechtman (1990), for instance, claims, plausibly it seems to me, that “presuppositions about who has a psychological state come in at a level deeper than the level of the connections between states; they are necessary to defining those states as well” (Schechtman 1990, 84). Helpful in this context is Peter Goldie’s (2012) description of narrativity in The Mess Inside:

Our lives have a narrative structure—roughly speaking, they comprise an unfolding, structured sequence of actions, events, thoughts, and feelings, related from the individual’s point of view. A narrative, of course, can be recounted in vastly varying degrees of detail: I can summarize my whole life in ten minutes; or I can take an hour to tell you what happened to me in the last twenty-four hours. But, however much detail is provided, to be faithful to the narrative of my life I must show how its parts fit together in a structured way—making sense from my point of view as part of the whole. (This, as we all know, is often a difficult and painful thing to do.) Similarly, with emotional experience, it is the notion of narrative structure which ties together and makes sense of the individual elements of emotional experience—thought, feeling, bodily change, expression,
and so forth—as parts of a structured episode; and in turn it underpins the way that individual emotional episodes relate to the emotion of which the episode is a part, and this emotion to mood, to character trait and to character, and to the person's life seen as a whole. To make sense of one's emotional life, including its surprises, it is thus necessary to see it as part of a larger unfolding narrative, not merely as a series of discrete episodes taken out of, and considered in abstraction from, the narrative in which they are embedded. A true narrative, as I understand it, is not simply an interpretive framework, placed, so to speak, over a person's life; it is, rather, what that life is. (Goldie 2012, 1).

The sort of view Goldie proposes with regard to emotions is one in which emotions cannot be defined exclusively in terms of ‘episodes’ or the “individual elements of emotional experience”, i.e. in terms of whatever is at the lower end of the cascade, but requires that the subject is able to relate to and understand these episodes in light of things at the higher end of the cascade. A holist approach to self-knowledge may appeal to a similar ‘narrative’ view and might see the question of what it is to have, understand and know one’s attitudes as structurally similar to how Goldie proposes what is required for one to have, understand and know one’s emotions.

Note that this does not mean that the subject must see or experience herself as a ‘protagonist in a plot’, or that she must see her life as having a ‘beginning, middle and end’, as some narrativists have claimed, and which has, rightly it seems to me, been criticized (G. Strawson 2004; see also Jongepier 2014). I rather mean for the label ‘narrative’ to bring out the idea that knowing one’s attitude requires self-understanding, i.e. understanding the role that that attitude plays is seen as something like a node in the larger web of one’s life, and where the person whose attitude it is has a specific (reflexive) relation towards her own attitude.

The upshot of this is that when we ask what is required for someone to ‘know’ her own ‘attitude’, we should carefully consider the question of what one takes ‘attitudes’ to be and what it would mean to ‘know’ them. In other words, to properly understand the question of self-knowledge, and whether atomism or holism is plausible, we must

---

24 Note that what I mean by ‘narrative’ is meant to be even more minimal than, for example, Christman’s narrative view, according to which “the person must be able to look upon the factors and events of her life with a certain interpretive reflection” (Christman 2004). Perhaps some readers will worry that what I’m calling narrative isn’t narrative, but I will steer clear of these terminological disputes here. See also Slors (2001).

25 Cf. the brief discussion of reflexivity in Chapter 3.4.
first try to get a proper understanding of the ‘self’ component in self-knowledge, as well as the ‘knowledge’ component. In terms of a motto, on what I take to be the atomist view, self-knowledge does not require ‘self-understanding’, whereas on a holist view, knowing your attitude does require understanding your attitude.\footnote{Katsafanas for instance writes, in his discussion of Kant and Nietzsche on self-knowledge: “I might take pity to be a state aimed at helping those in need, whereas Nietzsche suggests that it can constitute a covert attempt to extend my power over others (see D 132-8, GS 338, GM III.14, BGE 260). Thus, even if I manage to identify one of my motives, I may not understand it.” (Katsafanas 2015, 121) Similarly, perhaps, I might be able to identify and thus ‘know about’ certain of my beliefs and intentions, but fail to understand them, i.e. have self-knowledge without self-understanding. Unfortunately, I cannot offer a fuller discussion of the relation between self-knowledge and self-understanding at this point. Giving a minimal account of their relation will probably involve answering the question of what the distinction between ‘knowledge’ and ‘understanding’ more generally comes down to, if indeed there is such a distinction. Knowledge, as Duncan Pritchard claims, is “concerned with propositions, whereas understanding usually isn’t, at least not directly” (Pritchard 2009, 30). What is also distinctive about understanding is that understanding is something that can come in degrees (Kvanvig 2003, 196). This in turn raises the question of how the relation between knowledge and understanding must be understood, and whether an answer to this question has implications for understanding the relation between self-knowledge and understanding. Pritchard notes that “the standard view within epistemology is that understanding is distinctively valuable but that it is not a species of knowledge” (Pritchard 2009, 31; cf. Kvanvig 2003, 196), but that “outside of epistemology the consensus is clearly that understanding is a species of knowledge” (Pritchard 2009, 38). Pritchard argues that this view is “false” by arguing that “one can not only have understanding without the corresponding knowledge, but also knowledge without the corresponding understanding” (Pritchard 2009, 38). I take it that how the distinction is to be understood, and whether knowledge requires understanding or the reverse, depends on the subject matter. Pritchard’s claims do not necessarily apply to moral ‘knowledge’, for instance (Düring and Düwell for instance claim that “Understanding what it means to know something ... presupposes understanding what it means to understand something” (Düring and Düwell 2015, 945)), and indeed his claims do not necessarily apply to self-knowledge, either.}

question about what makes these attitudes one's own, i.e. what makes these states belong to one's self. A holist will see the question of what is required for someone to gain knowledge of her attitudes as a question to be answered not at the level of introspective self-knowledge but rather at the level of non-alienation or autonomy.

These more fundamental issues about what attitudes are and what it means for them to be one's own have consequences for how the notion of 'self-knowledge' should be understood and studied. A number of contemporary writers on the philosophy of self-knowledge begin their books with a list of disclaimers warning against disappointment for those readers who were expecting a book about self-knowledge to be about something like 'knowledge of the self' or interesting states of mind, such as those that "reflect depth of character or strength of will; the ones that motivate heroic striving or evil deeds; the ones that get repressed, create neuroses, and need psychoanalytical uncovering and treatment" (Bilgrami 2012, 1–2). Instead, the innocent reader who stumbles upon the self-knowledge debate is confronted with literature that is concerned with "a kind of self-knowledge that nearly everyone thinks is easy to come by, almost to the point of triviality" (Carruthers 2011, xi). Rather than getting a theory of how we know our 'selves', what we get is a theory of how we know our own minds, more specifically our mental states, or actually only our propositional attitudes, in particular our beliefs, or indeed, what we get is a theory of how we know our conscious beliefs or occurrent thoughts.

I am one of those disappointed readers. But my disappointment is not due to the fact that I expect philosophers of self-knowledge to provide me with a 'theory of wisdom' or a guide to what makes me happy or what living a good and fulfilled life consists in. I suspect that questions about how we know our own tragic or heroic mental states or what makes us happy are not the sort of questions that we should expect a philosophical theory of self-knowledge to provide us with in the first place. Maybe it isn't something we should expect philosophy to deliver the answers to but something we should turn to novels for.

My disappointment rather has to do with the fact that most of the literature on self-knowledge seems to overlook the question regarding the preconditions of self-knowledge, more specifically the circumstances in which people self-ascribe desires, beliefs, intentions, emotions, hopes and expectations and other attitudes, i.e. the sort of states that we normally take to have a reasonable life expectancy and which lead us to think and act in various ways. The fact that questions regarding the circumstances have been overlooked would be understandable if we could
assume that the question of how one acquires knowledge of one's own attitudes can be reduced to the question of how one knows about one's own judgements or occurrent thoughts. In order to know that I am currently consciously thinking about pink elephants, it does not seem particularly relevant to take into consideration where I am, whether I'm excited, angry or sad, or who I'm with; indeed, it does not seem necessary to know what sort of person I am and what's important to me. I can be thinking about pink elephants, and know that I am, when I'm confused, tipsy or depressed and indeed without understanding myself at all. But are the circumstances equally trivial or irrelevant when it comes to the question of how we know our own attitudes?

The following are attitudes that I take myself to have: the intention to drink less coffee, the desire to have another one all the same, the belief that the outcome of the Brexit referendum is a disaster, the hope that things aren't as bad as they seem, and a particular fondness for my cats. These mental states, and my knowledge of them, appear to be tied up with 'who I am' in a way that my occurrent thoughts are not, or need not be. However, the above mental states, as well as the question of how I acquire knowledge of them, cannot be set aside as being too 'heroic'. My intention to drink less coffee does not motivate 'heroic striving'; my desire to have another coffee all the same isn't an attitude that 'creates neuroses'; my belief that the outcome of the Brexit referendum is a disaster and the hope that things aren't as bad as they seem do not require 'psychoanalytical uncovering and treatment'; and my being particularly fond of my cats does not reflect any particular 'depth of character'.

One of the leading concerns underlying this thesis has been this: that we can legitimately draw a distinction between knowledge of occurrent thoughts on the one hand and the gnōthi seauton or 'know thyself' sort of self-knowledge that the ancient Greeks were concerned with, or what Cassam (2014) has referred to as 'substantial' self-knowledge, on the other, that is, the sort of self-knowledge that "matters in a practical or even a moral sense" (2014, 31). More specifically, the worry is that a person's attitudes are seen as falling squarely on the 'occurrent' end of the divide rather than on the 'substantial' end. Though I have not argued for any of this explicitly, providing reasons to think that this way of looking at attitudes and knowledge thereof is misleading has been one of the underlying aims of this thesis.
9. Conclusion

In this chapter I have addressed the question of whether someone in autonomy-undermining circumstances can be said to know her own mind and, if so, in what sense. I have explored, specifically, Moran's account of self-knowledge, given that Moran, contrary to other theorists of self-knowledge, offers a distinctively normative theory of self-knowledge that is principally concerned with notions such as alienation, autonomy and ownership. I have argued that Moran's account bears similarities to so-called 'proceduralist' theories of autonomy and have therefore contrasted and compared his view with a so-called 'relational account' of autonomy that has criticized proceduralist views by arguing that having certain reflective abilities - such as the capacity to conform to transparency - is not sufficient for autonomy. I argued that if the objections raised by relational accounts of autonomy hold, then this means the rationalist account that focuses on transparency is subject to similar criticisms.

The constructive aim of this chapter has been to suggest that apart from distinguishing between different objects of and routes to self-knowledge, we should further distinguish between different types of self-knowledge. To this end, I distinguished between introspective, non-alienated and autonomous self-knowledge. I have not, however, offered much more than a rough sketch. Clearly, much more needs to be said about the relation between autonomy and self-knowledge, such as the relations between the three types of self-knowledge and the relation between self-knowledge and first-person authority. One of the most important questions, though, is what, exactly, it means to (fail to) be autonomous, i.e. what it means (not) to be governed by other people or institutions, or what is required in order to be resistant to someone else's will. An answer to these questions may, in turn, go some way towards answering the question of whether someone who is, say, not oppressed but is 'merely' manipulated by propaganda, or someone who is nudged, lacks self-knowledge in some sense of that term, or not. 27

27 Sophie Grace-Chappell has recently written on political deliberation under conditions of deception, Brexit in particular: "is deliberation binding when its results are arrived at by deception? As a question in the individual case, we have no difficulty at all in seeing that the answer to this question is 'Of course not'. Ever since Plato, perhaps longer, it has been commonplace to make an analogy between individual agency and the agency of the political community, ... If individual deliberation is regarded, as it clearly should be, as invalid when it has been warped by deception, there is no reason why political deliberation should not equally be invalidated by deception" (Chappell 2016) A question in this context would be to ask whether voters can be said to 'know what they want'
Despite the fact that many questions remain unanswered, I hope to have made a case for the idea that if we adopt a theory of self-knowledge that is intimately related to normative notions, such as Moran's, then it's legitimate to evaluate how these normative notions themselves are defined and whether the definitions stand up to criticism. Even if the threefold distinction between introspective, non-alienated and autonomous self-knowledge as I have described it is not the best way of addressing these topics, I hope the question of what the preconditions of self-knowledge are and the suggestion that there are interesting and important connections between self-knowledge and autonomy are issues at least worth putting on the agenda.

if their deliberative conclusions were arrived at by deception. Interesting in this context is Jason Stanley's work on propaganda. Stanley suggests that certain group identities that are formed under a regime of propaganda can "channel rational and affective streams in specific ways, creating obstacles to self-knowledge, as well as to the free flow of deliberation required in a healthy democracy" (Stanley 2015, 4). Stanley does not elaborate on how propaganda might create obstacles for self-knowledge, but it does not seem implausible to think that irrational feelings of fear and beliefs based on ideologies, for example, might make it harder for one to know what one wants or thinks.
I started this thesis with the question of what is required for a subject to know her own attitudes. A natural response to this question would be to say “Well, it depends”. The main task of the thesis has been to address the question of on what, exactly, it depends. To this purpose, I have described two different approaches to this question, which I have referred to as ‘atomism’ and ‘holism’. I used these lenses to look, specifically, at expressivist (Chapters 2-3) and rationalist (Chapters 4-7) theories of self-knowledge.

The principal aim of this thesis was to point out the limits of the atomist view, and in that respect the contribution of this thesis has been primarily negative. Atomism, as I have characterized it, is the view according to which answering the question of what is required for a subject to acquire self-knowledge involves focusing (almost) exclusively on the procedures or methods that she followed. With respect to expressivism and rationalism, this means that a subject necessarily or typically acquires self-knowledge by speaking her mind or by making up her mind. The question I have been concerned with is what is actually required in order for self-expression (expressivism) or transparent deliberation (rationalism) to deliver self-knowledge; in other words, I have asked what the preconditions are for such methods to be knowledge-conducive.

I described two versions of atomism, radical and moderate, both of which I have described in terms of what they have to say about the circumstances of self-expression and deliberation, respectively. I have taken the circumstances of self-knowledge in a broad sense, to include various psychological and sociopolitical circumstances in which the subject might find herself and how the subject’s self-ascription as made under these circumstances relates to her actions, values, character, other attitudes and self-conception. The radical atomist thinks that questions of circumstance are entirely irrelevant, or that all circumstances are ‘good’ circumstances, whereas the
moderate atomist appeals to normal or standard circumstances of self-expression or deliberation, i.e. by including certain (implicit) *ceteris paribus* clauses. The moderate atomist therefore thinks that we can mostly ignore the circumstances of self-knowledge.

Throughout the chapters I have tried to make a case for the idea that the moderate atomist view is untenable by showing that there is no way of defining what the right or normal circumstances of procedures should be such that they lead to self-knowledge rather than self-ignorance or self-deception that does not beg the question. My argument throughout the chapters has been that so-called 'abnormal' circumstances of self-knowledge are actually quite 'normal' for us, and that, at the end of the day, the appeal to (implicit) *ceteris paribus* clauses, not the procedures, ends up doing most of the work. If moderate atomism is an unstable position, then this leaves us with two alternatives: one is to retreat to radical atomism and avoid appealing to normal circumstances from the start, and the other is to develop a holist approach to self-knowledge.

I have mostly assumed, rather than provided arguments for, the idea that radical atomism is implausible, even though I have argued (in Chapter 6) that the (im) plausibility of radical atomism is something that depends on what answers one gives to more fundamental metaphysical and meta-theoretical questions. In any case, an important task for future research is to evaluate radical atomism in relation to holism in greater detail, and, in particular, to ask on the basis of which desiderata we should choose one over the other.

I have not directly argued for a holist account of self-knowledge, though I hope that the shape of what a holist approach would look like has emerged as a result of my consideration of the limits of moderate atomist versions of expressivism and rationalism. On a holistic approach to self-knowledge, the question of what is required for a subject to know her own attitudes depends on the circumstances in which she speaks or makes up her mind and on how her self-ascription as made under these circumstances relates to her other attitudes, values, character and self-conception. The result is that on a holist view, the question of what makes a self-ascription expressive of self-knowledge can only be answered by tailoring this question to the specific individual, the life that she leads and the psychological, social and moral-political context in which she finds herself.

To develop a holist account of self-knowledge, I don’t think it's necessary to start from scratch. There are interesting developments in, for example, feminist
Taking Stock and Looking Forward

and social epistemology that are concerned with questions of whether knowledge is connected to practical concerns, stakes, standards and social contexts and how the loss and gain of knowledge is connected to various forms of (in)justice and (in) equality that we may expect a holist approach to self-knowledge to be able to draw on. Apart from drawing on research in epistemology, developing a holistic view of self-knowledge is a project that will presumably benefit from going beyond the confines of epistemology by considering insights from the philosophy of language, social and political philosophy and (meta-)ethics. I hope to have been able to make a case for the relevance of various discussions outside epistemology to self-knowledge in Chapter 7.

I have left many important questions unanswered, even unaddressed, but I hope to work on these questions in the future. One such question is whether denying someone (autonomous) self-knowledge has any implications for first-person authority and, if so, which ones, and if not, why not. The more general issue in the background concerns the relation between self-knowledge and first-person authority, in particular whether the latter depends on the former or whether it should be given a non-epistemic explanation instead. Also, the relation between autonomy and self-knowledge deserves more attention than I have been able to give it. If we should distinguish, as I have claimed, not just between different objects and different routes but also between different types of self-knowledge – introspective, non-alienated and autonomous self-knowledge – then what a plausible account of these types of self-knowledge comes down to depends on what the right account of non-alienation and autonomy is; I have only scratched the surface of this matter. A final question for future research concerns the value of self-knowledge in light of the threefold distinction between introspective, non-alienated and autonomous self-knowledge. Which, if any, of these types of self-knowledge is especially important to have, and why? Is self-knowledge necessary for autonomy and/or well-being or rather the other way around? These are the sorts of questions I hope to work on in the future.

Despite the fact that I have not provided a positive account of what concrete shape a holist approach to self-knowledge can or should take, my aim in this thesis has, nonetheless, been constructive. The constructive aim has been to get a picture of the theoretical landscape of self-knowledge, in particular to show that answers to important questions about self-knowledge appear to be implicitly presupposed and that atomism with respect to self-knowledge is only plausible in light of certain
assumptions that are not unproblematic. One such question is the metaphysical question of what, exactly, we are trying to explain knowledge of, how things such as thoughts are related to attitudes and how the latter are connected to such things as persons and societies. The ‘metaphysical face’ of atomism would involve thinking of attitudes as equivalent or sufficiently similar to things like thoughts, and hence that we can give an account of how we know the latter by giving an account of the former. Another question is how we should understand ‘knowledge’ in the term ‘self-knowledge’. The ‘epistemic face’ of atomism would involve thinking of self-knowledge as something that one can have independently of self-understanding, where we might take the latter to be a capacity for understanding how one’s self-ascription relates to one’s larger biography, self-conception, past and future actions, and so on.

The constructive aim of getting a better view of the theoretical landscape of self-knowledge has negative consequences for the atomist, though. For it is only on a particular picture of what attitudes are and what it would mean to know them that atomism is plausible, i.e. that it would be legitimate to think that the question of what is required for someone to have knowledge of her attitudes can be explained exhaustively, or at least satisfyingly, by answering the procedural question of what the subject was doing when she ascribed a mental state to herself. To put it metaphorically, the more positive contribution of this thesis is to have made plausible the notion that the question of self-knowledge is a bit like a mobile hanging above a baby’s cot: if you touch one of the items dangling from it, the whole thing begins to twist and turn. Less metaphorically put, my hope is to have convinced at least some readers of the fact that when studying self-knowledge, we cannot suspend judgement on or ‘bracket’ larger metaphysical and epistemic questions.

I have argued in this thesis that any plausible account of self-knowledge needs to take holism seriously – not, perhaps, as a defining characteristic of the account itself but because any plausible account of self-knowledge will hinge on a proper understanding of these larger metaphysical, epistemic and moral questions and will therefore need to imply a comprehensive set of articulate answers to them.
Bibliography


Borgoni, Cristina. 2015. ‘On Knowing One’s Own Resistant Beliefs’. Philosophical Explorations 18 (2): 212–25.


Byrne, Alex. 2005. ‘Introspection’. Philosophical Topics 33 (1).


—. 2015b. ‘What Do You Really Want?’ The Philosopher’s Magazine Online.


—. 2017. ‘Diagnostic Error, Overconfidence and Self-Knowledge’. Palgrave Communications (3).


—. 2015. ‘Kant and Nietzsche on Self-Knowledge’. In Nietzsche and the Problem of Subjectivity. Berlin: De Gruyter.


—. 2016b. ‘Propaganda and the Authority of Pornography’. Theoria 31 (3).


——. 2015a. "Self-Knowledge and Communication". Philosophical Explorations 17 (2).


Schaubroeck, Katrien. unpublished manuscript. "Self-Confidence and Self-Knowledge".


Schwitzgebel, Eric. unpublished manuscript. "The Pragmatic Metaphysics of Belief".


Strijbos, Derek, and Fleur Jongepier. forthcoming. ‘Self-Knowledge in Psychotherapy: Adopting a Dual Perspective on One’s Own Mental States’. Philosophy, Psychiatry and Psychology.


Stel dat je met een enorm ochtendhumeur wakker wordt. Eigenlijk was je van plan om een boswandeling te maken maar, humeurig als je bent, besluit je dat je eigenlijk liever binnen wilt blijven. Heb je in zo’n geval zelfkennis of houd je jezelf voor de gek? Of stel dat je erg onzeker voelt of depressief of juist hoteldebotelt verliefd. Kun je in die omstandigheden weten wat je echte verlangens, overtuigingen en intenties zijn? Weet je wat je wilt wanneer je tijdens een woede-uitbarsting aangeeft een scheiding te willen? Weet je dat je een gezonde salade wilt wanneer het evident is dat je ‘genudged’ wordt? Ken je je eigen overtuigingen wanneer je bent gemanipuleerd door propaganda?

Het doel van dit proefschrift is niet zozeer een concreet antwoord te formuleren op dit soort vragen over zelfkennis, maar de meer fundamentele vraag te stellen wat er überhaupt voor nodig is om ze te beantwoorden. Meer specifiek is het doel van dit proefschrift te onderzoeken of hedendaagse theorieën van zelfkennis, expressivisme en rationalisme in het bijzonder, voldoende gereedschap hebben om dergelijke vragen op een bevredigende manier te behandelen, dat wil zeggen, op een manier die recht doet aan hun complexiteit.

1. Introspectieve en persoonlijke zelfkennis

Voordat ik in ga op wat ik concreet heb ondernomen in de hoofdstukken van dit proefschrift, wil ik eerst iets zeggen over het concept van ‘zelfkennis’ dat in het debat en in dit proefschrift wordt gehanteerd. Deze term is namelijk ambigu: het kan verwijzen naar wat we kunnen noemen ‘introspectieve’ zelfkennis enerzijds en ‘persoonlijke’ zelfkennis anderzijds.

Je wordt genudged wanneer de overheid (of een bedrijf) met behulp van inzichten uit, bijvoorbeeld, de (sociale) psychologie, je een duwje in de ‘goede’ richting geeft.
Introspectieve zelfkennis is kennis van je eigen (bewuste) mentale toestanden. Denk hierbij bijvoorbeeld aan kennis van je overtuigingen, verlangens, angsten, voorkeuren of intenties. Vragen die gaan over introspectieve zelfkennis zijn vragen als: “Ik geloof dat partij X mij het beste kan vertegenwoordigen, maar hoe weet ik dat ik dit geloof?”, maar ook vragen als: ‘Ik weet dat ik denk dat het regent, maar hoe weet ik dat ik deze overtuiging heb?’. Persoonlijke zelfkennis, aan de andere kant, kan begrepen worden als de kennis over jezelf als persoon, zoals bijvoorbeeld kennis van je handelingsmotieven, kennis van je karakter, gewoontes, ambities, wilskracht (of gebrek daaraan), diepste verlangens, onderdrukte angsten, enzovoorts. Het gaat niet simpelweg over kennis van je mentale toestanden, maar kennis van de elementen die cruciaal zijn voor wie je bent. Persoonlijke zelfkennis is het soort van zelfkennis waar de oude Grieken zich in interesseerden (denk aan de woorden “Ken uzelf” (gnothi seauton) die geschreven stonden op de Tempel van Apollo in Delphi) en waar het bovendien de meeste mensen buiten de filosofie om te doen is.

Met één of twee uitzonderingen zijn alle (analytisch) filosofen bezig met introspectieve zelfkennis en niet met persoonlijke zelfkennis. Een aantal filosofen opent zelfs hun boek met waarschuwingen tegen te hoge verwachtingen van lezers. Hun strekking: lezers die geïnteresseerd zijn in persoonlijke zelfkennis hebben niets te zoeken in het filosofische debat over zelfkennis. De nietsvermoedende lezer die in het zelfkennis debat verzeild raakt, leest dus niets over hoe we ons ‘zelf’ zouden kunnen kennen, of zelfkennis überhaupt belangrijk is en zo ja, waarom dan. In plaats daarvan wordt deze lezer geconfronteerd met argumenten en theorieën over hoe we onze eigen geest (mind) eigenlijk kennen, specifiek onze mentale toestanden, meer specifiek onze overtuigingen, meer specifiek onze bewuste overtuigingen en de dingen die ons in het hier-en-nu te binnen schieten. Kortom, niet het soort van zelfkennis dat we kunnen herformuleren in termen van een plicht zoals “Ken uzelf”. De inzet is eerder: “ik weet wat ik denk, maar goh, hoe weet ik dat eigenlijk?”.

Introspectieve zelfkennis is het vertrekpunt van dit proefschrift, in zoverre dit proefschrift zich engageert met het filosofische debat van zelfkennis. Echter, één van de onderliggende ambities van dit proefschrift is om te laten zien dat het onverstandig, zo niet onmogelijk, is om introspectieve zelfkennis onafhankelijk van persoonlijke zelfkennis te begrijpen en te analyseren. Dit is niet in alle gevallen

---

2 Twee uitzonderingen op de regel in de analytische filosofie zijn Quassim Cassam en Eric Schwitzgebel.
Om bijvoorbeeld te weten dat ik nu denk aan roze olifantjes, lijkt het niet nodig dat ik weet ‘wat voor persoon’ ik ben. De meeste filosofen zijn echter niet zozeer geïnteresseerd in dit soort vluchtige gedachten, maar in onze meer ‘stabiele’ mentale toestanden zoals onze overtuigingen en intenties. De onderliggende boodschap van mijn proefschrift is dat om kennis te verkrijgen van deze meer stabiele mentale toestanden een belangrijke mate van persoonlijke zelfkennis noodzakelijk is, en dat elke filosofische analyse die een strikt onderscheid hanteert tussen deze twee vormen van zelfkennis uiteindelijk tekortschiet.

2. De procedurele vraag en atomisme over zelfkennis

Het huidige filosofische debat is vooral bezig met de vraag of zelfkennis een kwestie is van ‘naar binnen kijken’ (introspectionisme), jezelf interpreteren (interpretationalisme), jezelf uitdrukken (expressivisme), of delibereren (rationalisme). Wanneer je aangeeft in het zelfkennis-debat werkzaam te zijn, is dan ook een van de eerste vragen waarmee je geconfronteerd wordt: ben je een introspectionist, interpretationalist, expressivist of rationalist of misschien een ‘pluralist’? Met andere woorden, het debat concentreert zich met name op de specifieke ‘methodes’ of ‘procedures’ van zelfkennis – hun onderlinge overeenkomsten en verschillen, en (in)comptabiliteiten. De focus ligt op wat ik noem de ‘procedurele vraag’ naar zelfkennis: dit is de vraag welke procedure of methode iemand moet volgen om tot een zogenaamde ware ‘zelf toeschrijving’ te komen.3

Er zijn een aantal andere belangrijke vragen die tot nu toe spijtig genoeg weinig aandacht hebben gekregen. Bén daarvan is de vraag naar de omstandigheden van zelfkennis. Wanneer Peter een zelf toeschrijving maakt, kunnen we niet alleen de vraag stellen welke procedure Peter volgde (“was Peter aan het introspecteren of aan het delibereren?”), maar kunnen we bijvoorbeeld ook vragen: wie is Peter? Wat is Peter van plan? Wat heeft Peter gedaan? Hoe voelt Peter zich? Wat weet Peter? Met wie praat Peter? Hoe wordt Peter door anderen gezien en behandeld? Hoe ziet Peter zichzelf? Enzovoorts.

---

3 Wanneer je een zelf toeschrijving maakt, schrijf je (in gedachten of in taal, bewust of onbewust) een bepaalde mentale toestand aan jezelf toe. Denk bijvoorbeeld aan: “Ik geloof dat Brexit een slecht idee is”, “Ik wil een koffie”, “Ik weet dat het gaat regenen” of “Ik hoop dat de vergadering niet doorgaat.”
Volgens velen zijn dergelijke vragen die betrekking hebben op de omstandigheden van iemands zelftoeschrijving (grotendeels) irrelevant wanneer we de vraag willen beantwoorden wat er voor het verkrijgen van zelfkennis nodig is. Volgens mij zijn ze echter wel degelijk van belang. De reden hiervoor is dat het beantwoorden van de vraag wat voor procedure iemand volgt, niet per se de vraag beantwoordt of iemands zelftoeschrijving ook zelfkennis oplevert. Misschien dat Peter wel volgens de regels had gedelibereerd, maar leverde dat alsnog geen zelfkennis op, bijvoorbeeld omdat Peter boos, verliefd, of onzeker was. Het is in dergelijke omstandigheden goed mogelijk dat je op een verkeerde manier delibereert (bijvoorbeeld met oogkleppen op). Om te weten of bijvoorbeeld Peters zelftoeschrijving een instantie is van zelfkennis, hebben we dus veel meer nodig dan een antwoord op de vraag wat Peter aan het doen was in termen van de procedures die hij al dan niet volgde.

Dit proefschrift hoopt een impliciete en hardnekkige aanname bloot te leggen, namelijk, de aanname dat vragen die betrekking hebben op de omstandigheden van zelfkennis van marginaal belang zijn voor zelfkennis. In dit proefschrift gebruik ik de term ‘atomisme’ om deze aanname aan te duiden. De notie van ‘omstandigheden’ begrijp ik daarbij opzettelijk in zeer brede zin. Tot mogelijk relevante omstandigheden voor het verkrijgen van zelfkennis reken ik bijvoorbeeld de aanwezigheid of afwezigheid van (1) ‘lichamelijke’ omstandigheden (zoals honger, dorst, pijn), (2) emotionele omstandigheden (woede, blijheid, oplichting, stress, verliefdheid), (3) sociaal-politieke omstandigheden (zoals vrijheid van meningsuiting, propaganda, onderdrukking). De alternatieve ‘holistische’ benadering ten aanzien van zelfkennis kan vervolgens begrepen worden als het omgekeerde van atomisme: het slechts volgen van een specifieke procedure of methode is geen garantie voor zelfkennis.

S. Samenvatting

Vasthouden aan het idee dat procedures het belangrijkst zijn wanneer we het over zelfkennis hebben. Of, met andere woorden, dat wanneer we een analyse willen geven van zelfkennis, we het vooral over de procedures moeten hebben. Waar de gematigd atomist denkt dat we het kunnen hebben over ‘uitzonderingen op de regel’, verdedig ik een holistische positie die stelt dat uiteindelijk de uitzonderingen de regel zijn.

3. Samenvatting van de hoofdstukken

De strategie van dit proefschrift is de limieten van atomisme bloot te leggen door een interne kritiek te leveren op twee van de meest dominante theorieën in het debat van zelfkennis: expressivisme en rationalisme, welke respectievelijk de twee delen van dit proefschrift behelzen. Het doel is daarbij aan het eind van het proefschrift een patroon te laten zien. De structuur van het proefschrift is daarom ook niet lineair: ik hoop aan te tonen dat, ondanks de belangrijke verschillen tussen expressivisme en rationalisme, deze theorieën geconfronteerd worden met soortgelijke problemen. De oorzaak van deze problemen schrijf ik toe aan hun gedeelde vooronderstelling van het atomisme.

3.1 Expressivisme

In het eerste deel van dit proefschrift (hoofdstukken 2 en 3) bespreek ik expressivisme. Volgens expressivisten zijn zelfoeschrijvingen zoals “Ik wil koffie”, “Ik hoop dat de vergadering komt te vervallen” of “Ik vrees dat het gaat regenen” geen rapportages over innerlijke mentale toestanden, noch slechts tekenen of mogelijke evidentie van wat voor mentale toestand je hebt, maar uitdrukkingen daarvan.4

In hoofdstuk 2 begin ik met het uiteenzetten van expressivisme. Opvallend genoeg wordt expressivisme vaak uitsluitend gezien als een (niet-epistemische) theorie over ‘eerste-persoons autoriteit’ in plaats van als (epistemische) theorie over zelfkennis. Ik laat daarom ten eerste zien dat expressivisme wel degelijk de middelen heeft om uit te leggen waarom we doorgaans zelfkennis verkrijgen wanneer we onszelf op een talige manier uitdrukken. De vervolgvraag is hoe we deze theorie van zelfkennis moeten evalueren, en, in het bijzonder, in hoeverre expressivisme een vorm van gematigd atomisme veronderstelt.

4 Er bestaat helaas geen goede vertaling voor de meer intuitieve Engelse frase ‘to speak one’s mind’.
De centrale vraag van hoofdstuk 3 is dan ook: maakt het uitoefenen van omstandigheden je jezelf uitdruk? Een radicaal atomistische variant van expressivisme zou deze vraag negatief beantwoorden. Maar het lijkt evident dat wanneer je jezelf uitdrukt tijdens een woedeaanval, je zelf-expressie niet correspondeert met wat je echt denkt of wilt ("Het was maar een momentopname"). Ik ga daarom in op de gematigd atomistische variant van expressivisme, volgens welke je zelfkennis verkrijgt door jezelf uit te drukken, behalve in ‘buitengewone’ omstandigheden, zoals wanneer je slaapwandelt, dronken bent, of een woedeaanval hebt. Ik problematiseer vervolgens deze gematigde variant van expressivisme door te laten zien dat er enorm veel omstandigheden zijn waarin je evenmin zelfkennis verkrijgt, zonder dat deze omstandigheden nu zo ‘buitengewoon’ zijn (zoals stress, vermoeidheid, verliefdheid of een ochtendhumeur). Bovendien is het zo dat wat voor de één als zelfkennis-beperkende omstandigheid geldt (zoals dronkenschap, verliefdheid of woede) voor een ander nog niet het geval hoeft te zijn. Dit lijkt te pleiten voor een holistische benadering van expressivisme waarbij de vraag of iemand zelfkennis heeft niet zozeer te maken heeft met of iemands zelftoeschrijving een instantie is van zelf-expressie, maar met de context waarbinnen die zelf-expressie plaatsvindt en hoe de persoon zich tot zichzelf en haar eigen verlangens, overtuigingen, intenties, angsten, enzovoorts, verhoudt. Echter, willen we expressivisme op holistische lijnen ontwikkelen, dan kunnen we niet anders dan zelf-expressies te zien als instanties van zelf-interpretatie. Met andere woorden: zelftoeschrijvingen als “Ik geloof dat ze het leuk vindt”, “Ik wil ermee stoppen” of “Ik hoop dat hij nog langskomt” vormen slechts mogelijke aanwijzingen voor wat daadwerkelijk onze overtuigingen, verlangens en andere mentale toestanden zijn. Maar zelftoeschrijvingen als zodanig begrijpen is helaas incompatibel met één van de centrale stellingen van expressivisme, namelijk, dat zelftoeschrijvingen niet slechts mogelijke indicaties zijn voor je mentale toestanden, maar een uitdrukkingen daarvan. De conclusie is dat holisme incompatibel lijkt te zijn met een expressivistische theorie over zelfkennis. Als de holistische benadering ten aanzien van zelfkennis de juiste is, dan moeten we expressivisme dus laten varen.

3.2 Rationalisme
In deel twee van het proefschrift (hoofdstukken 4 tot en met 7) bespreek ik rationalisme. Rationalisme is de theorie volgens welke zelfkennis een kwestie
Samenvatting

is van ‘praktisch delibereren’. Grofweg houdt dit in dat zelfkennis een actieve in plaats van een passieve bezigheid is. Om erachter te komen wat je overtuigingen zijn, bijvoorbeeld, moet je volgens de rationalist niet ‘naar binnen’ kijken maar juist naar ‘buiten’: je moet reflecteren op gebeurtenissen of feiten in de wereld. Centraal hierbij is de zogenaamde ‘transparantie procedure’. De procedure is als volgt: om erachter te komen of je de overtuiging hebt dat X (een vraag over het mentale domein) moet je de vraag beantwoorden of X waar is (een vraag over de wereld). Wanneer je de vraag over de wereld beantwoordt, kom je tot een oordeel, en dat oordeel is nu eenmaal niets anders dan je overtuiging. Kortom, je kunt er niet achter komen wat je gedachtes zijn door je eigen gedrag te observeren of door een hersenscan te laten maken, maar je moetactief bepalen wat je ergens van vindt door te reflecteren op situaties in de wereld.

In het eerste introducerende hoofdstuk geef ik allereerst een overzicht van de theorie zoals deze uiteen is gezet door Richard Moran, waarna ik inga op zijn centrale stelling, namelijk, dat we kennis kunnen verkrijgen van onze eigen mentale toestanden door de eerder genoemde transparantie procedure te volgen. Ik bediscussieer een aantal veelvoorkomende tegenwerpingen op rationalisme en bespreek kort manieren om deze te weerleggen. Ik eindig door in te gaan op de relatie tussen oordelen aan de ene kant en overtuigingen aan de andere kant.

In hoofdstuk 5 onderscheid ik een ‘radicale’ en een ‘gematigde’ atomistische variant van rationalisme. Ik leg mij vervolgens toe op de gematigde variant, volgens welke het volgen van de transparantie procedure alleen zelfkennis oplevert wanneer de persoon deze procedure in de ‘juiste’ of ‘normale’ omstandigheden heeft gevolgd. Vervolgens is de vraag wat we dan onder de juiste of normale omstandigheden moeten verstaan. Hiertoe bespreek ik het voorbeeld van de Boze Echtgenoot. De Boze Echtgenoot is iemand die overtuigd is geraakt van Moran’s theorie, en besluit om de vraag te beantwoorden “Geloof ik dat een echtscheiding een goed is?” (een vraag over haar mentale toestand) door de corresponderende vraag “Is een echtscheiding een goed idee?” te beantwoorden (een vraag over de wereld). Omdat de Boze Echtgenoot deze procedure volgt in een opwelling van woede, kan ze alleen maar redenen vóór een scheiding bedenken, en ziet ze geen enkele reden om samen te blijven met haar partner. De vraag is: heeft de Boze Echtgenoot nu zelfkennis verkregen door deliberatie, zoals we volgens rationalisme zouden moeten veronderstellen, of leidt het juist tot zelfbedrog? Ik bespreek de mogelijke reacties

5 Ook hier leent het Engels zich beter: ‘to make up one’s mind’.
Samenvatting

die de rationalist zou kunnen geven, en laat zien dat geen van die reacties afdoende zijn. Ik concludeer dat de gematigde versie van rationalisme, die een beroep doet op (impliciete) *ceteris paribus*-clausules, uiteindelijk geen bevredigend verhaal kan vertellen over wat nu de ‘juiste’ of ‘normale’ omstandigheden van deliberatie zijn. Althans, zonder zelf-deceptie vanaf het begin uit te sluiten en daarmee aan te nemen wat precies op het spel staat.

Een voor de hand liggende optie, voor de rationalist, is om dan toch voor de radicale variant van rationalisme te kiezen en simpelweg geen beroep te doen op juiste of normale omstandigheden van zelfkennis. In hoofdstuk 6 bespreek ik hiertoe de theorie van Matthew Boyle. Volgens Boyle zijn er geen gevallen – en dus ook geen omstandigheden – waar er een kloof zou kunnen ontstaan tussen iemands oordeel (*judgement*) enerzijds en iemands overtuiging (*belief*) anderzijds. Boyle’s lijkt theorie daarom lijkt te kwalificeren als een vorm van radicaal atomisme. In het hoofdstuk bespreek ik vervolgens een aantal voorbeelden, waaronder een voorbeeld over impliciete oordelen, waaruit lijkt te volgen dat het wel degelijk mogelijk is om tot een oordeel te komen zonder de corresponderende overtuiging te hebben. Mijn doel in dit hoofdstuk is om te laten zien dat de vraag of Boyle’s theorie van zelfkennis plausibel is afhankelijk is van de metafysische vraag wat we onder mentale toestanden, overtuigingen in het bijzonder, verstaan. Om die reden contrasteer ik Boyle’s metafysische opvatting over mentale toestanden met een alternatieve ‘dispositionele’ metafysica, waarbij ik mij baseer op het werk van Eric Schwitzgebel. Mijn conclusie luidt dat het idee dat radicaal atomisme implausibel is alleen volgt gegeven een (impliciet) antwoord op de metafysische vraag wat we überhaupt verstaan onder, bijvoorbeeld, een ‘overtuiging’. Deze metafysische vraag is op haar beurt afhankelijk van meer fundamentele kwesties, zoals wat de rol en relevantie is van intuities die we hebben wanneer we geconfronteerd worden met bepaalde voorbeelden. Mijn conclusie wat betreft radicaal atomisme is dus voorzichtiger dan die over gematigd atomisme. De conclusie die ik trek is dat radicaal atomisme bij nader inzien niet zonder argumentatie terzijde kan worden geschoven, omdat de verwervende ervan afhankelijk is van bepaalde metatheoretische aannames. De meer genuanceerde conclusie is dan ook dat deze metatheoretische vraagstukken relevant zijn voor het bestuderen van zelfkennis.

In hoofdstuk 7 bespreek ik een zogenaamde normatieve interpretatie van het rationalistisch project, en ga ik meer in het algemeen in op de vraag hoe zelfkennis en autonomie zijn gerelateerd. Ik doe dit aan de hand van een aantal voorbeelden,
zoals zelfkennis verkregen in omstandigheden van brainwashing en onderdrukking. Vervolgens vergelijk ik de normatieve lezing van rationalisme met zogenaamde ‘procedurele’ en ‘relationele’ theorieën van autonomie. Op basis van deze discussie concludeer ik dat we niet slechts een onderscheid moeten maken tussen verschillende objecten van zelfkennis en verschillende routes naar zelfkennis maar dat we ook een onderscheid moeten maken tussen verschillende soorten zelfkennis. Ik stel voor dat we een onderscheid moeten maken tussen (1) introspectieve, (2) authentieke, en (3) autonome zelfkennis. Mijn hoop is dat inzichten uit discussies over autonomie ons iets leren over (de grenzen van) atomisme over zelfkennis; ik sluit het hoofdstuk daarom af door een stap terug te doen en te reflecteren op de onderliggende metafysische en epistemische bronnen van atomisme.

In het concluderende hoofdstuk maak ik de balans op, vat ik de belangrijkste punten uit het proefschrift samen, reflecteer ik op de vraag hoe dit proefschrift bijdragt aan ons begrip van zelfkennis, en zet ik mogelijke lijnen uit voor toekomstig onderzoek.

4. Tot besluit

Ik heb in dit proefschrift, met opzet, geen ‘nieuwe’ theorie van zelfkennis geïntroduceerd of verdedigd. Er zijn mijns inziens al genoeg theorieën van zelfkennis. Mijn hoop is de grenzen van atomistische, procedure-gerichte theorieën bloot te leggen en om een belangrijke impliciete aannamie van het debat zelf te bevragen. Ik heb beargumenteerd dat iedere plausibele theorie van zelfkennis holisme serieus moet nemen, omdat een plausibel verhaal over zelfkennis afhangt van hoe we tal van metafysische, epistemologische, morele en metatheoretische vraagstukken beantwoorden.
Fleur Jongepier received her Bachelor's degree in Philosophy from Utrecht University (2009) and her Research Master's degree in Philosophy from the Radboud University Nijmegen (2001, summa cum laude). Her thesis was supervised by prof. dr. Marc Slors, prof. dr. Jan Bransen and prof. dr. Quassim Cassam. During her PhD project, Fleur spent one term at the Heinrich Heine Universität Düsseldorf and two terms at the University of Warwick. During her PhD, Fleur co-authored a Dutch book on the philosophy of the self, worked as an editorial assistant for Philosophical Explorations and co-founded a Dutch philosophy blog for a broader audience, Bij Nader Inzien. Currently, she teaches courses in theoretical and practical philosophy at the Radboud University Nijmegen and the Erasmus University of Rotterdam. Starting autumn 2017, she will do a postdoc on the role and value of self-knowledge in contemporary liberalism in Cambridge (UK).